EAGLES

Expert Advisory Groups for Language Engineering Systems

Spoken Language Working Group

Handbook of Audiovisual, Multimodal and Spoken Dialogue Systems

Resources and Terminology for Development and Product Evaluation

Dafydd Gibbon, Inge Mertins, Roger Moore (eds.)

Dedicated to the memory of our colleague, co–author and friend

Christian Benoît

Contents

List of Figures

List of Tables

Editorial preface

The present volume is a companion to the *Handbook of Standards and Resources for Spoken Language Systems* (Gibbon et al. 1997), and addresses decision makers, developers and advanced students in the human language technologies who are looking for guidance in areas such as the evaluation of consumer off-the-shelf products, multimodal and audiovisual systems, terminology, and current resources for system development and evaluation.

The publication of the *Handbook* marked a milestone in the development of the spoken language technologies. Standards, in the sense of consensus about the best laboratory practice of the time, were formulated by leading experts and negotiated interactively with a large number of European laboratories working in the field. As with any handbook, even during the authoring phases it was clear that the *Handbook* could not keep pace with all developments in a rapidly expanding field, and the more modest aim of creating a standard common platform for best practice in the field was pursued. A mode of dissemination in which knowledge can be continually updated was considered; maintenance cost would, however, have been prohibitively high.

To the specialist, a number of gaps in the *Handbook* are evident, and full coverage of the field is currently an extremely elusive goal. The editors of the first *Handbook* selected a number of topics for further treatment, in consultation with project partners and other colleagues in the field. These topics formed the substance of work in the EAGLES Phase II Project (LE3-4244 10484/0), Work Packages 4, 5 and 6 (constituting the Spoken Language Working Group, SLWG), and were developed using the consensus negotiating procedures which had already been tried and tested during the development of the first EAGLES *Handbook*. The present volume presents the results of this work, which is indicated in the title *Handbook of Audiovisual, Multimodal and Spoken Dialogue Systems: Resources and Terminology for Development and Product Evaluation*. Chairman of the SLWG was Roger Moore, coordinator and general editor SLWG was Dafydd Gibbon, and technical editor was Inge Mertins. Editorial workshops were held in Bielefeld, Germany, in January 1998 and in Leusden, Netherlands, in June 1998, attended by the SLWG technical authors and academic and industrial reviewers.

The Chapter *Representation and annotation of dialogue* was coordinated by Geoffrey Leech (Lancaster University, UK) and the main technical author was Martin Weisser (Lancaster University, UK) with substantive contributions from Andrew Wilson (Chemnitz University of Technology, Germany), and Martine Grice (Universität des Saarlandes, Saarbrücken, Germany). The Chapter was designed and edited in close consultation with Kerstin Fischer (Computer Science Department, Universität Hamburg, Germany), Susanne Jekat (Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany), Elisabeth Maier (SBC / IT Camp, Basel, Switzerland), Paul McKevitt (Aalborg University, Denmark), Jean Carletta (University of Edinburgh, UK), Joaquim Llisterri (Universitat Autònoma de Barcelona, Spain). Help from the following is also gratefully acknowledged: Anton Batliner, Niels Ole Bernsen, František Čermak, Alain Couillault, Paul Dalsgaard, Mika Enomoto, Ulrich Heid, Arne Johnsen, Magne

Johnsen, Andreas Kellner, Gerry Knowles, Klaus Kohler, David Milward, Norbert Reitlinger, Paul Rogers, Geoff Sampson and Fernando Sánchez-León.
Chapter *Audio-visual and multimodal speech-based systems* was initially coordinated by Christian Benoît, whose untimely death in early 1998 left the project without one of its most active and stimulating contributors. We are very grateful to Catherine Pelachaud, Universita di Roma "La Sapienza", Italy, who took on the daunting task of coordination and technical authoring and brought the work to an impressive conclusion. Considerable portions of the Chapter were contributed by Bernhard Suhm, Universität Karlsruhe, Germany and Jean-Claude Martin, LIMSI-CNRS, Orsay, France, and Lambert Schomaker, Nijmegen Institute for Cognition and Information, Nijmegen, Netherlands.
The Chapter on *Consumer off-the-shelf (COTS) speech technology product and service evaluation* was coordinated by David van Leeuwen and Herman Steeneken; the main technical author was David van Leeuwen, with additional contributions by the participants in the Bielefeld and Leusden workshops.
The Chapter *Terminology for spoken language systems*, and the design and construction of the *EAGLET* terminology database was coordinated by Dafydd Gibbon and Inge Mertins, Universität Bielefeld. Extensive technical and editorial assistance was provided by Thorsten Trippel, Silke Kölsch and Michaela Schulte, Universität Bielefeld. Substantive contributions were made by Lou Boves (Katholieke Universiteit Nijmegen, Netherlands), Melvyn Hunt (Dragon Systems UK), John McNaught (University of Manchester Institute of Technology, UK), and Klaus–Dirk Schmitz, Fachhochschule Köln, Cologne, Germany. The *EAGLET* database interface was designed by Holger Ulrich Nord and Dafydd Gibbon, Universität Bielefeld, and implemented in JavaScript and mSQLlite by Holger Ulrich Nord. The *EAGLET* terminological database can currently be consulted at
`http://coral.lili.uni-bielefeld.de/EAGLES/SLWG/TERMBANK/interface.shtml`
The Chapter containing *Reference materials* was coordinated and authored by Christoph Draxler, University of Munich, Germany, with support from many colleagues around the world, in particular from Metin Erdogan, Middle East Technical University of Ankara, Turkey, and Russ Wilcox (E Ink Corporation, Cambridge, USA).
Inge Mertins was responsible for technical editing and production, including coordination with technical authors. Holger Ulrich Nord and Thorsten Trippel took care of numerous technical problems which arose in LaTeX and HTML conversion. We hope to have achieved an acceptable combination of quick publication in a rapidly developing field with painstaking reference quality, and ask the reader's indulgence for any remaining infelicities or errors.
We wish to thank the coordinators of the entire EAGLES project in Pisa, Antonio Zampolli and Nicoletta Calzolari, for their foresight, patience, advice and organisational skills, John McNaught, University of Manchester Institute of Science and Technology, the EAGLES general editor, for his continuing expert support, and, *sine qua non*, to the staff of CEC DG XIII in Luxembourg, in particular Roberto Cencioni and Norbert Brinkhoff–Button.

Main technical authors

Representation and annotation of dialogue	Martine Grice Geoffrey Leech Martin Weisser Andrew Wilson
Audio-visual and multimodal speech-based systems	Christian Benoît Jean-Claude Martin Catherine Pelachaud Lambert Schomaker Bernhard Suhm
Consumer off-the-shelf (COTS) speech technology product and service evaluation	David van Leeuwen Herman Steeneken
Terminology for spoken language systems	Dafydd Gibbon Inge Mertins
Reference materials	Christoph Draxler
SAMPA and X-SAMPA phonetic symbols	Inge Mertins
The EAGLET Term Database	Dafydd Gibbon Silke Kölsch Inge Mertins Michaela Schulte Thorsten Trippel

1 Representation and annotation of dialogue

1.1 Introduction

1.1.1 Goals

The main purposes of this chapter are to present a survey of current and developing work in the areas of research and development with respect to integrated spoken and written language resources, and to provide preliminary guidelines for the representation or annotation of dialogue in resources for language engineering (see also Gibbon et al. 1997, pp. 146–172).

The terms *representation* and *annotation* have distinct conventional uses in this chapter. 'Representation' is used for the orthographic transcription of a dialogue, giving the basic information about what was said, by whom it was said, and other necessary details. The term 'annotation', on the other hand, is used for the additional levels of linguistic information which are added to the orthographic transcription. This conventional usage needs some brief preliminary explanation.

In reference to corpora of written language, the distinction is relatively clear: the *representation* of a text is the encoding of the orthographic form of the text itself, either as straight *ASCII* text, or in some mark-up system such as is provided by the TEI (Text Encoding Initiative: see Sperberg-McQueen and Burnard (1994)). On the other hand, *annotation* constitutes additions to that basic representation, providing various levels of linguistic analysis (such as morphosyntactic, syntactic, semantic levels: see Garside et al. (1997), pp. 1–19).

However, with a corpus of spoken language, the orthographic transcription does not have the same status of basic representation of the data, being itself a level of linguistic abstraction from the speech signal. (The term *transcription* above corresponds to *representation* in the sense that an orthographic transcription, say, undertakes to represent, as a verbatim record, what was said by the speakers in a dialogue.)

Traditionally, users of the transcription have treated it as a useful substitute for the actual sound recording, in deriving from it the wording and sense of the spoken message. It is clear, however, that this substitute use is not a desirable use of an orthographic transcription in spoken language resources for language engineering (LE). From the point of view of speech analysis, an orthographic transcription is more remarkable for what it excludes than for what it includes. Moreover, it is assumed, with modern technological progress, that all users of a spoken language corpus will have ready access to the sound recording, which can therefore be regarded as the basic record of any spoken language data.

Although this means that the orthographic transcription loses its observational primacy, there is still an important sense in which the orthographic transcription is the primary level of abstraction from the data, involving as little interpretation as possible. A common format for orthographic *representation* of dialogue is therefore highly desirable for the exchange (and automatic processing) of the data. Other levels of information, *annotations*, are added to this baseline

verbatim record, without which it would be difficult to make sense of them. The goals of the present chapter are:

1. to identify and describe linguistic phenomena specific to spoken language and in particular to dialogue, which require special provision for annotation.
2. to survey, compare and analyse methods, solutions and practices proposed to represent and annotate these phenomena.
3. to propose guidelines for annotating the identified dialogue-specific phenomena at various levels.
4. to integrate these recommendations or guidelines, in a coherent way, into the overall annotation guidelines.

The present chapter primarily addresses the second and third of these goals, while not overlooking the other goals where relevant.[1]

1.1.2 What is meant by 'Integrated Resources'?

In the 1980s, the *speech community* and the *natural language community* were effectively two research communities working on a common subject matter – human language – but otherwise having little communication with one another. Towards the end of the twentieth century this situation changed, simply because many of the emerging new applications of language engineering (LE) involve both the domain of 'speech' and that of 'natural language'. It has become evident that these communities have to pool their specialist knowledge and to strive to become a single research community (see Llisterri 1996, Section 2.2 on the need for such convergence). And not only this: with the advent of multimodal systems, the requirement for multidisciplinary exchange and cooperation is becoming even stronger (see Chapter 2).

The Natural Language (written language) community has in the past concentrated both on (a) written language processing, and/or on (b) the processing of language at those levels of analysis (e.g. syntactic, lexical, semantic, pragmatic levels) which in general apply both to written and spoken language, and where the distinction between the two channels is relatively unimportant. The speech community, on the other hand, has in the past tended to concentrate on levels of analysis which relate fairly directly to the spoken signal.

However, it has already become clear that this division of interest can no longer be maintained: many of the most forward-looking and challenging applications of LE today (e.g. high-quality speech synthesis, large-vocabulary speech recognition, speech-to-speech translation, dialogue systems, multimodal systems) involve both low-level and high-level processing. A parser, for example, is needed for processing both spoken and written language data. Moreover, current R&D (research and development) is working towards integrated spoken language systems undertaking all levels of speech understanding and speech synthesis, such as are needed for the appropriate understanding and production of speech in dialogue.

[1] A complementary European project in the dialogue area is MATE (Multi-Level Annotation Tools Engineering). Two areas handled by MATE but not dealt with in this chapter are co-reference and multilingual annotation. Further details may currently be be found at the following URL:
"`http://www2.echo.lu/langeng/projects/mate/summary.html`" -Ed.

1.1.3 Limitations

Hence *integrated resources for spoken and written language* refers to LE resources which are to be shared by both speech and natural language processing research. They include *corpora*, *lexica, grammars* and *tools*. For example, lexica for integrated resources should provide for the integration of lexical information as a common resource relating to both spoken and written language (while allowing for their expedient separation where the need arises). There is also need for integration in a further sense: resources such as lexica and corpora should be consistent with one another so that information can be easily exchanged between them. Similarly, tools should be capable of processing data in terms of the representations used for other resources. What can be achieved within the scope of this chapter, however, is limited in several ways.

1.1.3.1 Focus on corpora

In this chapter we restrict our attention primarily to (a) *corpora*, because this is the area in which the need for standardisation arises most compellingly. We have not been able to consider (b) lexica (but see Gibbon et al. (1997)), (c) grammars and (d) tools in any detail. On the other hand, (d) *tools* have been given some attention here (see especially 1.8.1.9), since the transcription and annotation of spoken corpora are in part constrained by what tools exist or can be developed to facilitate and integrate these tasks. The other Chapters of this volume should also be consulted.
A *corpus* in this context is defined as a body of spoken language data which has been recorded, has been transcribed (in part or in toto) and documented for use in the development of language engineering (LE) systems, and in principle at least, is available for use by more than one research team in the community. The need for *standards*, or rather *guidelines*, for the representation and annotation of spoken language data arises primarily because of the need to ensure interchangeability of data, between different sites, in a multilingual community such as the European Union, so that progress in the provision of resources can be shared and can provide a springboard for further collaboration and advances in the future.

1.1.3.2 Focus on dialogue corpora

Apart from the focus on corpora, there is an additional restriction on the scope of this chapter, which is the decision to limit the treatment of integrated resources to *dialogue* corpora. For present purposes we define a dialogue as a discourse in which two or more participants interact communicatively, and where at least one of the participants is human. This covers cases of human–machine as well as human–human dialogue. In principle, this can include not only spoken dialogue, but also written dialogue, where for example a human participant interacts with a machine via a keyboard. However, in practice, this chapter will mainly focus on spoken dialogue.
Walker and Moore (1997), p. 1 point out the important role dialogue now plays in LE:

> In the past, research in this area focused on specifying the mechanisms underlying particular discourse phenomena; the models pro-

> posed were often motivated by a few constructed examples ... Recently however the field has turned to issues of robustness and the coverage of theories ... this new empirical focus is supported by several recent advances: an increasing theoretical consensus on discourse models; a large amount of on-line dialogue and textual corpora available; and improvements in component technologies and tools for building and testing discourse and dialogue testbeds. This means that it is now possible to determine how representative particular discourse phenomena are, how frequently they occur, whether they are related to other phenomena, what percentage of the cases a particular model covers, the inherent difficulty of the problem, and how well an algorithm for processing or generating the phenomena should perform to be considered a good model.

Research in this field can be either close to or distant from practical commercial or industrial applications. Less applications-oriented studies may concentrate on certain modules or levels of analysis to the exclusion of others. All such studies can, however, be valuable in leading to richer and more precise models of human dialogue behaviour. What is particularly significant, in task-oriented dialogue annotation, is that all levels of analysis can be seen as culminating in the pragmatic level, where the communicative function of the dialogue is characterised in terms of dialogue acts. Dialogue, in this perspective, is the nexus which gathers all areas of integrated resources research and development into a practical focus.

1.1.3.3 Focus on applications-oriented task-driven dialogue

Third, a third limitation on our study of integrated resources is that we focus attention primarily on applications-oriented task-driven dialogue, bearing in mind that the present objective is to promote standards in LE, rather than more generally in linguistics or social science, in such fields as dialectology, sociolinguistics, discourse analysis or conversational analysis. In recent years, corpora of spoken dialogue have been compiled for a wide variety of reasons. For example, one well-developed initiative is the CHILDES database (MacWhinney 1995), which sets standards for the interchange of data between researchers in the area of child language acquisition. Another instance of incipient standardisation is the spoken subcorpus of the BNC (British National Corpus) (see Burnard 1995), which contains about 10 million words of spoken English, all transcribed and marked up in accordance with the guidelines of the TEI (Text Encoding Initiative) (see Johansson 1995). The need for a standard in this case had to be reconciled with the requirement of a corpus large enough to be usable for dictionary compilation and other wide-ranging fields of linguistic research. Other examples could be added: there can be many reasons for introducing standards/guidelines for representation of dialogue, apart from those which are most salient to the LE community. While it is instructive to take note of these other initiatives, especially where they come to conclusions of value to LE specialists, they cannot be treated unquestioningly as models to be followed in this chapter.

1.1.3.4 Restriction to certain tiers of representation/annotation

A final limitation of this task is the following. We have restricted attention to certain levels or tiers of representation/annotation where it is felt that there is a particular need to propose guidelines. The levels for which a representation or annotation of dialogue can be provided are many: see Gibbon et al. (1997), p. 149 ff., for a reasonably complete list. However, for the present purpose we disregard semantic annotation, which is being dealt with elsewhere, and we also largely ignore phonetic/phonemic and physical levels of transcription, on which considerable standardising work has been done already (see, for example, Gibbon et al. (1997), pp. 688–731, on SAMPA). We confine our attention to the following levels:

- *general* (Section 1.3) – general coding issues (i.e., Standard Generalized Markup Language (SGML), eXtensible Markup Language (XML), etc.)
- *orthographic* (Section 1.4) – constructing a verbatim record of the dialogue
- *morphosyntactic* (Section 1.5) – part-of-speech or word-class tagging
- *syntactic* (Section 1.6) – treebanks (either partially or fully parsed)
- *prosodic* (Section 1.7) – representation of suprasegmental phenomena such as accentuation and phrasing, using annotation systems such as ToBI, TSM or INTSINT and automatic analysis of acoustic parameters (e.g. fundamental frequency)
- *pragmatic* (Section 1.8) – functional units at macro-, meso- or speech act levels in dialogue

At the same time, we assume that all the different levels of annotation above need to be integrated in a multi-layer structure, and linked through relative or absolute time alignment to the sound recording.
It has to be admitted that these levels (particularly the orthographic, pragmatic and prosodic) do not yet show a highly developed trend towards standardisation. Consequently, this chapter concentrates heavily on surveying current practices, and on identifying those which may be considered good models for others to follow. Further, it is not at present possible to give guidelines for dialogue in multimodal systems. Inevitably, we will have overlooked some significant current research, and will have also drawn tentative conclusions which others will contest.

1.2 A preliminary classification of dialogue corpora

Before we turn to the different levels of representation or annotation, it is important to consider the various types of dialogue which have been investigated or modelled for LE purposes. This section contains an outline of some of the different types of dialogue that occur in different research projects and that are to some extent the basis for finding ways of categorising and identifying dialogue acts in Section 1.8 below.
For example, one of the most general types of dialogue concerns airline, train timetable or general travel inquiries. The German VERBMOBIL project specifically deals with *appointment scheduling* and *travel planning* tasks, while the TRAINS corpus developed at the University of Rochester, USA, deals with developing *plans to move trains and cargo* from one city to another. One of the major dialogue projects in the US, the ATIS (Air Travel Information Service)

project, deals strictly with *providing air travel information* to customers, and major companies, such as *Texas Instruments* and *AT&T*, have been involved in the collection and evaluation of the corpus.[2] Other dialogue projects involve *furnishing rooms interactively* (COCONUT, University of Pittsburgh), *giving directions on a map* (HCRC, University of Edinburgh) and *explaining cooking recipes* (Nakatani et al. 1995). These are just a few of the tasks to which dialogue projects have devoted attention up to the present.
As yet, there does not seem to exist any complete or systematic typology of dialogues, which makes it difficult (for example) to establish a complete list of all the goals that might be involved in the annotation and use of dialogue material.[3] Broadly, dialogues can be classified and described by reference to either *external* or *internal* criteria. The former include situational and motivational factors. The latter include formal or structural factors, especially how the dialogue breaks down into smaller units or segments such as turns and dialogue acts (see 1.8). However, there seems to be a definite need for such a classification in order to establish a valid list of criteria that are to be used for annotation: one that is based on actual experience and not on pure introspection. Such a list of criteria can then serve as a basic reference model that would need to be expanded only for special purposes that did not fit any of the existing criteria. A starting point for establishing such a typology is suggested in 1.2.2.

1.2.1 Dialogue acts

However, first it will be convenient to introduce here the term *dialogue act*, which will recur in this chapter, and will be more fully explained in Section 1.8. Dialogue acts are the smallest functional units of dialogues, and are utterances corresponding to speech acts such as 'greeting', 'request', 'suggestion', 'accept', 'confirm', 'reject', 'thank', 'feedback'. When considering the overall communicative function of dialogues, it is as well to bear in mind that for annotation as well as for processing purposes, they are seen as decomposable into such basic communicative units.

1.2.2 Towards a dialogue typology

In principle, we need a typology of dialogues geared towards the needs of LE as they can be foreseen at present. In practice, present research not surprisingly shows a heavy concentration on certain rather straightforward kinds of dialogue: those with the features marked ** below.

A. NUMBER OF PARTICIPANTS

A.1 TWO PARTICIPANTS **
A.2 MORE THAN TWO PARTICIPANTS

[2] More information on some of the work that has so far been done on the ATIS corpus can be found in Section 1.7.1.6.

[3] Useful background for both external and internal aspects of dialogue description are to be found in the sociolinguistic literature of the past 30 years, for which Dell Hymes's work on the 'components of speech' and 'rules of speaking' is a seminal starting point (see Hymes 1972/1986).

Most dialogues in LE research have two participants only (at any one stage).[4] More than two participants greatly complicate the task not only of collecting data, but of modelling all levels of analysis and synthesis. The number of overlaps is likely to increase, thereby influencing the quality and analysability of speech and the complexity of annotation.[5]

B. TASK ORIENTATION

B.1 TASK-DRIVEN **
B.2 NON-TASK-DRIVEN

Almost all dialogues in LE research are task-driven; that is, there is usually a specific task (or possibly more than one task), which at least one participant aims to accomplish with the aid of the other(s). An example is the Edinburgh Map Task Corpus (Anderson et al. 1991) in which one participant guides another to trace a route on a map. Others are the TRAINS corpus (Allen et al. 1996), in which speakers develop plans to move trains and cargo from one city to another and the VERBMOBIL dialogues that deal with appointment scheduling and travel planning. In contrast, most conversational dialogues would be classified as non-task-driven.

C. APPLICATIONS ORIENTATION

C.1 APPLICATIONS-ORIENTED **
C.2 NON-APPLICATIONS-ORIENTED

Applications orientation is a relevant parameter particularly among dialogues which are task-driven. The Map Task corpus may be cited as an example of a non-applications-oriented dialogue type. However valuable its contribution to research, it cannot be seen to have direct commercial or industrial applications. In contrast, dialogues which have clear application to useful human–machine interfaces, such as those dealing with airline or hotel reservations, may be classified as applications-oriented.

D. DOMAIN RESTRICTION

D.1 RESTRICTED DOMAIN **
D.2 UNRESTRICTED DOMAIN

Again, most dialogues in LE are restricted to a relatively tightly-defined domain of subject-matter. All three of the examples in 2. above belong to a restricted domain. (On the other hand, an everyday dialogue at the dinner table would be an example of unrestricted domain.)
A typology of domains follows naturally, at this point, under D.1. The following are purely exemplificatory:

[4] An exceptional case is the three-participant dialogue scenarios used in some VERBMOBIL projects, involving two negotiators and an interpreter/intermediary (see Jekat et al. 1997).

[5] It is of interest to mention, however, that large spoken corpora such as the 4.2-million-word demographic component of the BNC (British National Corpus), although of little value to LE, often contain dialogues with many participants (see Burnard 1995).

D.1 RESTRICTED DOMAIN

D.1.1 TRAVEL **
D.1.2 TRANSPORT **
D.1.3 BUSINESS APPOINTMENTS **
D.1.4 TELEBANKING
D.1.5 COMPUTER OPERATING SYSTEMS
D.1.6 DIRECTORY ENQUIRY SERVICES
D.1.7 (etc.)

Subclassification may also be needed: e.g., under 'travel', air travel, hotel bookings, and rail travel are subdomains.

E. ACTIVITY TYPES

E.1 COOPERATIVE NEGOTIATION **
E.2 INFORMATION EXTRACTION **
E.3 PROBLEM SOLVING
E.4 TEACHING/INSTRUCTION
E.5 COUNSELLING
E.6 CHATTING
E.7 (etc.)

Alongside domain, the *activity type* (Levinson 1979) to which the dialogue belongs is another variable defining the type of dialogue, particularly in terms of the constraints on the dialogue roles adopted by participants. For example, under E.1 in the VERBMOBIL three-agent dialogues the participants may be characterised as two 'negotiators' and one 'interpreter/intermediary'. In E.2, the two participants may be characterised as 'customer' and 'service-provider'. In current dialogue research, there is a major division between two leading paradigms: *cooperative tasks* between human participants (such as negotiating appointments) (E.1) and *information extraction tasks* (such as obtaining information on a computer operating system) in which a human agent interrogates a computer system (or a human surrogate for a computer system) (E.2) (see Gibbon et al. (1997), p. 598 on 'dialogue strategies'). Other task-driven activity types include problem-solving (as in the Map Task Corpus), teaching/instruction, counselling, chatting and interviewing.
Relations between variables (C.) 'applications orientation' and (E.) 'activity type' are obvious. On the whole, applications-oriented dialogue corpora at present will be characterised as either E.1 or E.2. Similarly, constraints on (D.) domain and (E.) activity type are clearly interrelated variables. They help to delimit the nature of the *task* (see B.1 below). However, they can be considered independently: the Linguistic Data Consortium (LDC) Switchboard Corpus has dialogues in which speakers share a pre-determined topic or domain of discourse; however, the activity type is not constrained in any specific way.
At this point, we turn to a classification of *tasks*, which logically could have been slotted in earlier, after 'B. Task Orientation'. The reason why it has been postponed is to show the relation of interdependence between, on the one hand, task and domain, and on the other hand, task and activity type.

B.1 TASK

B.1.1 Negotiating appointments and travel planning (VERBMOBIL) **

B.1.2 Answering airline/travel inquiries (ATIS) **

B.1.3 Developing plans for moving trains and cargo (TRAINS) **

B.1.4 Furnishing rooms (COCONUT) **

B.1.5 Giving directions to find a route on a map (Map Task)

B.1.6 (etc.) ...

Distinct tasks can be informally defined by the intention(s) of participants, the illocutionary function(s) of their utterances (Mc Kevitt et al. 1992) or by the end state which defines the successful accomplishment of the task. The number of tasks for which dialogue takes place is very large. Also, the amount of detail which may be specified to define the task for a particular dialogue is open-ended. Hence no closed set of 'task attributes' can be reasonably specified. As an example, consider the following as a succinct definition of the Map Task scenario (Thompson et al. 1995, p. 168):

> Each participant has a schematic map in front of them, not visible to the other. Each map is comprised of an outline and roughly a dozen labelled features (e.g. 'white cottage', 'Green Bay', 'oak forest'). Most features are common to the two maps, but not all. One map has a route drawn in, the other does not. The task is for the participant without the route to draw one on the basis of discussion with the participant with the route.

It is sound practice to keep 'task' and 'domain' as separate parameters, recognising that when a dialogue system has to be built for a particular application, the two parameters need to be combined for the specification of that particular system. The separation of task and domain is particularly useful for the typology both of dialogues and of dialogue acts (see Section 1.8 below): it enables generalisations across indefinitely many different tasks and different domains to be built into the typology, and into the construction of suitably generic dialogue system software.

F. HUMAN/MACHINE PARTICIPATION

F.1 HUMAN–MACHINE DIALOGUE

F.1.1 SIMULATED (WIZARD OF OZ) **

F.1.2 NON-SIMULATED

F.2 HUMAN–HUMAN DIALOGUE

F.2.1 MACHINE-MEDIATED **

F.2.2 NON-MACHINE-MEDIATED

In corpus-driven methodology, there is always a problem of matching the naturally-collected data to the needs of the artificial LE system. One problem of dialogue research where this shows up strongly is in our lack of knowledge of how human beings will behave when conversing with computer dialogue systems. How far will they adapt, when talking to a machine, so that their dialogic behaviour is 'unnatural' by the standards of human–human dialogue? To answer this question, *Wizard of Oz (WOZ) experiments* (see Gibbon et al. 1997, pp. 104–105, 143, 375–379) have been set up to simulate the behaviour of a machine in dialogue with a human being, and to record both the behaviour of the machine and the behaviour of the human being who believes he or she is interacting with a machine.
The other option under F.1, non-simulated human–machine dialogue, is clearly of limited value for R&D purposes, unless the computer system has already attained a basically satisfactory level of functionality. This has been described as a system-in-the-loop method (see Gibbon et al. 1997, p. 581).
To understand the way in which humans interact with machines is also important because there are many types of machine-mediation that may each influence the way dialogue is conducted in a particular way, both when communicating with the computer and with another human via the computer. Even using the telephone may be considered a form of machine-mediation restricting the transmission channel, although it is something we accept as part of our everyday lives and tend not to consider. Other forms of mediation may include or exclude other channels, such as video-conferencing systems or chat programs on the computer.

G. SCENARIO

- G.1 SPEAKER CHARACTERISTICS
- G.2 CHANNEL CHARACTERISTICS
- G.3 OTHER ENVIRONMENT CONDITIONS

By *scenario* we mean the various practical conditions and attendant circumstances which affected the collection of the dialogue data. Such conditions are important to keep track of, since they might have an effect (foreseen or unforeseen) on the value of the corpus as a basis for further research and development.
Speaker characteristics are often stored in a speaker database, and include information on how speakers were sampled; the age and gender of each speaker, the speakers' native language, their geographical provenance, their drinking and smoking habits (see Gibbon et al. 1997, pp. 110ff.); whether speakers are known to one another; whether speakers are practised in the dialogue activity. Speaker characteristics also include (a) what language(s) was/were spoken, and (b) what the native language of each speaker is.
Channel characteristics include use of the spoken versus written medium; recording characteristics (e.g. whether multi-channel recording was used); use or non-use of a telephone line; availability of visual channel; recording in studio vs. recording on location; and so on.
Other environment conditions include not only general contextual factors, but also special design features used in the collection of data and affecting the nature of the outcome: e.g. a signal button was used in some VERBMOBIL recordings

to request a turn, thereby eliminating turn overlaps and allowing speakers to formulate their ideas before speaking. Another dialogue manipulation strategy is the Wizard of Oz (WOZ) scenario mentioned above (under F.1.1).

1.3 General coding issues

We will shortly turn to the examination and recommendation of representation and annotation practices at the specific levels listed towards the end of 1.1.3 above. But first, we should give attention to general coding issues which affect all these levels. Perhaps the overriding issue is whether all levels should follow the same general encoding standards. There is much to be said for adhering to existing or emerging standardisation initiatives, where scientific considerations permit, since this would make information exchange or display much easier and reduce the need for (re)-writing individual tools for each application. The best candidates to consider are the SGML-based TEI standardisation initiative and the more recent emergence of the XML conventions. In principle, they could apply to all levels of transcription and annotation. However, it is necessary to avoid being dogmatic on this issue. In the following sections, we discuss and exemplify TEI mark-up where appropriate, but at the same time we illustrate other forms of encoding where the data we are illustrating happen to be in these alternative forms. For future projects, we recommend that as much use as possible should be made of standardised encoding schemes such as those of the TEI, extending them or departing from them only where necessary for specific purposes.

Another issue is the degree to which different levels of transcription or annotation make use of information provided by other levels. Here again, it would be premature to insist on too great a degree of conformity. Let us consider briefly the requirement of segmentation or 'chunking' at various levels. The orthographic transcription (1.4) will divide the dialogue up in the first instance into *turns*, within which further units will typically be signalled, where necessary, by the use of full stops or other punctuation marks. The *'orthographic sentence'*, if indicated at this level, may be regarded as a pre-theoretical unit, arrived at more or less impressionistically by the transcriber, who may not have the expertise to make use of prosodic or other levels of information. At the syntactic level, a similar unit (termed a *C-unit* in 1.6) may be recognised, but may not correspond one-to-one with the 'orthographic sentence' of the basic transcription. Equally, at the prosodic (1.7) and pragmatic (1.8) levels, segmentation may lead to the delimitation of *tone groups* or *utterances* which are important at those levels. Whereas in the longer run we may anticipate more integration of these units at different levels of analysis, it would be better at this stage to regard them as independent though correlated. The degree to which one level of annotation depends on another rests on factors such as the ordering of the procedures of annotation and the kinds of expertise the transcribers or annotators make use of. For purposes of implementation, however, segmentation at the orthographic, syntactic and/or prosodic levels may be seen as subservient to the task of isolating key pragmatic dialogue-units representing the communicative goals of the participants.

1.4 Orthography

1.4.1 Orthographic representation

The aim here is to represent the macro-features of the dialogue, including a verbatim record of what was said. A 'verbatim record' is a useful abstraction for many purposes, but it must naturally not be confused with the speech event itself. Some kind of hierarchy of priority seems to be needed in what kinds of macro-features of the dialogue to represent orthographically, and at what level of detail to represent them: see the recommendations at the end of this section.[6]

1.4.1.1 Background

This section takes account of the recommendations made by Llisterri (1996), by Gibbon et al. (1997) within the EAGLES framework, and of those made by Johansson et al. (1991) for the TEI, now largely codified in TEI P3 (Sperberg-McQueen and Burnard 1994). The corpus survey on which the following discussion is based comes partly from the document of Johansson et al. (1991) and partly from a fresh extension of it, which pays particular reference both to corpora produced for dialogue projects and to corpora in European languages other than English.

We try in particular to address the issue of integrating spoken and written resources – e.g., making representations of spoken corpora accessible to the language engineering (not just the speech technology) community. For this reason, we sometimes focus on *processibility* of texts (e.g., by stochastic or rule-based taggers and parsers) as an issue.

There is, at present, no strong consensus as to the means of representation, so that, for example, whilst we may use examples based on the TEI, we do not assume the necessity of TEI conformance. Rather, we concentrate on the *features* that should be represented. However, some forms of representation naturally capture certain phenomena more easily than others: for instance, the start and end tags used in SGML/TEI are particularly useful for indicating the duration of a speech-simultaneous phenomenon such as a non-verbal noise. It might also be noted that, in choosing a representation scheme, individual symbols that could be confused with other markup should perhaps be avoided: for example, the `@` character used by VERBMOBIL to mark overlapping speech could possibly be confused with the SAMPA representation of the schwa character. The use of tags with whole-word representations (e.g., the Spanish <`simultáneo`>) would minimise this kind of confusion. However, with multi-layered 'stand-off' annotation that separates the annotated material from the actual annotation (cf. Thompson 1997), this would be less of an issue. The labels for the various tags can be standardised for any given language, but it is not necessary that a single specified language be adopted as a universal 'metalanguage': tools may be developed to translate between different language versions, where this is necessary for processing (e.g., in multilingual research).

[6] There is a large literature on both practice and principle in the transcribing and coding of spoken language data. Particularly relevant to this section are the transcription conventions for SPEECHDAT corpora in Gibbon et al. (1997), pp. 824–828. Two collections of studies of transcription more from the point of view of general linguistics and discourse analysis are those of Edwards and Lampert (1993) and Leech et al. (1995).

The issue of obligatory vs. recommended vs. optional levels (cf. the recommendations on morphosyntax in Leech and Wilson 1994) is one that should also be addressed. Obviously, some applications will require more detailed transcription and analysis than others.

1.4.1.2 Documentation on texts

There are three primary ways of documenting information about texts:

1. in a separate set of documentation – e.g., a manual
2. in a header within the text itself, which may be
   1. structured – e.g., a TEI header
   2. relatively unstructured – e.g., a few lines of COCOA[7]references.
3. in separate documentation files with links (pointers) into the text. Those files may contain
   - pointers into a text, such as a transcription
   - pointers into a soundfile
   - a speaker database
   - etc.

Amongst the corpus linguistics community, a header has for some time been considered the minimum requirement for text documentation. An in-text header – as opposed to external documentation – makes it less easy to confuse texts: it can be used as part of an automatic analysis, to output background information; and it enables quick reference, especially when a manual is for some reason not to hand. On the other hand, in-text headers make for redundancy, if the same information has to be repeated in the head of each text using the same transcription scheme. This redundancy can be avoided by including in the header a reference or (better) a link to external documentation, in the form of a manual. Alternatively, data and documentation files may be integrated within a database management system (DBMS; see Chapter 5).
Whether a header or external documentation is used, as a bare minimum it should normally contain an *identifier* for the specific text and basic *information on the speakers*. We recommend that additional information should include:

1. Speaker characteristics
   - number of participants
   - individual speaker attributes – e.g., age; sex; social class; native languages; regional accents
3. Channel characteristics
   - use of telephone line or other channels
   - recording details – e.g., time and date; technical specifications
3. General environmental conditions
   - contextual information – e.g., where the dialogue took place; under what physical conditions

[7]COCOA was an early computer concordance program used for extracting indexes of words in context from machine readable texts, whose conventions were used by several corpus annotation projects. - Ed.

- human or machine or simulated (Wizard of Oz)
- etc.

4. Other information
   - activity type (see Section 1.2 above)
   - degree of spontaneity
   - matters under discussion (domain/task)
   - details of the orthographic transcription
   - details of levels of linguistic annotation
   - contact details for obtaining additional information, for reporting difficulties or errors, etc.

The speech community, especially according to the decisions agreed on during the SAM-project, favours external files which can be distinguished via different extensions and are linked together via pointers (cf. Gibbon et al. 1997, pp. 732 ff.). There are good practical reasons for separating a speech file (containing waveforms only) from associated descriptive files, though this very much depends on the developer's working environment.

1.4.1.3 Basic text units

The most common text units in dialogue corpora are the *text* (i.e., a self-contained dialogue or dialogue sample with a natural or editorially created beginning and end) and the *turn* (or contribution). Tone groups are also sometimes marked. 'Orthographic sentences' (that is, units delimited by conventional written punctuation) are also often present (see 1.4.1.7.2), but these should probably be viewed as artefacts of transcription, rather than as real observable units *per se*.
We suggest that the *text* and the *turn* should be the basic text units in orthographic transcription, together with the intuitively-identified 'orthographic sentence'. There is no reason to include tone groups in orthographic transcription, as these are difficult to identify reliably (see Knowles 1991): any marking of tone groups belongs to the interpretative stage of prosodic markup (Llisterri's S3 level (Llisterri 1996)). Similarly, there is no reason to include *utterances*, whose identification belongs rather to the level of dialogue act annotation (see 1.8). The notion of turn is itself not wholly unproblematic, since interruptions and overlaps can occur, but there are methods for representing these aspects (see, for example, 1.4.1.6 below). As noted, 'orthographic sentences' are often used in transcription for greater intelligibility and processibility (e.g., by taggers that assume the sentence as the basic processing unit), but it should be emphasised that the turn is a basic unit of *spoken dialogue* transcription, and that the 'orthographic sentence', delimited by turn boundaries and/or sentence-final punctuation, is, as a unit of *written* language, merely a convenient impressionistic unit providing useful preliminary heuristic input to other levels of annotation.

1.4.1.4 Reference system

A reference system – i.e., a set of codes that allow reference to be made to specific texts and locations in texts – may be absent from transcribed spoken corpora. This is partly due to the fact that multiple versions of spoken corpora

often exist, with a basic transcription being stored as one file and a time-aligned version being stored as a different file. A time-aligned file has, in essence, already a reference system, in that the time points can be used to refer to specific locations in the dialogue. Nevertheless, it is both useful and straightforward to introduce a basic reference system into ordinary orthographic transcriptions also. The references may be encoded either as a separate field, as in the TRAINS corpora:

```
58.3  : load the tanker
58.4  : then go back
```

or merged with speaker codes as in VERBMOBIL:

```
TIS019: gut , bin mit einverstanden , dann ist das klar .
HAH020: danke sch"on <A> .
```

1.4.1.5 Speaker attribution

Speaker attribution is most often indicated by a letter code at the left-hand margin, but may sometimes be inferred from the turn, especially if there are only two participants in the dialogue. The code may or may not be enclosed in some kind of markup delimiting notation. Also, a speaker's turn may or may not be closed by an end tag. Sometimes, the code may be longer than a single letter; in VERBMOBIL, it also includes digits to indicate the turn number – see 1.4.1.4 above. Some examples are:

From TRAINS:

```
57.1 M: puts the OJs  in the   tanker
58.1 S:       +southern route+
```

Based on the TEI Recommendations:

```
<u who=A> Have you heard that she is back?</u>
<u who=B> No.</u>
```

From CREA[8]

```
<u who="anat00001.PER002" trans="smooth">Ha llamado.</u>
<u who="anat00001.PER001" trans="smooth">No, la hemos llamado
nosotros.</u>
<u who="anat00001.PER002" trans="smooth">Bueno.</u>
```

The speaker identification codes used, such as
`<u who="anat00001.PER002" ...>`,
relate to information already given in the text header or accompanying documentation.

[8]CREA is the Corpus de Referencia del Español Actual, a 10-million-word corpus containing a million words of transcribed speech compiled at the Real Academia Española. The corpus is SGML encoded and follows closely the conventions of the TEI and CES (Corpus Encoding Standard: see Ide et al. (1996)). Further information can be obtained from "`joaquim.llisterri@cervantes.es or mpino@crea.rae.es`".

Cases where there is more than one speaker, or where the transcriber is unsure who is speaking, are normally explicitly indicated. The TEI, for instance, recommends the following practices:

- for uncertainty:
  `<u who=A1 uncertain=medium>`
  where `uncertain` can take various values such as a comment on the degree or cause of uncertainty.
- for multiple speakers:
  `<u who='A1 B1 C1'>`
- for unknown speakers:
  `<u who=unknown>`

The same features can be marked with slightly different conventions in non-TEI markup schemes.

1.4.1.6 Speaker overlap

Speaker overlap, i.e., synchronous speech by more than one participant in the dialogue, is one of the most important issues in dialogue transcription. An examination of existing corpora demonstrates that the most common method of indicating overlapping speech is by '*bracketing*' the relevant segments of both interlocutors' speech, although the choice of bracketing characters varies considerably (e.g., @ preceded or followed by an overlap identifier number in VERBMOBIL, plus signs in TRAINS, SGML *tags* in the Corpus of Spoken Contemporary Spanish (Marcos-Marín et al. 1993) – hereafter 'CSCS'). Sometimes, the speech of only one of the two or more overlapping interlocutors is bracketed, although this is potentially less clear than the marking of *all* overlapping speech.
Three other methods of handling overlap may also be encountered:

1. Vertical alignment, as in a musical score, of overlapping segments (widely used in conversation analysis and sociolinguistic transcription).
2. Reorganisation of overlaps into separate turns, without representing where overlaps occur (as used, for example, in the *Czech national corpus*).
3. The TEI practice of using time pointers, for example:

```
<timeLine>
     <when id=P1 synch='A1 B1 C1'>
     <when id=P2 synch='A2 C2'>
</timeLine>
     ...
<u who=A>this is <anchor id=A1> my <anchor id=A2> turn</u>
<u who=B id=B1>balderdash</u>
<u who=C id=C1> no <anchor id=C2> it's mine</u></u>
```

The first method is technically problematic, as it often does not delimit with markup the precise stretches of speech that overlap: often only the start of an overlap is marked. Thus this information can easily be lost, especially when different display or print fonts are used that alter the visible alignment. The second is an idealisation: it obliterates evidence of overlap in favour of neat, drama-like turns. The third (TEI) option has the advantage of dealing very well with multiple overlaps: e.g. where three speakers are talking simultaneously,

and cross-bracketing would otherwise occur. For most purposes, it is perhaps a little too cumbersome in comparison with bracketing; however, a multi-layered approach to transcription and annotation – e.g., Thompson's suggestions using eXtensible Markup Language (XML) (Thompson 1997) – can make it far less cumbersome for human users. Unpublished work by Steven Bird and Mark Liberman on *annotation graphs* proposes a method for consistently formalising multi-level (multi-tier) annotations; see:
"`http://morph.ldc.upenn.edu/annotation/`".
Occasionally, overlap bracketing crosses turns. In the CSCS, for example, a single overlap tag encloses the stretch of overlapping speech across speaker boundaries:

```
<H1> <simultáneo>Sí, sí.
<H2> ...había</simultáneo> sido mucho más compleja la posición
```

This is, however, perhaps less clear than if the overlap markup were nested within the turns, thus:

```
<H1> <simultáneo>Sí, sí.</simultáneo>
<H2> <simultáneo> ...había</simultáneo> sido mucho más
compleja la posición
```

CREA uses `<overlap> ...</overlap>` tags, as has already been seen in the preceding section.

1.4.1.7 Word form

Most corpora transcribe speech using the standard (or dictionary) forms of words, regardless of their actual pronunciation. The use of standard word forms has a huge advantage, in that annotation and retrieval tools, for example, may be applied relatively unproblematically to speech as well as to writing.
Furthermore, everything (including numbers) is typically written out in full. Thus it is important to distinguish different ways of saying the same numeral: in German *2* may be pronounced as either *zwei* or *zwo*. Similarly, in English there are different ways of saying the same string of numerals: *1980* can be said as 'nineteen eighty' (the year) or as 'one nine eight oh' (a telephone number) or as 'one thousand nine hundred and eighty' (an ordinary number). Units of time, currency, percentages, degrees, and so on are normally transcribed in full to capture their pronunciations – e.g., *two hundred dollars and fifty cents* rather than *$200.50*; or *ten to twelve* rather than *11.50*. However, in some cases, it may be more straightforward to transcribe numbers simply in arabic numerals: for example, in a restricted domain such as airline travel dialogues, the majority of numerical expressions may be flight numbers, which will conform to a uniform system of pronunciation. A further argument in favour of the more 'simplified' form of transcription (e.g., *$200.50*) is that the actual pronunciation may be represented at another (phonemic) level, if a multi-layered form of transcription and annotation is employed.
Common contractions and merges that are also encountered in written texts (e.g., *can't, gonna*) are usually allowed, but otherwise dictionary forms are used, with special pronunciations indicated instead by editorial comments (see 1.4.1.13 below). In projects such as the BNC, a supplementary list was drawn

up of those common allowable contractions, etc., that were not included in a standard dictionary. Spelling of interjections (e.g. the choice in English between *okay* and *O.K.*) can also be a problem: see Section 1.4.1.8.2 below. In practice, all lexical items that appear in a corpus should also appear in a lexicon, be it either an external, pre-existing standard dictionary or a lexicon specially generated from the corpus.

In some languages, compound words are also an issue for transcription. This is a problem even for languages such as German which have fairly regular orthographic conventions for representing compound words as single words in writing, and it is a problem for languages such as English, which, historically, have a more flexible approach to the representation of word compounding. For instance, in English, one may find *keyring*, *key ring* or *key-ring*. It would be difficult, if not impossible, to lay down strict rules for the representation of compound words. The key essentials for a good transcription system, therefore, are *internal consistency* of practice in representing compound words and *explicit documentation* of the practice adopted. If compound words *are* represented as multi-word units, it is possible to tag them as compound words at the morphosyntactic level (see 1.5.1.4).

Pseudo-phonetic/modified orthographic transcription tends to be reserved for oddities such as non-words or neologisms that have no true dictionary form. Letters of the alphabet that are pronounced individually are normally demarcated by spaces, to distinguish, for example, the two different pronunciations of *VIP* — /vɪp/ vs. /vi: aɪ pi:/. In CREA, the tag <distinct> is used for spelled-out words, with the attribute 'dele' (for 'deletreado'):

```
<distinct type='dele'>pe-e-erre-erre-o uve-e-erre-de-e</distinct>
```

(Here the speaker spells out the two words *perro verde*.) It is probably sufficient to separate these with spaces (e.g. `V I P`), but sometimes additional markup is encountered, as in VERBMOBIL: `$V $I $P`.

It has been suggested that a standard dictionary should be employed for each language as an arbiter, wherever needed, for these dictionary forms. The Duden has already been used in this way for German in VERBMOBIL, and the dictionary of the Real Academia Española has similarly been used for CREA. However, this may be a little too idealistic. Often, dictionaries present more than one possible spelling of a word – e.g. *analyse* vs. *analyze*. Also, it is difficult to conceive of transcribers checking spellings in a standard dictionary when they feel confident of how to spell something. It may be that a style guide, such as Hart's *Rules* for English (Hart 1978), would help with restricting common variant spellings. For languages with less spelling variation and/or one standard 'academy' dictionary, the situation could be more straightforward. Where available, a better alternative would be to use special dictionaries that have already been developed during projects in the speech community. These tend to be based on experience and actual requirements for systems, and normally take into account all the problems encountered during system development.

For example, to reduce error rates in testing and training signal recognition systems based on a particular language model, frequently occurring assimilations between individual words have to be integrated into the dictionary in addition

to the canonical orthography, because the system has to read and understand the transcriber's representation of the utterance, e.g. in German the spoken form *hamwanich* vs. the written form *haben wir nicht*.

1.4.1.7.1 Word fragments

Word fragments, also known as unfinished or truncated words, are typically transcribed as follows: as much of the word as is pronounced is transcribed, followed by a 'break-off' character – for instance a dash or an asterisk. Sometimes a tag is used instead of a special character, for example, `<distinct type='titu'>` (for Spanish 'titubeo') in CREA. For example:

```
<distinct type='titu'>es*</distinct> estamos
```

In this case, an asterisk (*) is added to the end of the incomplete word.
Some guidelines (e.g., the Gothenburg corpus of spoken Swedish) also allow for word-final fragments, in which case the 'word fragment' character may occur at the beginning rather than the end of a string. Most transcriptions of word fragments use standard or modified orthography, but this can be confusing in cases like the English digraph *po–*, which may represent either the diphthong of *poll* or the simple vowel of *pot*. It may thus be better to use some form of phonetic representation, such as SAMPA, for word fragments; however, if there is a further level of phonemic transcription, then this is unnecessary.
An interesting aspect of the guidelines used by the TRAINS project is that an interpretation (or expansion to full form) of word fragments is added where possible. This has both advantages and disadvantages. Where a fragment is not part of a repeated sequence that includes a full form, it enables more content to be extracted for language understanding and so on, but, on the other hand, it may be argued that to interpret such fragments – even when they seem unambiguous – is to read additional (and perhaps unwarranted) information into the transcript beyond what needs to be represented. Such interpretative information should preferably not appear at the level of orthographic transcription. Furthermore, word fragments may also at times serve a communicative function, indicating that the speaker has changed his/her mind about what to say next or how to interpret something, and expanding them may thus lead to misinterpretation.

1.4.1.7.2 Orthography, including punctuation

As to the more general form of transcription, the use of a basic canonical subset of the standard orthography is both normal and desirable. Sentence-initial capitals may be omitted, but, otherwise, normal capitalisation and at least full stops tend to be used. This improves readability for the human user and improves processibility for taggers, parsers, and so on. Obviously, it is understood that such standard orthography is, to a considerable extent, interpretative when applied to speech, but its advantages outweigh its disadvantages. The use of punctuation characters other than full stops is an open question, but commas may sometimes have certain advantages as well. In English, for example, using a comma before a tag question is unambiguous and may actually help to identify

the purpose of this particular phrase type as communicating a possible request for feedback: e.g. *Two o'clock, is it*. There is also a case for using question marks where the transcriber clearly perceives an utterance as a question. This can be useful especially where the structure of the utterance does not mark it as interrogative. There are many questions which lack such marking (e.g. *Next week?*), and their import is not clear to a reader who does not have access to the prosodic level of annotation.
Whatever punctuation scheme is adopted, the general rule must be to explain it in the text documentation, e.g. in the header. For example, if punctuation has been used, it should be explicitly stated which punctuation marks have been employed, and how they have been assigned (whether impressionistically or otherwise).

1.4.1.7.3 Unintelligible speech

It is sometimes impossible to decipher – at least in part – what a participant is saying, because of unclarity in the recording. Normally a single code is used – e.g, <inintelligible> in the CSCS or <%>, added directly to the word, in VERBMOBIL. Sometimes a form of bracketing is employed instead, with the number of unintelligible syllables given. An estimate of the number of unintelligible syllables is desirable, but it is emphasised that this estimate can only be approximate.

1.4.1.7.4 Uncertain transcription

In other cases, the transcriber can hazard a guess as to what was said, but wishes to indicate the existence of uncertainty. Normally, such uncertain transcriptions are bracketed in some way, but with conventions different from those used for truly unintelligible speech. Here are two examples of ways of marking uncertain transcriptions:

- Uncertain *syllables or sounds*: in the CSCS, these are bracketed within the word, thus: `burri<(t)>o`.
- Uncertain *words and phrases*: in the TEI, these are placed inside a set of start and end tags, e.g., `<unclear> burrito </unclear>`. The TEI tag shown here also has an optional attribute `reason`.

1.4.1.7.5 Substitutions

Also to be considered under this heading are those cases where words – normally proper nouns – are to be replaced for confidentiality or other reasons. These may be marked with codes, since this makes it more clear where an original text word has been replaced. The practice of simply substituting an alternative name without comment is sometimes encountered, but should perhaps be avoided; however, a replacement could be used if it is commented, for example, by the use of a TEI regularisation tag:

```
<reg>Bert</reg>
```

Obviously, in circumstances of confidentiality, the `orig` attribute, which normally encodes the original form of words, cannot be used.

1.4.1.8 Speech management

By *speech management* we understand the use of phenomena such as quasi-lexical vocalisations, pauses, repairs, restarts, and so on.
Although speech management is normally an issue for transcription, it should be noted that sometimes phenomena included under this heading are instead *annotated* at a separate level of processing – cf. the so-called *dysfluency annotation* of the Switchboard corpus in the Penn Treebank project.[9]

1.4.1.8.1 Pauses

Unfilled pauses (by which we mean *perceived* pauses, rather than silence in the speech signal) are typically marked with suspense dots (. . .) or some other special punctuation such as an oblique slash. It is important to distinguish short pauses from longer pauses or silences, which may indicate an interruption by some non-conversational event, activity, etc. The Gothenburg Swedish corpus uses various numbers of slashes (/, //, or ///) to give an impression of the length of a pause. Sometimes a tag is used instead of punctuation – e.g., <P> in VERBMOBIL. Both methods may allow additional comments to be added as to the length of a pause.

1.4.1.8.2 Quasi-lexical vocalisations

Most corpora make some attempt to standardise the transcription of quasi-lexical vocalisations, such as interjections and filled pauses such as *um, uh-huh, oi, ooh* and *ah*. In contrast, the CSCS avoids the use of invented/idealised word forms and instead uses markup to indicate where quasi-lexical vocalisations occur. The down side of this, however, is that such features can confuse transcription with dialogue-act annotation: they require an interpretation of the function of a vocalisation (e.g., agreement, negation). A possible third way, which is mentioned as an option by the TEI guidelines, would be to merge the two systems, so that quasi-lexical vocalisations have standardised forms but occur in the form of markup to indicate that standardisation has occurred. For example:

```
<vocal type=quasi-lexical desc=uh-huh>
```

However, this approach may be found to be too verbose and cumbersome. It may be better simply to use a standard list of orthographic forms for these phenomena, without any additional markup, and this approach is also sanctioned by the TEI. Whichever approach is adopted, it is useful to draw up a standardised and generally acceptable list of these quasi-lexical forms for each language, so that unwanted variants do not proliferate, causing retrieval problems.

[9] For an example, see "`http://www.cis.upenn.edu/~treebank/switch-samp-dfl.html`"; for the manual, see "`ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps`".

1.4.1.8.3 Other phenomena

Many corpora do not explicitly identify repetitions, repairs, etc. However, for the purpose of activities such as part-of-speech tagging or speech recognition (cf. Section 1.7.1.6), it may be important to do so, so that, for example, repetitions can be taken into account when developing a dialogue model and training of dialogue category transition models. If repetitions and so on are identified in the transcription, it is probably desirable that one full-word transcription should be retained in the main running text and the rest marked up with some kind of bracketing. The TEI's `<del>` tag is one possible way of representing this and allows the various types of phenomena to be noted:[10]

```
<del type=truncation>s</del>see
<del type=repetition>you you</del>you know
<del type=falseStart>it's</del>he's crazy
```

1.4.1.9 Paralinguistic features

By 'paralinguistic features' we mean concomitant aspects of voice such as superimposed laughter, tempo, loudness,. We exclude features that do not accompany speech but rather occur in isolation (e.g., laughter not superimposed on speech), for which see 1.4.1.10 below.
Paralinguistic features tend to be encoded with a finite set of standard features, but sometimes also free comment is allowed. A standard list of codes will enable features to be retrieved and counted with concordancing software. Unconstrained comment tags should be avoided as much as possible. The TEI has already produced a basic list of paralinguistic features, which can be used or amended for LE purposes; these are reproduced in Appendix A of this document.
The use of balanced start and end tags will enable the duration of a paralinguistic phenomenon to be encoded more clearly.

1.4.1.10 Non-verbal sounds

Non-verbal sounds are typically transcribed as a form of comment. Sometimes, a standard set of codes is defined in place of free comment.[11] However, it may be advisable for at least one more general feature to be retained (e.g., `noise`), to allow for unattributable sounds or those for some reason omitted from the standard list. It is possible, following the practice of the CSCS, to combine standard features and free comment, so that additional information is available as well as a basic indication of broadly what kind of noise has occurred.
Minimally, four types of non-verbal sound might be differentiated:

1. non-verbal but *vocal* utterances attributable to the speaker (e.g., a laugh, or audible intake of breath)
2. non-verbal but *vocal* utterances not attributable to the speaker (e.g., an unattributed grunt)

[10] But see 1.4.1.7.1 above for a preferred method of transcribing truncations (phonetic representation rather than orthographic characters).

[11] Good starting points for a typology of non-verbal noises would be the two noise databases, 'Noise-ROM-0' and 'Noisex' (see Gibbon et al. 1997, p. 8)

3. non-vocal noises attributable to the speaker (e.g., snapping fingers)
4. non-vocal noises not attributable to the speaker, including noises that are not humanly produced (e.g., a dog barking, a doorbell ringing)
5. technical noises, e.g. microphone noise, click.

Again, as with paralinguistic features, the use of start and end tags allows a continuous noise to be represented.

1.4.1.11 Kinesic features

Kinesic features comprise what is, in informal speech, termed 'body language' – e.g., eye contact, gesture, and other bodily movements. Few corpora represent these features, since corpus transcription is typically from audio rather than from video data or a live performance. In the past, kinesic features have been of less relevance to natural language and speech research than have the other features discussed in this document; however, as work on audio-visual speech synthesis progresses, they are likely to become much more relevant (see Chapter 2). But, since these have been investigated by the Multimodal Working Group of EAGLES, guidelines on such features belong to another chapter. We may note, however, that in an auditory transcription they can be included as editorial comments or using the TEI's <kinesic> tag, which has attributes to indicate the 'actor', a description of the action, and whether or not it is a repeated action.

1.4.1.12 Situational features

Basic information about the context of a dialogue (e.g., the participants, location, etc.) tends to be included in the text header or equivalent descriptive documentation (see Section 1.4.1.2). More 'short-term' information, such as the arrival or departure of a participant, is normally introduced as editorial comment. For these features the TEI suggests a special comment tag (<event>) , with the same attribute set as <kinesic>.

1.4.1.13 Editorial comment

Editorial comment comprises a number of cases where interpretative information needs to be added over and above the transcription of the phenomena described above. These cover several types, discussed below.

1.4.1.13.1 Alternative transcriptions

Pseudo-phonetic or modified orthographic transcription should be avoided as a general rule, and canonical (lexical) orthography should be preferred, reserving variants for explicit markup in cases where it may be desirable to indicate, separately from a full phonetic/phonemic transcription, how a word or phrase was pronounced, for example, because it is a dialect form or ('orthographic noise') a homograph. Modified orthography in the transcription itself may cause difficulty in concordancing or processing the text and may, in any case, be misleading – e.g., for non-native speakers using the corpus. An approach similar to that adopted by VERBMOBIL might be preferable, namely that alternative transcriptions should be enclosed within markup

brackets. A similar approach is recommended by the TEI using the <reg> tag:

```
<reg sic='booer'>butter</reg>
```

If more than one standard orthographic word is included in a variant pronunciation, VERBMOBIL also adds a number indicating how many of the standardly transcribed words are represented by a given pronunication. This feature is not part of the TEI syntax for <reg>, but might be an optional addition. It would be less important in a TEI representation than in VERBMOBIL, since VERBMOBIL does not use start and end tags to bracket the stretch of speech. If using a number, *whatcha* in English, for example, might be represented with something like:

```
<reg words=3 orig='whatcha'>what are you</reg>
```

The SAMPA conventions for encoding phonetic (IPA) transcriptions in 7-bit ASCII permit the representation of alternative pronunciations in SAMPA format rather than in an idiosyncratic modified orthography:

```
<reg orig='bU?@'>butter</reg>
```

The 7-bit character set is still the most dependable encoding format for information interchange, and it is advisable, for the time being, to stick with SAMPA rather than attempting to use other forms of encoding such as Unicode. It is strongly recommended that proprietary encoding (e.g. with fonts specific to an operating system or a word processor) should be avoided, as these require dedicated and highly specialised conversion software for interchange, and are therefore not freely exchangeable.

1.4.1.13.2 General comments

General in-text comments are typically introduced within some form of distinctive bracketing. In addition to the comment itself, the Gothenburg corpus of spoken Swedish encloses the stretch of text to which the comment refers. Comments in this scheme can also be numbered. We feel that enclosing the text commented on does make the comments more transparent. Numbers are probably not essential (in the Gothenburg corpus, comments occur on a different line from transcribed text, which is why they are used there). In an SGML (but non-TEI conformant) representation, this would look something like the following:

```
That is what <note comment="Which one?">Geoff</note> said.
```

1.4.2 Recommendations

We conclude with recommendations regarding the priority of information to be included in the orthographic representation of a dialogue. We provide for three levels of priority: 'Highest priority recommendations', 'Strong recommendations' and 'Recommendations'. These lists are by no means exhaustive, and

features may be added to them, or moved from one list to another, according to the needs of this or that project.

1.4.2.0.3 Highest priority recommendations

- Text header (or equivalent documentation) with text identification and identification of speakers (1.4.1.2)
- Text header documentation to include information on (a) speaker characteristics, (b) channel characteristics and (c) environmental conditions, as recommended at the end of Section 1.2 (1.4.1.2)
- Dialogue divided into turns (1.4.1.3)
- Speaker of each turn made explicit (1.4.1.5)
- Standard (canonical) spellings used wherever possible (deviations from standard spelling practices to be justified and documented in the markup) (1.4.1.7)
- Numbers, currency expressions, dates, clock times etc. written out in full (except where there is no risk of ambiguity, and where there are overriding reasons for economy) (1.4.1.7)
- Spelled-out letters (e.g. V I P) to be separated by spaces (1.4.1.7)
- Normal use of capitalisation and full stops (but sentence initial capitals optional) – avoid use of abbreviatory stops (1.4.1.7.2)
- Overlapping speech in indexed brackets or tag pairs, these to be closed within turns (1.4.1.6)
- Word fragments marked, with use of phonetic representation where needed (1.4.1.7.1)
- Quasi-lexical vocalisations transcribed using standard representations (1.4.1.8.2)
- Unintelligible speech tagged as such (1.4.1.7.3)
- Uncertain transcriptions tagged as such (1.4.1.7.4)

1.4.2.0.4 Strong recommendations

- Text header (or an independent but linked document) to specify transcription conventions (1.4.1.2)
- Pauses tagged, and long and short pauses distinguished (1.4.1.8.1)
- Repetitions and false starts tagged (1.4.1.8.3)
- Paralinguistic features tagged using standard list of features (1.4.1.9)
- Non-verbal sounds tagged using standard list of features (1.4.1.10)

1.4.2.0.5 Recommendations

- Comments tagged (1.4.1.13.2)
- Pause lengths marked (1.4.1.8.1)
- Alternative pronunciations tagged and represented with SAMPA (1.4.1.13.1)
- Kinesic features tagged (1.4.1.11)
- Punctuation other than full stops (usage to be explained in header) (1.4.1.7.2)
- Short-term situational features to be tagged in-text where appropriate (1.4.1.12)

1.5 Morphosyntax

1.5.1 Morphosyntactic (POS) annotation

Morphosyntactic annotation is also known as word-class tagging, POS (part-of-speech) tagging, or grammatical word tagging. It takes the form of associating a word-class label with each word token in a corpus. The set of *tags* used for labelling words in a particular language and in a particular corpus is known as a *tagset*. The list of tags, together with their definitions and the guidelines needed to map them on to a corpus, is known as a *tagging scheme*.

Previous work on morphosyntactic annotation within the EAGLES framework has primarily focussed on written language corpora and their relation to lexica (see Appendix E, p. 89). Although in practice only a few European languages have been exemplified, in intention the framework adopted has been multilingual and both language and application indepedendent. A number of EAGLES or EAGLES related documents are relevant. Leech and Wilson (1994) provide a set of preliminary recommendations for the morphosyntactic tagging of corpora; exemplary tagsets are provided for Italian and for English. This document has been closely coordinated with work on another document, Monachini and Calzolari (1996), which proposes a set of morphosyntactic guidelines for both lexica and corpora, and which exemplifies tagsets in some detail for Dutch, English, Italian and Spanish. Three documents which provide draft morphosyntax guidelines for Italian, English and German respectively are Monachini (1995), Teufel (1996) and Teufel and Stöckert (1996). Of these, the German scheme (Teufel and Stöckert) is worked out in considerable detail.

Morphosyntactic information can typically be represented as a type hierarchy, with features and their values. The major POS (part of speech) feature has values such as noun, verb, adjective, pronoun, adverb and interjection. More peripheral word categories are included under the values 'unique/unassigned' (e.g. infinitive and negative markers) and 'residual' (e.g. formulae, foreign words). Each of these values (except 'interjection', which tends to be undifferentiated) is then represented as a hierarchy table within which subcategories are shown as subsidiary features and values. For example, for nouns, the following features and values may commonly occur: Type (common, proper); Number (singular, plural); Case (nominative, genitive, dative, etc.); Gender (feminine, masculine, etc.). The range of features and values can obviously vary from one language to another, as can their hierarachical dependencies. But it is proposed that the morphosyntactic inventory for each language should be mappable into an *intermediate tagset* (Leech and Wilson 1994, Section 4.3), which shows what is common between languages, while enabling the differences to be captured by optional extensions and omissions.

The actual formal representation or encoding adopted for morphosyntactic annotation can vary from one tagging scheme to another. One proposal for tagging within the SGML-based TEI guidelines is found in the CDIF implementation for the BNC (Burnard (1995); Garside et al. (1997), pp. 19–33). Another, known as CES has been put forward for implementation as a general EAGLES standard by Ide et al. (1996), Section 5.2. The follow example illustrates the SGML-based CDIF tagging scheme for the BNC:

```
<w AV0>Even <w AT0>the <w AJ0>old <w NN2>women
<w VVB>manage <w AT0>a <w AJ0>slow <w UNC>Buenas
<c PUN>,<w AV0>just <w CJS>as <w PNP>they<w VBB>'re
<w VVG>passing <w PNP>you<c PUN>.</PUN>
```

In this model, the primary textual data and the annotations are combined in a single file, the annotations being encoded as SGML tags. However, in the Corpus Encoding Standard (CES) model of Ide et al. (1996), preference is given to the mechanism of placing annotations in a separate file, with its own document type definition (DTD). In this case, cross-reference between the text itself and the annotation document is achieved by using HyTime-based TEI addressing mechanisms for element linkage. In effect, the text document and the annotation documents associated with it are handled as a single hyper-document (Ide et al. 1996, Section 5.0).

Our particular concern here, however, is with the linguistic decisions involved in morphosyntactic annotation of dialogue. It could be argued that this is not a special problem area for dialogue corpora, since the same word-class categories are likely to appear in both spoken and written texts (even 'ums' and 'ers' occur in fictional dialogue, albeit in stylised versions). That there is no great difficulty here is suggested by the fact that the whole of the BNC, for example, has been tagged using the same tagset for the spoken data (ca. 10 million words) as for written texts (ca. 90 million words).

However, most tagsets have been devised primarily for written language, and the fact that the same tagset can be applied to spoken and written data should not lead us to ignore the fact that frequency and importance of word categories vary widely across the two varieties of data, or that at a given level of generality tags may be applicable to related but different categories. Interjections and hesitators (or filled pauses) (*um, er* etc.) are vastly more frequent in speech than in writing. There are, in fact, two aspects of morphosyntactic tagging which need to be considered in adapting a tagset from written to spoken language:

(a) Dysfluency phenomena:
   (i) How to tag pause fillers (*um, er,* etc.);
   (ii) How to tag word fragments (e.g. where a speaker is interrupted in mid-word).

(b) Word-classes which are characteristic of speech, but not of writing:
   (i) How to tag discourse markers etc.
   (ii) How to tag peripheral adverbials

1.5.1.1 Dysfluency phenomena in morphosyntactic annotation

There are two problems to consider under this heading. The first is how to tag hesitators, i.e. filled pauses such as *um* and *er* in English. The second is how to tag word fragments or fragments (see 1.4.1.7.1 above) which result from repairs and incomplete utterances. In the so-called intermediate tagset proposed in EAGLES preliminary guidelines (Leech and Wilson 1994), there is, as already noted, a 'catch-all' peripheral part-of-speech category U ('unique' or 'unassigned') that can be used for these quasi-lexical phenomena. The guidelines also allow for the subdivision of this category U into subcategories such as

Ux 'hesitator' and Uy 'word fragment' (where x and y are digits). It is highly recommendable that the morphosyntactic annotation of spoken language make use of such subcategories. Alternatively, the guidelines would allow the I (interjection) part-of-speech category to be subclassified to include hesitators (see (ii) below). Hence in this respect, although the existing morphosyntactic annotation guidelines are adequate, devising optional extensions such as the inclusion of new subcategories should be seriously considered.
On the other hand, an alternative solution is not to assign morphosyntactic tags to these items at all, but to mark them in the orthographic transcription as non-word vocalisations comparable to laughs and snorts (see 1.4.1.10 above). This solution is in tune with the proposal, discussed further in 1.6.1.1 below, to treat dysfluency phenomena as extraneous to the grammatical annotation of speech, on the assumption that they belong to a distinct level of dialogue control.

1.5.1.2 Word-classes which are characteristic of speech, but not of writing

Tagsets may need to be augmented to deal with spoken language phenomena such as discourse markers (*well, right*), pragmatic particles (*doch, ja*), and various kinds of adverbs (especially stance or modal adverbs and linking adverbs) which are strongly associated with the spoken language. Most of these forms might in a very general sense be termed 'adverbial' in that they are peripheral to the clause or sentence, are detachable from it, and may often occur in varying positions, particularly initial or final, in relation to any larger grammatical structures of which they are a part. They tend to have an important role in marking discourse functions and therefore in providing criteria for dialogue act classification (see Section 1.8 below).

1.5.1.2.1 Interjections in morphosyntactic annotation

The interjection POS category (I) is badly served in the current EAGLES documentation, since no subcategories are recommended. However, analysis of spoken language corpora reveals the high frequency of a number of rather clear subcategories which are also relatively distinct in their syntactic and discoursal distribution. It is suggested, therefore, that these might be distinguished by different tags, all beginning with the part-of-speech prefix I. Something like this proposal, put forward in two earlier articles (Stenström 1990; Altenberg 1990), was adopted by Sampson (1995), pp. 447–448, in his seminal discussion of the grammatical annotation of spoken English. His subcategory tags (which begin with U rather than I) include, in addition to familiar exclamatory interjections such as *oh* and *wow* (tagged UH), those shown in Table 1.1.

This list is simply presented here as an illustration, showing that the *interjection* category in spoken language is both broader and more finely structured than is allowed for in traditional grammar. This should not be worrying in that the Latin etymology of *interjection* suggests that it is something 'thrown between', in a sense that applies more or less happily to all the items above. They are grammatically 'stand-alone' items, capable of occurring on their own in a turn,

Table 1.1: Sampson's subcategories for interjections

UA	Apology	(e.g. *pardon, sorry, excuse_me*)
UB	Smooth-over	(e.g. *don't_worry, never_mind*)
UE	Engager	(e.g. *I_mean, mind_you, you_know*)
UG	Greeting	(e.g. *hi, hello, good_morning*)
UI	Initiator	(e.g. *anyway, however, now*)
UL	Response Elicitor	(e.g. *eh, what*)
UK	Attention Signal	(e.g. *hey, look*)
UN	Negative	(e.g. *no*)
UP	*please*	as discourse marker
UR	Response	(e.g. *fine, good, uhuh, OK, all_right*)
UT	Thanks	(e.g. *thanks, thank_you*)
UW	*well*	as discourse marker
UX	Expletive	(e.g. *damn, gosh, hell, good_heavens*)
UY	Positive	(e.g. *yes, yeah, yup, mhm*)

or else of being loosely attached (prosodically speaking) to a larger syntactic structure, normally either at the beginning or, less commonly, at the end.

1.5.1.2.2 Adverbs in morphosyntactic annotation

Like interjections, adverbs are dealt with cursorily by existing EAGLES guidelines and practices. Leech and Wilson (1994) simply include recommended subcategories for base, comparative and superlative forms, as well as for interrogative adverbs such as *when, where* and *how*. Apart from these, various syntactico-semantic functions of adverbs (such as place, frequency and manner) can easily be recognised through optional extensions. On the whole, however, tagset makers have avoided subcategorising adverbs, on the following grounds: Adverbs constitute a loosely organised word class, in which even well-known subcategories, such as time, place, degree, manner and stance adverbs, are notoriously difficult to distinguish by hard-and-fast criteria, and certainly difficult to recognise and tag automatically. Yet it is worth noting that two tagsets for English, which have been devised with spoken corpora in mind, do subcategorise adverbs in considerable detail. These are the London-Lund Corpus tagset (Svartvik and Eeg-Olofsson 1982) and the International Corpus of English tagset (Greenbaum and Ni 1996). Table 1.2 with brief extracts from the London-Lund Corpus tagset gives an impression of how the adverb part of speech can be usefully subcategorised for spoken language.
Again, this (incomplete) list illustrates the diversity and importance of adverbial components in speech, and may serve as a starting point for a more sophisticated tagset. In any case, it is clear that one must consider carefully the addition of subcategories to the tagset before undertaking a morphosyntactic tagging of spoken data.

Table 1.2: Some adverb subcategories from the London-Lund Corpus

TAG	CATEGORY	SUBCAT	SUBSUBCAT or ITEM	EXAMPLE
AApro	adverb	adjunct	process	*correctly*
AAspa	adverb	adjunct	space	*outdoors*
AAtim	adverb	adjunct	time	*how*
...	...	...	...	...
AQgre	adverb	discourse item	greeting	*goodbye*
AQhes	adverb	discourse item	hesitator	*now*
AQneg	adverb	discourse item	negative	*no*
AQord	adverb	discourse item	order	*give over*
AQpol	adverb	discourse item	politeness	*please*
AQpos	adverb	discourse item	positive	*yes, [mm]*
AQres	adverb	discourse item	response	*I see*
...	...	...	...	...
Asemp	adverb	subjunct	emphasiser	*actually*
ASfoc	adverb	subjunct	focusing	*mainly*
ASint	adverb	subjunct	intensifier	*a bit*
...	...	...	...	...

1.5.1.3 Extending the part-of-speech categories in EAGLES morphosyntactic guidelines

Returning to the *interjection* category, one tentative proposal is for an extended use of the I ('interjection') POS category in the EAGLES morphosyntactic guidelines (Leech and Wilson 1994), with the subcategories presented in Table 1.3, which are based on those in Biber et al. (1999, forthcoming), Chapter 14.

Table 1.3: Extended interjection POS categories.

TAG	CATEGORY	EXAMPLES (English)
I1	exclamations	*oh, ah, ooh*
I2	greetings/farewells	*hi, hello, bye*
I3	discourse markers	*well, now, you know*
I4	attention signals	*hey, look, yo*
I5	response elicitors	*huh? eh?*
I6	response forms	*yeah, no, okay, uh-huh*
I7	hesitators/filled pauses	*er, um*
I8	polite formulae	*thanks, sorry, please*
I9	expletives	*God, hell, shit*

These subcategories cover the major 'interjection' phenomena which occur in spoken English generally. However, there is one major caveat over their use in morphosyntactic annotation: many of the words in these classes are liable to occur in more than one of the subcategories, so that ambiguity, in fact rather

sophisticated polysemy, can be a major headache for automatic tagging, or even for manual tagging. For example, *oh*, classified above as an exclamation, in many instances behaves more like a discourse marker; *okay*, classified as a response form, can also occur as a response elicitor and as a discourse marker. A way out of this problem is to regard all the subcategory names in the table as preceded by the word 'primarily': e.g. *oh, ah,* etc. are designated as 'primarily exclamations', leaving any ambiguities at this level unresolved.

1.5.1.4 Residual problems

The sections on interjections and adverbs above illustrate two further difficulties to bear in mind when tagging spoken data.

One is the extremely unclear boundary between these two peripheral parts of speech. We note, in fact, that the two tagsets above, that of Sampson for the SUSANNE Corpus, and that of Svartvik and Eeg-Olofsson for the London-Lund Corpus, are somewhat inconsistent with one another in where they draw the boundary: whereas Sampson places greetings such as *good-bye*, response forms such as *yes* and the politeness marker *please* among interjections, Svartvik and Eeg-Olofsson place them among adverbials. This is an area where drawing the line between categories appears to be little more than an arbitrary decision.

Another phenomenon of spoken language illustrated above is the tendency for lexicalised multi-word expressions such as *I see, I'm sorry, thank you* and *sort of* to occur with greater density than in written texts. It might be argued that this phenomenon of *multi-words* (or *multi-word unit*) can be ignored, if one really wants to, in tagging written language (as indeed it is ignored by some well-known taggers). But it can scarcely be ignored in tagging spoken language. The problem, for morphosyntactic annotation, is whether these expressions should be decomposed into their individual orthographic words for tagging purposes, or whether they should be assigned a single tag labelling the whole expression, as in the lists above. If a single *multi-tag* is used, this raises the question of how to represent, in the formal encoding of morphosyntactic tags, this discrepancy of 'more than one orthographic word = one morphosyntactic word' (see Garside et al. 1997, pp. 20–22).

1.5.1.5 An alternative solution

An alternative solution is to argue that the different kinds of 'interjection' in 1.5.1.3 above really differ on the functional plane, and that therefore these distinctions belong not to the level of morphosyntactic annotation, but to that of pragmatic annotation (see further 1.8.1.6.1 below). The rationale for this approach is provided by Fischer (1996), Fischer (1998) and Fischer and Brandt-Pook (1998), where it is shown that a broad class of *discourse particles* can be differentiated functionally and distributionally in a way that facilitates the automatic analysis of dialogue. The discourse functions which these particles perform comprise a limited list: *take-up, backchannel, frame, repair marker, answer, check, modal* and *filler*. On the morphosyntactic level, however, a broad differentiation between conjunctions, modal particles and discourse particles may be sufficient (with the possible addition of multi-word categories of speech routines (e.g. *you know*) and pragmatic idioms (e.g. *good-bye*). Of these, conjunctions (e.g. *but*) are connective, being outside the sentential unit

themselves, while modal particles (e.g. *schon*) are integrated into the sentential unit and the intonation contour, and discourse particles (e.g. *okay*) are not grammatically integrable, but are able to constitute entire utterances.

1.5.2 Recommendations

It is recommended that dialogue corpus creators as far as possible adopt existing EAGLES or EAGLES-related guidelines for morphosyntactic annotation, especially as specified in Leech and Wilson (1994) and Monachini and Calzolari (1996). The main areas where these guidelines need to be extended are in the number and definition of tags used for adverbials and interjections. It is proposed that these be dealt with along the lines proposed in this subsection, with attention to the alternative analyses suggested in 1.5.1.3 and 1.5.1.4. Standards for morphosyntactic annotation are still evolving, and there is room for further discussion and adaptation particularly in relation to the annotation needs of spontaneous dialogue.

1.6 Syntax

1.6.1 Syntactic annotation

Syntactic annotation, as distinct from morphosyntactic or POS annotation, has up to now taken the form of developing *treebanks* (see e.g. Leech and Garside 1991; Marcos-Marín et al. 1993) or corpora in which each sentence is assigned a tree structure (or partial tree structure). Treebanks are usually built on the basis of a phrase structure model (see Garside et al. 1997, pp. 34–52); but dependency models have also been applied, especially by Karlsson and his associates (Karlsson et al. 1995).

Until very recently, little spoken data has been syntactically annotated. There is an EAGLES document (Leech et al. 1996) proposing some provisional guidelines for syntactic annotation, but this again, while acknowledging their existence, does not handle the special problems of syntactically annotating spoken language material.

With syntactic annotation, as with tagsets, the inventory of annotation symbols has been generally drawn up with written language in mind. An example of syntactic annotation of written language is the following sentence from a Dutch journal, encoded minimally according to the recommended EAGLES guidelines of Leech et al. (1996):

```
[S[NP Begin juni NP] [Aux worden Aux] [VP[PP in [NP het Scheveningse
Kurhaus NP]PP] [NP de Verenigde Naties NP-Subj] [AdvP weer AdvP]
nagespeeld VP]. S]

(At the beginning of June the United Nations will again be enacted in
the Scheveningen 'spa'.)
```

The following is an example of a different syntactic annotation scheme, that of the Penn Treebank,[12] applied to a spoken English sentence:

```
( (CODE SpeakerB3 .))
```

[12] "ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/"

```
( (SBARQ (INTJ Well)
         (WHNP-1 what)
         (SQ do
             (NP-SBJ you)
             (VP think
                 (NP *T*-1)
                 (PP about
                     (NP (NP the idea)
                         (PP of
                             ,
                             (INTJ uh)
                             ,
                             (S-NOM (NP-SBJ-2 kids)
                                    (VP having
                                        (S (NP-SBJ *-2)
                                           (VP to
                                               (VP do
                                                   (NP public service work))))
                                        (PP-TMP for
                                                (NP a year)))))))))
         ?
         E_S))
```

Just as with morphosyntactic annotation (see Section 1.5), we note that in early development of syntactic annotation (especially the IBM-Lancaster treebank, 1987–1991 — see Leech and Garside (1991), there seemed to be nothing seriously inappropriate in the use of syntactically annotated *written texts* on a large scale as a training corpus for *speech* recognition applications. Recently, the development of treebanks including or comprising spoken language has confronted a number of research groups with the same problem of adapting syntactic annotation practices to spontaneous spoken language. The four research groups which have been tackling this problem for English data are:

- UCREL, Lancaster (see Eyes 1996) working on a sample treebank of the BNC
- Marcus and his associates working on the Penn Treebank[13]
- Sampson and his associates working on the CHRISTINE corpus at Sussex[14] (Sampson wrote an anticipatory Chapter 6 on treebanking spoken data in Sampson (1995), which reports on the earlier SUSANNE treebank of written data.)
- Greenbaum, Nelson, and others working on the International Corpus of English at University College London (Greenbaum 1996; Nelson 1996).

1.6.1.1 Dysfluency phenomena in syntactic annotation

Again as with morphosyntactic annotation, the adaptation of syntactic annotation is necessary in order to deal with dysfluency. The main phenomena requiring special treatment are:

- Use of hesitators or 'filled pauses'
- Syntactic incompleteness
- Retrace-and-repair sequences
- Dysfluent repetition
- Syntactic blending (or anacoluthon)

[13] "`http://www.cis.upenn.edu/treebank/home.html`".

[14] "`http://www.cogs.susx.ac.uk/users/geoffs/RChristine.html`".

In considering what solutions may be applied to the syntactic annotation involving these kinds of dysfluency, we will mainly refer to solutions adopted by Sampson (1995), Ch. 6, and for the UCREL syntactic annotation scheme by Eyes (1996). The other two research initiatives mentioned above (the Penn Treebank and the International Corpus of English) have taken a different approach, which bypasses the problem of syntactic annotation of dysfluencies entirely. They have adopted schemes for explicitly annotating dysfluencies. These features may then, if necessary, be excluded from the syntactically annotated material, by applying syntactic annotation only to a normalised version of the data. This normalised version may be represented, alongside a record of the dysfluent material, by the use of mark-up devices like the TEI deletion or regularisation tags (see for example 1.4.1.8.3 above). The approach of Sampson and of UCREL, on the other hand, is to include the dysfluent material in the syntactically annotated material, by means of a set of guidelines devised for that purpose.

1.6.1.1.1 Use of hesitators or 'filled pauses'

Hesitators such as *um* and *er* can be handled relatively unproblematically (in Sampson's terms) by treating them as equivalent to unfilled pauses. In syntactic annotation of written corpora, generally, punctuation marks are incorporated into the syntactic tree, being treated as terminal constituents comparable to words. For the training of corpus parsers, this is a useful strategy, since punctuation marks generally signal syntactic boundaries of some importance. Similarly, for spoken language, it is an advantage to adopt the same strategy, and to treat pause marks like punctuation, as in effect 'words' in the parsing of a spoken utterance. This strategy is then extended to filled pauses or hesitators.[15] The general guideline adopted by UCREL and by Sampson (SUSANNE) is that punctuation marks are attached as high in the syntactic tree as possible, i.e. they are treated as immediate constituents of the smallest constituent of which the words to the left and to the right are themselves constituents. This policy generalises very naturally to hesitators, regarded as vocalised pause phenomena.

1.6.1.1.2 Syntactic incompleteness

Syntactic incompleteness occurs where the speaker fails to complete an utterance, owing to self-correction, to interruption, or to some other disruption of the speech production process. In 1.5.1.1 above we discussed the case of word fragments (incomplete or truncated words) as a problem for morphosyntactic annotation. On the syntactic level there is a comparable problem of non-terminal constituent fragments, where a constituent is interrupted before its completion:

```
<pause> [NP you NP][VP ‘re [NP/ a British NP/]V] <pause>
```

[15] In a similar spirit, in the International Corpus of English morphosyntactic annotation, hesitators are tagged with a 'negative' label UNTAG, which signifies that the item so tagged cannot be assigned to any part-of-speech category (Greenbaum and Ni 1996).

This example from the BNC guidelines illustrates the use of a special marker (in this case a slash following the non-terminal constituent label) to indicate that the constituent is incomplete. In Sampson's scheme, instead, a marker is inserted *within* the incomplete constituent, to indicate the locus of the interruption:

```
[S [NP she ] [VP was going ] [PP into [NP the # ] ] ]
```

(adapted from Sampson 1995, p. 454)

It should incidentally be noted here that, as a matter of principle as well as of practice, the issue of the (un)grammaticality of syntactically incomplete sentences does not generally arise with treebanks (see Sampson 1987). In written data, as well as in spontaneous speech, ungrammaticality (by the standards of formally defined rule-driven parsers) is found to be of frequent and routine occurrence. Therefore any automatic syntactic annotation of spoken or written data has to cope with this phenomenon - for example, by the adoption of robust probabilistic parsing algorithms which will provide an adequate syntactic annotation for every sentence or utterance. No special dispensation is required for spoken data containing dysfluencies.

1.6.1.1.3 Retrace-and-repair sequences

We will use this term to refer to frequent cases (also known as 'false starts') where a speaker 'interrupts' the production process by discontinuing the construction of the current constituent, returning to an earlier point of the same utterance (thereby notionally deleting the sequence 'retraced'), and restarting from there. Sampson proposed the use of a marker (again #) to signal the interruption point, and the inclusion of both the retrace and the repair within a minimal superordinate constituent:

```
and that [NPs any bonus [RELCL he ] # money [RELCL he gets over that ] ]
is a bonus
```

This example, liberally adapted from Sampson (1995), p. 453, uses the minimum bracketing needed to demonstrate the point. The labels adopted are those in the EAGLES preliminary syntactic annotation guidelines (Leech et al. 1996). The example shows how, on either side of the interruption point #, two relative clauses, the former incomplete, are handled as co-constituents of the same noun phrase.

1.6.1.1.4 Dysfluent repetition

Repetition, as a manifestation of dysfluency, occurs where the speaker shows hesitation by repeating the same word, or the same sequence of words, before proceeding with the normal production process. The repetition can be iterated. In Sampson (1995) this repetition is again handled by the intervening use of the interruption-point marker #. It is treated, in effect, as a special case of a retrace-and-repair sequence, where the retrace and the repair are identical:

```
[O Oh [S [NP I ] [VP don't think ] # [NP I ] [VP don't think ] [NCL I ever
went to see mine ] S] O]
```

(This is again adapted from Sampson (1995), p. 457), with use of labelled bracketing in accordance with EAGLES syntactic annotation guidelines, to illustrate the point.)

1.6.1.1.5 Syntactic blends (or anacolutha):

These occur where, in the course of an utterance, a speaker changes tack, failing to complete the syntactic construction with which the utterance began, and instead substituting an alternative construction. E.g. the switch to a non-matching tag question in: *And there's an accident up by the Flying Fox, is it?* (example from the BNC - though, in isolation from header information, it is impossible to know whether this is, for instance, a Welsh dialectal variant). Since no test of grammaticality is generally applied to treebank annotations, the annotation of cases like the one above causes no problem and probably needs no special annotation. More drastically incoherent sentences, however, do occur quite frequently in spontaneous speech. An example (from the BNC) is:

```
(1)    And this is what the, the <unclear> what's name now
       now <pause> that when it's opened in nineteen ninety-two
       <pause> the communist block will be able to come through
       Germany this way in.
```

In this utterance, punctuated as a single sentence, there appear to be three word sequences between which there is no common superordinate constituent, and so a minimal analysis of the following general form is adopted according to the BNC guidelines (# is again added to indicate interruption points):

```
(1a)   [ And this is what the #, the <unclear> ] # [ what's
       name now # now  ] # <pause> [ that when it's opened in
       nineteen ninety-two <pause> the communist block will be
       able to come through Germany this way in ].
```

This example illustrates the effect of what the BNC guidelines call a 'structure minimisation principle', which specifies that a syntactic annotation should not contain more information than is warranted in the context. A possible source of inconsistent parsing practice is that different grammarians will interpret the incoherent sentence differently – one reading into the sentence a particular structure, and another another structure. This can often be avoided if annotators err on the side of omission rather than inclusion of uncertain information.[16] In example (1a) above, there is no clear warranty for making the three major segments fit into a single covering constituent. Similarly, it may

[16] Geoffrey Sampson's comment, however, (personal communication) is that although it is good to use the principle of "if in doubt, leave structure out", "again and again one seems forced to make a choice between different and equally defensible analyses."

be felt unwarranted to give particular syntactic labels to these segments. One option which is allowed in the BNC guidelines (again in line with the 'structure minimisation principle') is the omission of labels where there are no clear criteria for the assignment of a particular label. This option is followed in (1a) above. On the other hand, there are arguable grounds for labelling the three segments as sentence (S), sentence (S) and nominal complement clause (NCL) respectively. Hence the following is an alternative, slightly fuller annotation:

(1b) `[S And this is what the #, the <unclear> S] # [S what's name now # now S] # <pause> [NCL that when it's opened in nineteen ninety-two <pause> the communist block will be able to come through Germany this way in NCL].`

1.6.1.1.6 Concluding remarks on syntactic annotation and dysfluency

At present the syntactic annotation of spontaneous spoken language is at a pioneering stage, and the practices shown above should be regarded as tentative and incomplete. With this serious reservation, the above illustrations do show how syntactic annotation practices may be adapted to cope with dysfluent features of spontaneous speech. The two major methods employed – that of normalisation by excluding dysfluencies and that of stretching syntactic annotation to include the parsing of dysfluencies – have complementary advantages. The normalisation option enables spoken data to be automatically parsed with relatively little need to customise software for spontaneous spoken input, since major dysfluencies can be edited out. On the other hand, the inclusion option is preferable to the extent that it provides some parsing information even for incompleteness and repair phenomena. It can be pointed out, also, that the normalisation procedure cannot be applied to some markedly dysfluent utterances such as example (1) above. Here it is not at all clear what a normalised version of the utterance would be.

1.6.1.2 Unintelligible speech

Another problem related to that of syntactic incompleteness arises in dialogue when the circumstances of speech production or of recording leave passages of speech unclear or unintelligible (cf. 1.4.1.7.3 and 1.4.1.7.4 above). Example (1) in Section 1.6.1.1.5 above shows how an unintelligible passage (tagged <unclear>) may be incorporated into a syntactic phrase marker. The general treatment of unintelligibility is parallel to that of incomplete constituents. Just as a marker # was introduced by Sampson (see 1.6.1.1.2) to signal the location of a point of interruption, so a marker such as <unclear> may signal the point where the parsing information cannot be recovered because of unintelligibility. The tag <unclear>, unlike <pause>, refers to a verbal sequence. The only problem is that the annotators do not know which words the speaker used. The strategy here, then, is to *include* <unclear> within parse brackets wherever this appears appropriate, in order to 'complete' an otherwise incomplete constituent. Examples:

`So [NP all these [ families and <unclear> ]NP]`

```
No but [S <unclear> [NP twenty one NP]S] [S aren't you S]?
```

In the first case, it is obvious that <unclear> fills the gap in an otherwise incomplete coordinate construction. In the second case, the incompleteness arises from a gap at the *beginning* of the main clause. We can guess that the unclear words are *you are* or *you're*, because of the tag question which follows. So we have some warrant to include <unclear> within the [S ... S]. However, on the principle of minimising structure, we refrain from inserting any further brackets.[17].

1.6.1.3 Segmentation difficulties

The syntax of spoken dialogue may seem fragmentary or disorderly for reasons other than dysfluency or unintelligibility. Some reasons are:

(i) The canonical sentence of written language, as a structure containing a finite verb, is far from the being a satisfactory basis for the segmentation of speech into independent syntactic wholes. According to one count by Leech (Biber et al. 1999, forthcoming, Ch. 14), ca. 39% of the independent syntactic units of conversational dialogue have no finite verb: many are single-word utterances typically consisting of a single interjection in the extended sense of 1.5.1.2.1. The practice in the compilation of treebanks has often been to use parse brackets (conventionally [S ... S]) to enclose the whole parsable unit, but to make no assumption that what occurs within those brackets should have the structure of a canonical sentence. Thus a stand-alone noun phrase unit, such as *No problem*, should be parsed simply [S [N *No problem* N] S]. The [S ... S] brackets may be interpreted as 'sentence' or, say, as '(syntactic) segment', according to the annotator's or user's preference. For our present purpose, the term *C-unit*[18] will be used for a segment parsed as an [S ... S] which is not part of another [S ... S].

(ii) The criteria for what counts as a C-unit in speech are difficult to determine, and may have to rely on prosodic separation (for example the boundary of a major tone group or intonation phrase).

(iii) There are utterance turns in dialogue where one speaker completes a syntactic construction begun by another speaker.

There appear to be four methods of segmenting a dialogue into C-units:

(a) The C-unit should be delimited by criteria internal to syntax. That is, where no syntactic link can plausibly be established between one parsable unit and

[17] Again, however, Geoffrey Sampson comments (personal communication) that this strategy is not always satisfactory. "It is all right if the stretch of unclear wording was in fact a constituent or part of a constituent, but sometimes it manifestly includes material on either side of a constituent boundary."
Sampson proposes some guidelines to deal with this situation. For a discussion, see Rahman and Sampson (1998).

[18] In educational linguistics, the term *C*-unit has evolved on the model of Hunt's T-unit (Hunt 1965) as a measure of syntactic complexity in children's written language. It is an attempt to define a 'maximal parsable unit' for speech. One attempted definition (Chaudron 1988, p. 45) begins with a definition of the T-unit as 'any syntactic main clause and its associated subordinate clauses' and goes on to define a C-unit as 'an independent grammatical predication; the same as a T-unit, except that in oral language, elliptical answers to questions also constitute complete predication'. Although this definition is still influenced by written norms, the concept of a maximally parsable unit of spoken language underlies it.

another, they are treated as independent. This solution, however, does not address point (ii) above.

(b) The C-unit should be delimited by prosodic criteria, either alone, or in conjunction with syntactic criteria where these are clear. This solution, obviously, depends on the existence and quality of a prosodic level of annotation.

(c) The C-unit should be delimited by orthographic criteria: that is, by treating sentence-final punctuation marks (specifically periods and question marks) as boundaries. This is the simplest method to apply, assuming that the orthographic transcription is so punctuated. On the other hand, it is the most arbitrary, since punctuation marks are artefacts of the transcription, and do not have a warranted linguistic function.

(d) The C-unit should be delimited by pragmatic, functional or discoursal criteria. Apart from the turn boundary, which is no doubt the clearest delimiter one can use for parsing, pragmatic and discoursal criteria are probably no clearer in determining C-units than internal syntactic criteria. However, in the development of language engineering dialogue systems, considerable effort has been invested in the recognition of functionally-defined segments corresponding to dialogue acts. Moreover, in this context, the importance of syntactic annotation is in facilitating the automatic recognition and delimitation of such functional units, rather than parsing as an end in itself. Hence there is much to be said for relying on functional criteria as the most valuable guide to segmentation for purposes of dialogue annotation.

1.6.2 Recommendations

Dialogue corpus builders are recommended

1. to consult the existing EAGLES provisional recommendations on syntactic annotation (in Leech et al. 1996),
2. to bear in mind the need to extend and modify these recommendations in the light of the needs of syntactic analysis for spoken dialogue,
3. to be aware that syntactic annotation may need to be correlated with other categories, for example syntactic boundaries with prosodic (e.g. intonation) boundaries,
4. to note the complexity and, at the present time, very frequently the theory-dependence of tree-based syntactic annotation in comparison with other types of annotation,
5. to take into account the fact that syntactic annotation is at a less advanced stage of evolution than POS tagging, that there is relatively little consensus even on basic phrasal syntactic categories, and the area is a matter for ongoing research.
6. to remember that the notion of 'maximally parsable unit' is not a clear-cut notion for spoken language, and may range from word (e.g. interjection) length to turn length.

1.7 Prosody

1.7.1 Prosodic annotation

Prosodic labelling remains one of the major problem areas in the annotation of spoken data generally, and spoken dialogue in particular. This section takes the section on prosody in the *EAGLES Handbook* (Gibbon et al. 1997, p. 161 ff)

as its starting point, and brings it up to date in the light of recent work in the field.

In written text, as already noted in 1.4.1.7.2, use is sometimes made of punctuation marks to signal broad intonational distinctions, such as a question mark to indicate a final rise in pitch or a full stop to signal a final fall. Since it is well established that there is no one-to-one mapping between prosodic phenomena and syntactic or functional categories, it is important for a prosodic annotation system to be independent of syntactic annotation systems. In Southern Standard British English, for example, a rise in pitch may be used with a syntactically marked question, but this is not necessarily, and in fact not usually the case. On the other hand, questions with no syntactic marking often take a final rise, as, apart from context, it is the only signal that a question is being asked. A fully independent prosodic annotation allows for investigations into the co-occurrence of prosodic categories with dialogue annotations at other levels, once the annotations are complete.

Prosodic annotation systems generally capture two main types of phenomenon: (i) those which lend *prominence*, and (ii) those which divide the speech up into *chunks* or *units*. Words are made prominent by the accentuation of (usually) their lexically stressed syllable. Many Western European languages have more than one accent type. It is thus necessary to capture not only on which word an accent is realised but also which kind of accent is used. Since in some cases the accent may occur on a syllable other than that which is assigned the primary lexical stress of a word, some annotation systems tag explicitly the syllable (or the vowel in the syllable) upon which an accent occurs, rather than the word as a whole. Such a representation, however, requires a finer annotation of the corpus at a non-prosodic level than simple orthography, e.g. a segmentation into syllables or phoneme-sized units.

Common to all annotation systems is the division of utterances into prosodically-marked units or phrases, where prosodic marking may include phenomena such as *audible pause* (realised as either actual silence or final lengthening), *rhythmic change*, *pitch movement* or *reset*, and *laryngealisation*. Dividing an utterance into such units is usually the first step taken when carrying out a prosodic annotation, as many systems place restrictions on their internal structure. However, the size and type of prosodic units proposed by the systems described below differs considerably.

It is currently common practice for manual prosodic annotation to be carried out via auditory analysis accompanied by visual inspection of a time-aligned speech pressure *waveform* and *fundamental frequency* (F_0) track. This is the case for the *ToBI* annotation system described in 1.7.1.1 below. Additional information, e.g. spectrogram or energy, may also be available. Despite this, we report on one system, *Tonetic Stress Marks* (TSM) in 1.7.1.2, which originally used to rely entirely on auditory analysis, a well-established system which has been used for the annotation of a digitally available database.

Phenomena occurring across prosodically defined units, such as current pitch range, are not symbolically captured by any of the systems described below. A number of systems incorporate a means by which such information can be retrieved from the signal. For example, ToBI has a special label for the highest F_0 in a phrase. The F_0 value at this point may be used to give an indication of

the pitch range used by the speaker at that particular point in time. *INTSINT* marks target points in the F_0 curve which are at the top and bottom of the range. However, the range is determined for a whole file which might be one or more paragraphs long. Register relative to other utterances is only captured in cases where the beginning of a unit is marked relative to the end of the previous one (e.g. in INTSINT). Somewhat more flexible is the SAMPROSA system, which uses bracketing to indicate the extent of pitch register and range features. However, none of the manual annotation methods capture structures at a more macro level than the intonation phrase or its equivalent.
All existing representation systems for intonation have drawbacks. For a list and description of some of those systems, see Gibbon et al. (1997), p. 161 ff.

1.7.1.1 ToBI

The ToBI (Tones and Break Indices) system is an established *de facto* standard for the prosodic annotation of digital speech databases in General American English. It has been successfully applied to Southern Standard British and Standard Australian English, but, since it is an adaptation of a phonological model, it is not claimed to be applicable as it stands to other varieties of English or to other languages. However, it has been modified to apply to a number of other languages, including German and Japanese.
It has been made clear in the ToBI documentation that ToBI does not cover varieties of English other than those listed above, and that modifications would be required before it could be used for their transcription. In the ToBI guidelines it is stated that "ToBI was not intended to cover any language other than English, although we endorse the adoption of the basic principles in developing transcription systems for other languages, particularly languages that are typologically similar to English" (Beckman and Ayers Elam 1997, Section 0.4). The implication in Silverman et al. (1992) that ToBI aimed to meet the need for a suprasegmental equivalent to the IPA is therefore to be ignored. It is the basic principles behind ToBI, rather than a set of phonologically-motivated categories, which allow its adaptation to other languages.
A ToBI transcription consists of a speech signal and F_0 record, along with time-aligned symbolic labels relating to four types of event. The two main event types are tonal, arranged on a *tone tier* and junctural, arranged on a *break index tier*. There is additionally a *miscellaneous tier* for the annotation of non-tonal events such as voice quality or paralinguistic and extralinguistic phenomena, and a further tier containing an orthographic transcription, the *orthographic tier*. The tone and break index tiers are discussed below.

1.7.1.1.1 ToBI tones

As far as the tonal part of ToBI is concerned, the basic principles are taken from the phonological model of English intonation by Pierrehumbert (1980). This model has given rise to a substantial number of studies within what has been termed by Ladd (1996) as the *autosegmental-metrical framework*. Some of these studies have developed into similar ToBI systems for other languages. Others lay down the groundwork for such an adaptation, but have not yet been applied to the annotation of large-scale corpora.

Within the autosegmental-metrical framework, tones are used in two major ways: they can be part of an *accent* or they can be involved in the signalling of a *boundary*. Tones may be *high* (H) or *low* (L). Accents may contain one or more *tones*. If there is more than one tone in an accent, it is important that the tone which aligns with the prominent syllable be marked as such. This is done by means of an asterisk (or star) diacritic. By default, monotonal pitch accents have the star on their only tone. The inventory of pitch accents is language or dialect specific.
Tones signalling the boundaries of prosodically defined phrases may occur at their left or right edges. Whether a tone (or, in principle more than one tone) may occur at a boundary of a given domain is, again, specific to individual languages or dialects, as is the number and types of domain which allow for tonal marking.
The ToBI inventory for General American English (more recently referred to as E_ToBI) has five basic pitch accents, the glosses are taken from Beckman and Hirschberg (ToBI annotation conventions):

Table 1.4: ToBI Pitch Accents

H*	'peak accent'
L*	'low accent'
L+H*	'scooped accent'
L*+H	'rising peak accent'
H+!H*	'clear step down onto the accented syllable'

All of the H tones in the above inventory may be marked with a '!' diacritic which indicates that they are downstepped relative to the immediately prior H tone. The downstep diacritic is obligatory in the H+!H* accent. The others, if downstepped would be transcribed !H*, L+!H*, L*+!H, and, in principle, !H+!H*. The prerequisite for using a ! diacritic is that there must be at least one H tone prior to the downstepped tone from which it can be stepped down.
There are two domains at the right edge of which there is an obligatory tone: the *intermediate phrase* and the *intonation phrase*. Intonation phrases contain at least one intermediate phrase. The tones available at the right edge of the intermediate phrase are:

(1) L-
(2) H-

The right edge of an intonation phrase is automatically the right edge of an intermediate phrase. It is customary to label the sequence of tones at these two right edges together. Since there is also the choice of H or L tone at the intonation phrase boundary, there are four combinations to choose from:

(1) L-L%
(2) L-H%
(3) H-H%
(4) H-L%

The '-' diacritic is used for intermediate phrase boundaries and '%'for intonation phrase boundaries. One problematic aspect of the transcription of boundaries is the fact that the phonetic implementation of the tone sequences is far from transparent. The H% or L% is raised by an automatic 'upstep' if it follows H-. This means that H-H% symbolises a high rising boundary reaching a level very high in the current pitch range (H% is upstepped), and H-L% symbolises a high level boundary (the L% is upstepped to the same value as the previous H- tone).
One further edge tone may optionally be used. This is an intonation phrase initial boundary tone, transcribed: %H.

1.7.1.1.2 ToBI break indices

In the current ToBI system, there are five levels of perceived juncture, referred to as *break indices*, between words transcribed on the orthographic tier. They are numbered from 0 to 4. The lowest degree of juncture between two orthographic words is level 0, where the words are grouped together into a 'clitic group', e.g. between 'did' and 'you' pronounced as 'didya'. Level 1 is the default boundary between two words in the absence of any other prosodic boundary. Levels 3 and 4 correspond to intermediate phrase and intonation phrase boundaries. Since these latter two break index levels are linked to the tonal representation, the system might be argued to be circular. However, there is provision in the system for signalling cases where there is a mismatch between the tonal boundary transcribed and the perceived juncture. This is provided by a '-' diacritic, as in '4-', and by level 2. Level 2 can be used where there is tonal evidence to indicate a level 3 or 4 boundary but a lower degree of perceived juncture. Alternatively, it can also indicate a high degree of separation between the words without the corresponding tonal evidence.
It is important to point out here that the break indices are perceptual categories. In order to assign them, transcribers need make use of auditory and visual information only.

1.7.1.1.3 Using the ToBI system

A major advantage of the ToBI system is that there are extensive training materials and well-developed tools for carrying out the annotation. For instance, a transcriber can listen to cardinal examples of all of the pitch accent and boundary types at any point during transcription.[19]
The majority of ToBI users are also users of the proprietary commercial signal annotation software ESPS/waves+^{TM}, which includes a multi-tier annotation tool; there are also extensive training materials, and the ESPS format has become a near-standard.[20]

[19] The ToBI labelling guide, including electronic text and accompanying audio example files, is available at "`http://ling.ohio-state.edu/Phonetics/E_ToBI/etobi_homepage.html`".

[20] As this volume was going to press, we received the news that Entropic, the manufacturer of ESPS/waves+, has been acquired by Microsoft and that all current Entropic tool-kit and software development kit products have been withdrawn from the market. This clearly invalidates the recommendations in the text which refer to Entropic software. - Ed.

However, the fact that ESPS/waves+^{TM} is a high-end commercial product has been an obstacle for less the well-endowed laboratories in the majority of countries of the world who are looking for viable prosodic signal annotation systems.
There have been recent attempts to address this imbalance, in that training materials are now available over the world wide web with incorporated audio files and time-aligned transciptions, F_0 tracks and speech waveforms. A `.au`/`.gif` format version of the Guide is currently available in beta version at the ToBI homepage URL.
A less expensive commercial software package, Pitchworks, developed by SciCon in cooperation with the University of California, Los Angeles, fulfils basically the same requirements as ESPS/waves+, and is becoming the preferred prosodic annotation tool in many laboratories.
A public domain program, 'fish', which uses Tcl/Tk running under Unix, has been developed by Reyelt, Universität Braunschweig, a member of the German ToBI group.[21] It supports data exchange using Esprit SAM formats.
A freeware system with full multi-tier labelling facilities, and also a well-equipped set of phonetic analysis tools, is the 'Praat' phonetic productivity software developed by Boersma at the University of Amsterdam. Although the system has a more complex graphical user interface than most, it is also in many respects more powerful than other systems, and can be recommended for the more advanced user.
Praat is available at: "`http://www.fon.hum.uva.nl/praat`"
Provided that inexpensive or public domain software continues to be available, the ToBI system can be recommended. However, if required for a language for which ToBI has not yet been adapted, very careful adaptation of the tone inventory must be performed.

1.7.1.1.4 ToBI for other languages and dialects

J_ToBI is a transcription standard for Standard (Tokyo) Japanese, developed in collaboration between linguists at Ohio State University, USA, and speech engineers at ATR Interpreting Telecommuncations Research Laboratories, Japan.[22]
GToBI is a consensus transcription system for German developed by a multi-site group including universities in Saarbrücken, Braunschweig, Stuttgart, Erlangen and Munich.[23]

The training materials introduce basic pitch accents and edge tones along with tonal modifications such as upstep and downstep. For training purposes

[21] It is currently available at the following address: "`http://sbvsrv.ifn.ing.tu-bs.de/reyelt/`".

[22] An HTML version of the training materials containing audio (`.au`) and graphics (`.gif`) is available at "`http://ling.ohiostate.edu/Phonetics/J_ToBI/jtobi_homepage.html`". From here there is a link to an ftp site containing a postscript version of the Guide, audio files in ESPS and SUN `.au` format, and eps, `.gif`, and `.ps` files of F_0 track, waveform, and labels. A hard copy is also available (Venditti 1995).

[23] Information about the standard and a `.ps` version of the training materials (Benzmüller and Grice 1997) is available at the following address: "`http://www.coli.uni-sb.de/phonetik/projects/Tobi/gtobi.html`".

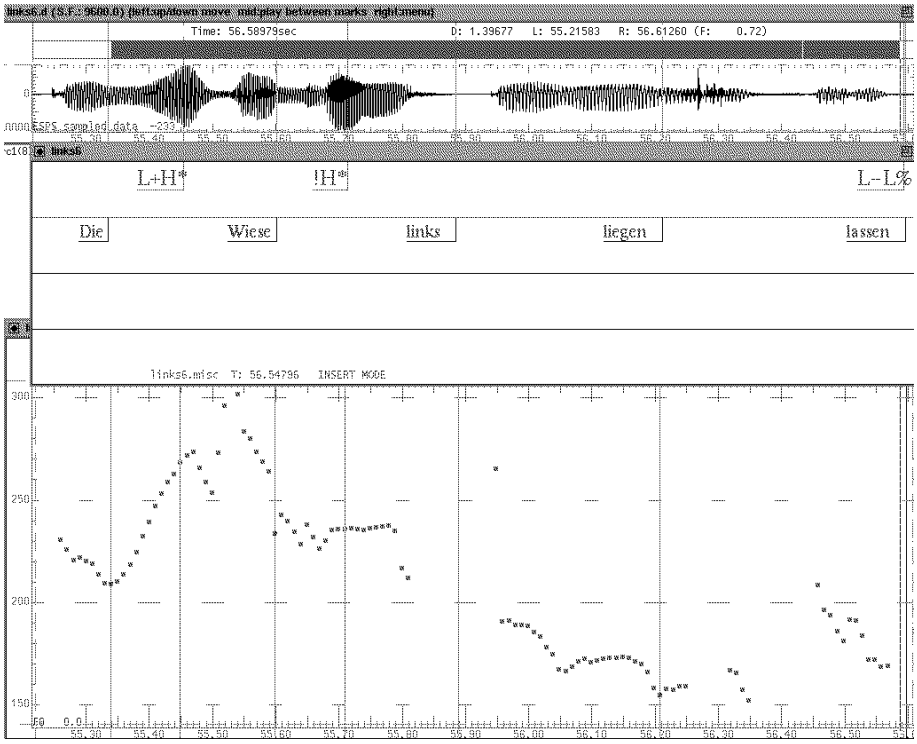

Figure 1.1: An example of GToBI transcription, time-aligned with an F_0-track.

schematic diagrams and lists of important criteria for each category are provided, along with pointers to speech files containing canonical examples. The speech signal files, available in headerless binary Unix and ESPS formats are available on demand at the address on the page. Inter-transcriber agreement ratings are reported in Reyelt et al. (1986) and Grice et al. (1996). Results show that GToBI is already adequate for large-scale database annotation with labellers of differing expertise at multiple sites.

In addition to the existence of ToBI systems for Japanese and German, an adaptation to the English ToBI has been made for the transcription of western Scottish (Glaswegian) English, GlaToBI, Mayo et al. (1997). Although no training materials are available, the system has been used in cross-transcriber consistency tests. The adaptations made include an L*H accent, representing a rise (rather than, say, a L valley as in L*+H) which is aligned with the accented syllable, and the elimination of automatic upstep of boundary tones after a H- intermediate phrase tone. In GlaToBI, H-L% represents a fall, rather than a level stretch as in E_ToBI.

It has been argued (Nolan and Grabe 1997) that ToBI, by which E_ToBI is meant, is too phonological for the comparison of dialects of English. This is to be expected, since it was not designed to do this. The adaptation necessary for GlaToBI illustrates this point.

Autosegmental-metrical analyses on related principles have been carried out in greater or less detail for many languages. These are, amongst others, Dutch (Gussenhoven 1984, 1993; Gussenhoven and Rietveld 1991), Bengali (Hayes and Lahiri 1991), American Spanish (Sosa 1991), Greek (Mennen and den Os 1993; Arvaniti 1994), Italian (Grice 1995; Avesani 1990; D'Imperio 1997), French (Post 1993), European Portuguese (Frota 1995).

1.7.1.2 TSM – Tonetic stress marks

The *tonetic stress marks system*, as used for the transcription of the SEC corpus (see Knowles et al. 1996, , p. 51–57 for a critical account) is based on the British school style of auditory intonation analysis. The TSM transcription system has two levels of intonation phrasing: the *major tone group*, the end of which is marked with a double bar, '‖', and the *minor tone group*, the end of which is marked with a single bar '|'. The TSM system indicates the presence and tonal characteristics of every accent by means of a diacritic before the accented syllable.
There is no internal structure to the major or minor tone groups, except that they must contain at least one accented syllable. The tones in the TSM inventory are:

(1) level
(2) fall
(3) rise
(4) fall-rise
(5) rise-fall,

each of which may be *high* or *low*, where *high* means that the starting point of the tone is higher than the previous pitch and *low* that the starting point is lower.
If an accented syllable is final in a tone group, marking it with a given tone determines the pitch from the beginning of that syllable up to the tone group boundary. The domain includes all syllables up to but not including the next accented syllable or end of tone group.
The corpus which has been auditorily transcribed using this method is the Lancaster IBM Spoken English Corpus (SEC), which has been digitised and is now also available as MARSEC (MAchine Readable Spoken English Corpus.)[24] The original SEC, transcribed by Briony Williams and Gerry Knowles, was completed in 1987 and comprises five different parts:

(1) Spoken recording
(2) Unpunctuated transcriptions
(3) Orthographic transcriptions
(4) Prosodic version
(5) Grammatically tagged version

MARSEC, developed by Peter Roach, Simon Arnfield and Gerry Knowles, contains a time-aligned version of the original corpus including annotations. Most of the files are in Entropics/waves+^{TM} format although there are also versions of the original files in PC format.

[24] cf. "`http://midwich.reading.ac.uk/research/speechlab/marsec/marsec.html`".

The British school type of analysis, using TSM at least for nuclear tones, has successfully been adapted to a number of languages. However, it is not, as far as the authors of this document are aware, currently being used for database annotation in any of these.

1.7.1.3 Conversion between ToBI and the TSM system

An attempt to devise a system for automatically converting nuclear intonation contours in the TSM transcription into E_ToBI has been made by Roach (1994). Although ToBI and the TSM system both have two levels of phrasing, these two levels do not map onto each other in a straightforward way. The minor tone group corresponds to the intonation phrase. There is no equivalent in ToBI of the major tone group. Furthermore, there is no equivalent in the TSM system of intermediate phrase boundary marking. However, as a first approximation, Roach suggests placing an intermediate phrase boundary after every kinetic (i.e. non-level) tone.
Of note is that the conversion uses only a subset of ToBI tone notations: those with the starred tone notation in initial position (i.e. H*, L*, and L*+H). This is because nuclear tones in the British system capture the pitch from the beginning of the accented (nuclear) syllable up to the end of the tone group. This precludes, in Roach's view, the use of leading unstarred tone notation (L in L+H* for instance) in the conversion; this would not necessarily be the case, however, if for instance an L+H* combination were analysed as a unit.
Roach's conversion table, slightly modified, is as follows:

Table 1.5: Conversion between TSM and ToBI, according to Roach (1994)

TSM description	at intermed. boundary	at inton. boundary
low level	(no level tones here)	L* L-L%
high level	(no level tones here)	H* H-L%
rise-fall	L*+H L-	L*+H L-L%
high fall-rise	?	H* !H-H%
high fall	H* L-	H* L-L%
low fall	!H* L-	!H* L-L%
high rise	H* H-	H* H- H%
low rise	L* H-	L* L-H%
low fall-rise	?	!H* L-H%

The main problem Roach finds is where fall-rises are transcribed in tone-unit medial position, converted into intermediate phrase final position. Here the ToBI system cannot capture the fall rise. It would need a sequence of HLH, and since the final H would have to be the boundary, then the pitch accent would have to be H*+L, an accent which is missing in the English inventory, falls being usually captured by a combination of H* and one or more low phrase tones.
Ladd maintains that "it is pointless to attempt to state a complete correspondence" (Ladd 1996, p. 82) between Pierrehumbert's analysis (the model upon

which ToBI is based) and the British school, though this is not exactly a valid argument, and he does in fact give a table of correspondences; Ladd's table differs from Roach's in a number of respects. Two major differences are as follows.

Ladd gives more than one equivalent for certain British-style nuclear tones as he also makes use of leading unstarred tones. For example, he lists L+H* L-L% as corresponding to a rise-fall, and L*+H L-L% as corresponding to an emphatic version of this tone. Roach on the other hand specifically rejects the possibility of using L+H* L-L% as a rise-fall because "perceptually the effect of rise-fall is of a pitch movement with strong prominence at the onset" (Roach 1994, p. 96). Roach uses downstepped tones as equivalents of the low versions of the tones. This is understandable, as the definition of the 'low' variants of the tones in the SEC TSM system is that they begin lower than a previous syllable. However, there are problems with this analysis, since in ToBI downstep can only be used on a non-initial H tone in a phrase. This means that a low fall which is the only accent in a phrase would be converted into !H* L-, which would be ruled out as illegal. Ladd does not use downstepped H tones as equivalents of the beginnings of low nuclear tones. Instead, he takes other options, such as L* L-L% to represent the low fall.
A short look at the differences in the correpondence tables leads to the conclusion that caution must be taken if any conversion is attempted in either direction. However, perhaps the mere fact that correspondences have been sought is an indication that of all the systems described here, the two most compatible are TSM and ToBI.
But the discussion also shows that the criteria for comparing systems have not yet been well worked out by the discussants, either in terms of their respective linguistic level of analysis, or in terms of their formal properties, or in terms of their phonetic (production, transmission and perception) properties, or in terms of actual data reliably transcribed and time-aligned in more than one system.

1.7.1.4 INTSINT

INTSINT (International Transcription System for Intonation) aims at providing a system for cross-linguistic comparison of prosodic systems. It has been developed by Daniel Hirst, and is based on a stylisation algorithm in which the F_0 (fundamental frequency) pattern is approximated by a spline interpolation function between F_0 target points.[25]
Transcription in INTSINT represents the prosodic target points aligned with an orthographic or phonetic transcription. It can be used at different levels of detail, allowing a narrow as well as a broad phonetic pitch transcription. Although it is conceived as a system for cross-language comparisons, language-specific subsets of elements can be recommended.
INTSINT is based on the postulate that "the surface phonological representations of a pitch curve can be assumed to consist of phonetically interpretable symbols which can in turn be derived from a more abstract phonological representation" (Hirst 1991, p. 307). In favour of the spline interpolation approach

[25] Information on INTSINT can be obtained from "`http://www.lpl.univ-aix.fr/~hirst/intsint.html`".

to the stylised representation of pitch curves, Hirst (1991) quotes evidence from acoustic modelling studies showing that pitch targets account better for the data than pitch changes, and evidence from perceptual studies claiming that pitch patterns are predominantly interpreted in terms of pitch levels. INTSINT aims therefore at the symbolisation of pitch levels or prosodic target points, each characterising a point in the fundamental frequency curve.

The symbolisation of prosodic target points is made by means of arrow symbols corresponding to different pitch levels. Higher, Upstepped, Lower, Downstepped or Same are tonal symbols describing relative pitch levels defined in relation to a previous pitch target or to the beginning of an intonation unit. Top or Bottom are tonal symbols describing absolute pitch levels described in relation to the operative range of the intonation unit; Mid is assumed to occur only at the beginning of an intonation unit, and is then considered unmarked. Hirst et al. (1991) have shown that, at least for French, the prosodic targets can be defined with respect to the speaker's F_0 (fundamental frequency) mean (Mid) to one point fixed at a half-octave interval above the mean (Top) and to one point fixed at a half-octave interval below the mean (Bottom). The F_0 modelling is carried out automatically by a program called MOMEL (Hirst and Espesser 1993) that, after F_0 detection, provides the best fit for a sequence of parabolas, dividing the F_0 curve into a microprosodic and a macroprosodic profile. The microprosodic component is caused by the individual segmental elements of the utterance, and the macroprosodic component reflects the intonation patterns produced by the speaker (Hirst and Espesser 1993). The output of the programme is a sequence of target points with a time value in ms. and a frequency value in Hz. Target points can be then automatically coded into INTSINT symbols, once the position of the intonation unit boundaries has been manually introduced.

An experiment comparing listener's evaluation of a synthesised text using original target points and INTSINT-coded target points has shown that the INTSINT version attained more than 80% of the score attributed to the version synthesised with the original target points (Hirst et al. 1991).

Within the MULTEXT project a tool is planned for the automatic symbolic coding of F_0 target points using INTSINT. A preliminary description of such an algorithm is given in Hirst (1991) (see also Hirst et al. 1994), which attempts to provide an optimal INTSINT coding of a given curve by seeking to minimise the mean squares error of the predicted values from the observed values. Absolute pitch values Top, Mid and Bottom are modelled by their mean values and relative pitch levels are modelled by a linear regression on the preceding target point.

One major difference between INTSINT and other models described so far is that symbols are aligned simply with a point in the signal. In the TSM system, a nuclear tone begins on a stressed syllable and is transcribed immediately before this syllable. In ToBI a tone is marked with a star to signal alignment with the lexical stress of a given word, allowing for the capture of timing differences such as that between L+H* and L*+H where the rise is earlier in the first than the second. ToBI also uses diacritics to signal alignment with a given boundary (although only loosely in the case of intermediate phrase edge tones). In INTSINT, on the other hand, target points are simply coded for their height,

of which there are five categories (as opposed to two in the ToBI and TSM systems). Information as to the alignment of the target point with a given constituent can be retrieved, if there is a parallel analysis of the utterance into such constituents. Distinctions regarding the timing of target points in relation to accented syllables (such as L+H* and L*+H above, or *early*, *medial* and *late* peak (see Kohler 1987)) are not captured in the tonal annotations. Again, actual alignment information is not explicitly coded, but retrievable through the linking up of different levels of annotation, assuming that they are available.

1.7.1.5 Automatic annotation of prosody in VERBMOBIL

Details of the prosodic annotation employed in the VERBMOBIL project are given in Gibbon et al. (1997), pp. 165–168. VERBMOBIL has two types of manual annotation, KIM and ToBI (as reported on in Reyelt et al. (1986) and Grice et al. (1996)). The prosodic labelling system PROLAB, based on the Kiel Intonation Model (KIM), is described in Kohler (1995). The model itself is described in Kohler (1991, 1996). Here we deal with automatic annotation, which is carried out separately.

Prosodic information is currently being used in the following analysis modules in VERBMOBIL: syntactic analysis, semantic construction, dialogue processing, transfer, and speech synthesis. Clause boundaries, for example, are successfully detected at a rate of 94%. A word hypothesis lattice or graph (WHG) and the speech signal serve as input for the prosodic analysis, which then enriches the WHG with prosodic information based on "the relative duration [...]; features describing F_0 and energy contours like regression coefficients, minima, maxima, and their relative positions; the length of the pause (if any) after and before the word; the speaking rate; [...]" (Niemann et al. 1997b, p. 2). Probabilities for accent on the word, clause (or sentence) boundaries and sentence mood are computed and used to facilitate syntactic analysis at clause or sentence level, to disambiguate sentence particles like *noch* ('still' vs. 'another') on the semantic level, to segment dialogue acts through the use of prosodic boundaries, to enable transfer from German to English by taking into account the sentence mood, and to 'imitate' the voice of the original speaker in speech synthesis by adapting pitch level and speaking rate.

Based on the results of this kind of prosodic analysis, the number of possible parse trees in the syntactic analysis can be reduced by 96% and processing time sped up by 92%. Below, we give one example each of prosodic disambiguation on the syntactic and the semantic level:

(1a) "*Vielleicht. Am Montag bei mir. Paßt das?*"
"*Maybe. On Monday, at my place. Is that OK?*"

(1b) "*Vielleicht am Montag. Bei mir paßt das.*"
"*Maybe on Monday. That's possible for me.*"
(Niemann et al. 1997b, p. 2)

(2a) "*Dann müssen wir noch einen* Termin *ausmachen.*"
"*Then we still have to fix a* date."

(2b) "*Dann müssen wir* noch *einen Termin ausmachen.*"
"*Then we have to fix* another *date.*"
(Niemann et al. 1997b, p. 3)

In (1), identifying the clause boundaries prosodically helps to delimit the ut-

terances automatically and to classify them according to dialogue acts. In (2), disambiguation of the particle *noch* is achieved by identifying the presence (1b) or absence (1a) of primary stress/accent on it.

1.7.1.6 Prosodic studies on the ATIS project

Various studies have been conducted on the corpus data collected on the ATIS project in order to determine whether prosodic features can be exploited in the automatic analysis of human–machine interaction. These studies deal partly with finding ways of automatically identifying structural elements of discourse (Wang and Hirschberg 1992) and partly with developing strategies for identifying and correcting dysfluencies (cf. Section 1.5.1.1) on the basis of prosodic information (Nakatani and Hirschberg 1994).
Wang and Hirschberg's study on the automatic clasification of intonational phrase boundaries had the explicit aim of detecting in which way structural/prosodic information predicted from the text can serve as a first step towards identifying the structural elements of texts, and determining how this information can be augmented and made more reliable by exploiting observed prosodic information in order to improve speech recognition and synthesis.
In order to achieve this aim, they used *classification and regression tree* (CART) techniques (Riley 1989) to determine the most salient features, after having first manually annotated the data prosodically according to Pierrehumbert's model of intonation (Pierrehumbert 1980).
Based on their analysis, they identified a combination of prosodic and (morpho-) syntactic features that can be used to detect prosodic boundaries more reliably:

1. Similar length of adjacent (preceding & current) phrases.
2. General length of a phrase. The occurrence of a boundary becomes more likely if a phrase is longer than $2\frac{1}{2}$ seconds, but less likely if the resulting phrase is less than half the length of the preceding phrase.
3. Accentuation, i.e. a boundary is more likely to occur after an accented word.
4. Syntactic constituency, e.g. the relative inviolability of an NP.
5. Word-class. A boundary is less likely to follow after function words other than *to, in* or a conjunction.

However, the results of their analysis also show that the success of automatic detection of phrase boundaries drops when dysfluent uttterances in the data are not 'normalised'. They therefore conclude that dysfluent boundaries cannot be phonologically categorised in the same way as fluent boundaries and may present a problem for automatic analysis. This is especially important as:

> The quality of the ATIS corpus is extremely diverse. Speakers range in fluency from close to isolated-word speech to exceptional fluency. Many utterances contain hesitations and other dysfluencies, as well as long pauses (greater than 3 sec. in some cases). (Wang and Hirschberg 1992, p. 12)

The problem that dysfluent utterances present for speech recognition or, more precisely, *spoken language understanding systems*, is treated in Nakatani and Hirschberg (1994). In their study, they try to identify *repair cues* and how those, in turn, may be used to detect and correct repairs efficiently in order to

facilitate the analysis of spontaneous speech. They define *repair* as "...the self-correction of one or more phonemes (up to and including sequences of words) in an utterance." (Nakatani and Hirschberg 1994, p. 7)
To illustrate how dysfluent speech can cause problems for a speech recognition system, they give the following examples of ill-recognised speech from the ATIS corpus:

> (1) *Actual string*: What is the fare *fro-* on American Airlines fourteen forty three
> *Recognised string*: With fare *four* American Airlines fourteen forty three
> (2) *Actual string*: Show me all *informa-* information about aircraft type, Lockheed L one zero one one
> *Recognised string*: Show meal *of make* information about aircraft flight Lockheed L one zero one one
> (3) ...Delta leaving Boston seventeen twenty one arriving Fort Worth *twenty two* twenty one forty and flight number ... (Nakatani and Hirschberg 1994, p. 2)

While all three examples here represent the occurrence of *false starts*, examples (2) and (3) represent dysfluent *speech fragments*, whereas example (3) is clearly different from the first two with respect to the fact that the utterance may be correctly recognised, but is nevertheless not correctly interpretable. Based on the frequent occurrence of fragments in dysfluent speech, they conclude that:

> ...the interruption of a word is a sure sign of repair, and so we expect the that the ability to distinguish word fragments from non-fragments would be a significant aid to repair detection. (Nakatani and Hirschberg 1994, p. 9)

In order to classify and help to correct the different types of repair, they set up a *Repair Interval Model* (RIM), based on their analysis, using CART techniques.
This model distinguishes between three sub-intervals, each interval possibly containing a number of features that may aid in the detection of repairs:

1. *Reparandum Interval*: Covers the lexical material that is to be repaired. May consist of word fragments, unfragmented words that are repeated or even (noun) phrases that are respecified. Fragmentation seems to occur more frequently in content words and most of fragments appear to be one syllable or less in length. Glottalisation may accompany fragmentation and when it does, seems to be distinct from *creaky voice*. One further distinction between fluent phrase boundaries and non- fluent ones is the absence of *final lengthening* in the latter.
2. *Dysfluency Interval (DI)*: Extends from the Interruption Site (IS) to the point of resumption of fluent speech. Characterised partly by silent, rather than filled pauses which are generally shorter than fluent pauses, whereby they tend to be even shorter for fragment repairs than non-fragment ones. However, pausal duration alone does not appear to be a reliable indicator of repairs and has to be examined in conjunction with other factors, such as a possible increase in F_0 and amplitudes from the last accented syllable of the reparandum to the first syllable of the correcting material and the possible occurrence of matching spectral-time or lexical patterns.

3. *Repair interval*: Contains correcting material.

The implications of their study are that, in order to detect different types of repair, different methods of analysis, such as spectral-time pattern matching, the analysis of pausal duration, the use of phone-based recognisers, etc. might be employed in conjunction with one another in order to improve the detection and subsequent correction of dysfluent utterances.

1.7.1.7 X-SAMPA and SAMPROSA

X-SAMPA is a computer-compatible version of the International Phonetic Alphabet, including all diacritics, and symbols for prosody and intonation (see Appendix A). It is well established in the phonetics and speech technology fields for the transcription and annotation of phoneme-sized segments. However, one of the main weaknesses of the IPA, and by extension also of its computer-compatible equivalent, is the provision for the transcription of prosody and intonation. The fact that there are many models currently in use, both in basic and applied research, makes standardisation an impossible task. It is not simply a matter of choosing which symbol to use, but rather of choosing which phenomena are to be captured. It is therefore necessary to have a computer-compatible alphabet for prosodic annotation which attempts to cover the breadth of the field.

An attempt to meet this need is SAMPROSA,[26] which was designed for application in multi-tier transcription systems. SAMPROSA requires that intonational annotations be transcribed on an independent tier from other transcriptions or representations of the signal. It is argued that symbolic representations on different tiers may be related in two different ways. They may be related through explicit association between prosodic and segmental units such as those on a phone, syllabic or orthographic tier. This is the autosegmental-metrical approach used in the ToBI system, and to some extent in the TSM system. Alternatively, they may be simply related by synchronisation: "The symbols may be assigned to the signal as tags or annotations; the temporal relations between symbols are then given empirically (extensionally) via their position with respect to the signal" (see footnote on SAMPROSA). This is the approach taken by the INTSINT system.

It is important to point out that neither X-SAMPA nor SAMPROSA are transcription systems as such. They are computer-compatible codes for use in formatting transcriptions for interchange purposes, once a model has been selected. Alternatively, they can be used for computer-coding extensions to existing models, leading to improved readability across the different approaches. Since they are working standards and not set in stone, if extensions to the underlying categories (i.e. the IPA, or an intonation transcription system) are introduced, they are open to extension.

1.7.2 Recommendations

It is not possible to make absolute recommendations in the field of prosodic annotation. The ToBI transcription system is to be recommended, if it is to be used for languages or dialects for which there is already a standard. However,

[26] "`http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm`".

it is not to be adopted wholesale for a new language or dialect. Rather it is to be adapted, where possible referring to existing autosegmental-metrical work in the literature. The INTSINT method of annotating intonational phenomena is a method which requires little adaptation for a new language, and can be recommended as an alternative, although the phenomena covered are not the same as those covered by ToBI. Although it was originally designed to be used on a purely auditory basis, the TSM system, as long as it is supported by making recourse to an F_0 track, provides a third prosodic transcription method. But of course a well-validated and reliable set of *auditory* transcription categories is valuable in its own right and may in some circumstances be preferable to a purely acoustic or production based set of categories.
Since the field is rapidly developing, it is advisable that anyone wishing to undertake prosodic annotation consult the links provided in this document before beginning work.

1.8 Pragmatics

1.8.1 Pragmatic annotation: functional dialogue annotation

1.8.1.1 'Historical' background

From a historical perspective, it should be mentioned that since the 1960s, there has developed a considerable body of linguistic research on the communicative structures and components of dialogue. On the one hand, linguistic philosophers such as Austin (1962) and Searle (1969, 1980) developed the concepts 'illocutionary act' and 'speech act', to explore and define the range of functional meanings associated with utterances. On the other hand, in sociolinguistics and discourse analysis, various segmental models of dialogue behaviour have been developed by Sinclair and Coulthard (1975), Ehlich and Rehbein (1975), Stubbs (1983) and Stenström (1994), among others. Such studies have often assumed that dialogues can be exhaustively segmented into units, and that these units can be reliably assigned a particular functional interpretation. Some have assumed that there is a hierarchy of such dialogue acts, analogous to the hierarchy of units (word, phrase, clause, etc) in syntax. These approaches have sometimes influenced the assignment of dialogue acts for automatic speech processing, and provide a foundation for general studies of dialogue analysis.
Another historical influence on dialogue research has been the work of the philosopher H.P. Grice on the understanding of spoken communication in terms of the intentions of the speaker (see Grice 1969). However, this is probably of little relevance to the applications-oriented R&D, which is the focus of this chapter.
In the more immediate context of LE, much of the work on dialogue analysis and annotation has up to now been done by the members of the Discourse Resource Initiative (DRI) and many links can be found on its homepage.[27] The DRI holds annual workshops in an attempt to unify previous and ongoing annotation work in dialogue coding. Out of the first workshop of the DRI, there evolved a coding scheme, called DAMSL (Dialog Act Markup in Several Layers), which served as a basis for annotation of the 'homework' material assigned to participants for

[27] "http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html"

the second workshop at Schloß Dagstuhl, Germany.[28] Since then the DAMSL scheme has been revised to incorporate at least some of the suggestions made by the participants of the workshop.[29] Further recommendations, especially with regard to the coding of higher-level discourse structures, are to be expected as the outcome of the third DRI workshop in May 1998 in Chiba, Japan (see Nakatani and Traum 1998).[30]
The DRI workshops may be seen as 'milestones' in the development of dialogue coding and represent a concerted effort to establish international standards in this field. Most of our recommendations are, at least to a considerable extent, based upon their workshop materials and reports.

1.8.1.2 Methods of analysis and annotation

The pragmatic annotation of dialogues constitutes a special case. Whereas the coding of all other levels of representation/annotation discussed so far may to an extent be performed independently, ideally pragmatic annotation makes use of information from all other levels.
Within LE projects, two different methods for the segmentation, annotation and analysis of dialogue are employed. Dialogues are segmented and annotated either automatically (VERBMOBIL, TRAINS) or manually using online marking tools (Instructions for Annotating Discourses, TRAINS, HCRC Map Task). None of the projects seem to rely on purely 'manual' annotation schemes, i.e. without the support of any online annotation tools, such as coders or SGML markup tools. Note that the term *segmentation* is sometimes used to refer to either structural or functional units, an ambiguity which is probably best avoided. We use the term unambiguously to refer only to the structural/textual level and not the functional one.
One of the main problems in analysing discourse is to separate form from content, in other words to distinguish the structural from the functional level. Although, for example, a speaker's turn may correspond to only one sentence on the structural/syntactic level, on the functional level it may correspond to more than one speech act or form only one part of a larger functional unit (see Section 1.8.1.4 for more details). This duality may sometimes lead to confusion if the same term is used to refer to both a structural and a functional unit within the dialogue, e.g. the term *turn* being used synonymously with *speech act*. In the context of this document, 'structural' may be understood as 'utilising information available from the orthographic, syntactic or prosodic levels of representation/annotation'.

1.8.1.3 Segmentation of dialogues

Before analysing any dialogue according to its functional elements, it is first necessary to segment it into textual units that serve as a basis for its repre-

[28] "http://www.dag.uni-sb.de/ENG/Seminars/Reports/9706/"

[29] "http://www.cs.rochester.edu/research/trains/annotation/RevisedManual/RevisedManual.html".

[30] However, at the moment of writing, we have only had very cursory information as to the outcome of this workshop, so that we can only give very sketchy details in the appropriate sections. We assume that more detailed information will be made available at the DRI website in due course.

sentation and annotation. This may have to be done manually, but in most cases will nowadays be done automatically according to the criteria outlined in Section 1.4. Within the turn (see 1.4.1.3 above), the most commonly used basic unit for this is the *structural utterance*, which will often, on the syntactic level, correspond to what we called in Section 1.6 a *maximally parsable unit* or *C-unit*. This may correspond to a traditional sentence, or, in many cases, to a single 'stand-alone' word or phrase. Note that some documents on dialogue coding may actually refer to structural utterances as *phrases* (see Nakatani et al. 1995). However, we recommend using the term *structural utterance* as *utterance* is the most commonly used term within the LE community.[31] However, we think that using it without the attribute *structural* may lead to confusion as the same term is often used to identify *functionally relevant* items as well and therefore propose a two-way distinction between *structural* and *functional utterances*.

In order to segment the turns of a dialogue into individual structural utterances, it seems to be more or less common practice to use mainly syntactic clues or pauses, sometimes supplementing them by making recourse to intonational clues. In fact, assuming that an orthographic transcription has already been undertaken (see Section 1.5), a pre-interpretative segmentation of the text will have been undertaken already, using such clues in the marking of full stops (see 1.4.1.7.2) or other punctuation marks. In this case, it will be the dialogue act annotator's task to refine those structural utterance units already tentatively identified in the orthographic text representation, splitting or merging such units where necessary.

When prosodic clues are used, they are still in practice usually based upon the transcriber's auditory interpretation and not on actual physical evidence. Two notable exceptions here are the VERBMOBIL project and some of the studies done on the ATIS corpus, which use pattern-matching techniques based on the F_0-contour and other prosodic features to establish structural utterance units, (see 1.7.1.5 and 1.7.1.6 above for more detail). Work of a similar kind is being undertaken within the framework of the TRAINS project as well.

Various different techniques are employed to represent structural utterances in the text. Most projects will initially make use of some kind of orthographic transcription as outlined in 1.4 and may later refine it according to more functional criteria. Some researchers prefer to store each functional utterance (no matter how short it may be) on one line by itself, whereas others group utterances according to 'intuitive sentences' and separate individual structural utterances from each other by using such symbols as a forward slash (/) (Condon and Cech 1995). However, important as the structural analysis may be, it may be seen as no more than a preliminary to functional annotation and great care has to be taken not to overemphasise the importance of structural elements such as line breaks, so that they may inadvertedly be confused as having functional significance.

As already noted, apart from the utterance, there is only one higher-order structural unit, which is generally referred to as *turn* (see 1.4.1.3). (It is also sometimes referred to as a *segment*; however, the use of the term 'segment' here may be slightly problematic, as it may be confused with segments identified at

[31] To add to the potential confusion, 'utterance' is sometimes used (e.g. in the TEI encoding of spoken texts) as equivalent to a *turn* (see 1.4.1.3).

the phonetic level.) A turn generally comprises the sequence of utterances produced by a single speaker up to the point where another speaker takes over. However, cases of overlap also have to be taken into account. Turns which totally overlap with another turn need to be coded separately since they may have functional significance, for example as expressions of (dis)agreement on the part of the interlocutor. In contrast to the structural utterance discussed immediately above, it is more important to mark turns at the pragmatic level because it is always important to be clear about who is speaking at any given time.

1.8.1.4 Functional annotation of dialogues

The functional annotation of dialogues, sometimes also referred to as *dialogue act annotation* , is a means of capturing and encoding different levels of discourse structure, and identifying how they relate to one another at the pragmatic level. Previously, there had been some debate as to whether this type of coding should try to capture information about a *speaker's intention* or the pragmatic *effect on the dialogue*, but this issue seems to have been resolved at the third DRI workshop at Chiba, in favour of coding with regard to the latter as a speaker's intention may not always be clear to the coder. Functional annotation plays an increasingly important role in current LE applications such as automatic translation systems, generation of summaries of dialogue content, etc. (see, for example, Alexandersson et al. 1997). Functional annotation will be examined first from the point of view of individual utterances (in Section 1.8.1.5) and secondly from the point of view of multi-level annotation (in Section 1.8.1.6).

1.8.1.5 Utterance tags

To characterise the function of individual utterances, the annotator may apply *utterance tags* that characterise the role of the utterance as a dialogue act. The revised DAMSL manual identifies four different dimensions according to which utterances may be classified: (1) 'Communicative Status', (2) 'Information Level', (3) 'Forward-Communicative-Function' and (4) 'Backward-Communicative-Function'. One additional dimension, that is not included in the DAMSL manual, but was discussed at the Dagstuhl conference, is that of 'Coreference'. This, however, may be regarded as a more general aspect of discourse annotation (including the annotation of written texts) and, as such, is beyond the scope of this document. We thus end up with a general four-way distinction for classifying dialogues (slightly expanded with respect to the DAMSL categories), which is discussed in more detail below.

1.8.1.5.1 Communicative status

Communicative status refers to whether an utterance is intelligible and has been successfully completed. If this is not the case, then the utterance may be tagged as either

(1) *Uninterpretable*,
(2) *Abandoned*
or
(3) *Self-talk*.

1.8.1.5.2 Information level and status

Information level gives an indication of the semantic content of the utterance and how it relates to the task at hand. The revised DAMSL manual offers a four-way distinction between

(1) *Task* ('Doing the task'),
(2) *Task-management* ('Talking about the task'),
(3) *Communication-management* ('Maintaining the communication') and
(4) *Other* (a dummy category for anything that is relevant, but cannot be categorised according to (1)–(3)).

The members of the Dagstuhl conference, however, decided that a three-way distinction would probably be more practical and proposed two alternative classifications:

a. (1) *Task*, (2) *About-task*, (3) *Non-relevant*
b. (1) *Task*, (2) *Communication*, (3) *Non-relevant*

Information status distinguishes between whether the information contained in an utterance contains *old* or *new* information. This distinction is not included in the DAMSL manual, but was discussed at Dagstuhl, where four alternative schemes where considered:

a. Retain a simple distinction between (1) *old* and (2) *new*,
b. Add a category (3) *irrelevant*,
c. Subdivide *old* into (a) *repetition* (including anaphora), (b) *reformulation* (or paraphrase) and (c) *inference* (to bridge anaphora)
d. Define four categories: (1) *repetition* (2) *reformulation* (3) *inference* and (4) *new*.

1.8.1.5.3 Forward-looking communicative function

Dialogue utterances that may be tagged as having forward-looking communicative function are those utterances that could constrain future beliefs and actions of the interlocutors and thus affect the subsequent discourse.
The four categories of the DAMSL manual are:

(1) *Statement*: e.g. *assert, reassert, other-statement*, etc.,
(2) *Influencing-addressee-future-action*: e.g. *request, question, directive*, etc.,
(3) *Committing-speaker-future-action*: e.g. *offer, commit*, etc. and
(4) *Other-forward-(looking-)function*: dummy category for fixed, relatively rare functions like *conventional-opening, conventional-closing*, etc.

No particularly noteworthy differences from the DAMSL manual emerged from the Dagstuhl conference, but note that category (4) may possibly be subsumed under information level category (3) *communication-management*.
One issue that has been raised at the Chiba workshop is the role of acknowledgements and dysfluency phenomena (see 1.5.1.1, 1.6.1.1 and 1.7.1.6) with regard to their possible forward-looking functions and how they may be integrated into a coding scheme.

1.8.1.5.4 Backward-looking communicative function

In contrast to those utterances that have a forward-looking communicative function, utterances that relate to previous parts of the discourse may be annotated as backward-looking. The DAMSL categories for this are:

(1) *Agreement*: e.g. *accept, maybe, reject, hold*, etc.,
(2) *Understanding*: e.g. *backchanneling, signal-non-understanding, signal-understanding*, etc.,
(3) *Answer*: generally signals compliance with a request for information,
(4) *Information-relation*: utterances expressing explicitly how an utterance relates to the previous one,
and
(5) *Antecedents*: any utterance may be marked as relating to more than just the preceding one.

1.8.1.5.5 General remarks on the above categories

The two final categories in 1.8.1.5.3 and 1.8.1.5.4 above do not seem to be mutually exclusive as there can be some overlap between them, i.e., it is sometimes difficult to decide whether an utterance is completely forward-looking or backward-looking. It might therefore be better to think of them as '*Primarily* Forward-looking (Communicative) Functions' and '*Primarily* Backward-looking (Communicative) Functions'. However, the DAMSL manual does not exclude the possibility of assigning multiple tags for forward- or backward-looking communicative functions and this concept was again reconfirmed at the Chiba workshop. Also, whereas the former two categories *communicative status* and *information level and status* primarily relate to the micro level of dialogue structure, the latter two can be seen as the building blocks for the higher-level structures discussed below.

1.8.1.6 Levels of functional annotation

In addition to a sequence of individual utterances, it is common to posit a hierarchy of dialogue units of different sizes. In conversational analysis the term *adjacency pair* (Sacks 1967–1972) has been commonly used for a sequence of two dialogue acts by different speakers, the second a response to the first. Similarly, in discourse analysis the term *transaction* has been used for a major unit of dialogue devoted to a high-level task, and the term *exchange* for a smaller interactive unit, not dissimilar to the adjacency pair (see Sinclair and Coulthard (1975); Stubbs (1983); see also Gibbon et al. (1997), pp. 568–569 on the application of these concepts to dialogue systems). It has further been proposed that such hierarchical groupings of dialogue acts can be modelled in terms of a *dialogue grammar* (see Gibbon et al. 1997, p. 185).
Multi-level functional annotation may be undertaken by determining the dialogue function of individual (meaningful) utterances and grouping them according to three different levels, the *micro*, the *meso* and the *macro levels*, although not all researchers make use of such a three-level distinction. These will be discussed in 1.8.1.6 below.

1.8.1.6.1 Micro-level annotation

Micro-level annotation seeks to identify the minimal meaningful functional units within the dialogue and to determine their functional value for the dialogue by assigning utterance tags to them. The annotation may be performed automatically as in the VERBMOBIL project or – most commonly – manually.
Both in the automatic and manual annotation of functional utterances/dialogue acts, we encounter similar problems, which were discussed in detail at the Dagstuhl & Chiba conferences. They are briefly outlined below and some recommendations as to their solution will be given in Section 1.8.1.7.[32] Since these problems concern annotation of content rather than of form, we shall refer to them as problems of *functional annotation*. They are related to, yet (at least in principle) distinct from, the problems of syntactic segmentation discussed under 1.6.1.3).

- *Pragmatic particles, discourse markers and interjections* (e.g. *well, okay, alright*): e.g. Where are these to be treated as utterances in their own right, and where as parts of others? (On 'interjections' used in a broad sense relevant here, see 1.5.1.2.1).
- *Hesitations*: What role do hesitations have in the delimitation of utterances?
- *Coordinated sentences* (e.g. sentences linked by *and*): When are coordinators like *and* to be regarded as beginning a new utterance?
- *Subordinate sentences* (e.g. *... so; ... because*): The same question arises with subordinated as with coordinated sentences.
- *Reformulations*: For example: Are they to be treated as constituting a different utterance or dialogue act from the utterances they reformulate?
- *Suggestions and requests for their confirmation*: For example: Should they be regarded as separate utterances?

Members of the Dagstuhl conference essentially identified the following three types of *functional boundaries*:

(1) *regular utterance-token boundaries* (suggested mark-up: @) correspond to what are referred to as *utterances* above.
(2) *weak utterance-token boundaries* (suggested mark-up: *) are optional subunits.
(3) *drop-in utterance-token boundaries* (suggested mark-up: $) serve to delimit phenomena such as self-repair and hesitations, which can interrupt other segments and do not have a functional role in relation to what precedes or follows.

However, category (3) is not necessarily to be taken at face value, since *self-repairs* or *hesitations* may actually fulfil functional roles, as pointed out earlier (see Section 1.4.1.7.1), and may therefore better be included under (1) or (2). Based upon the above categories, a set of five *annotation rules* was proposed:

(1) Annotate utterances that serve to perform an illocutionary function (@)
(2) When in doubt as to whether to annotate or not, do not annotate.
(3) If there are strong indicators, e.g. prosodic boundaries such as a long pause, annotate (@). (Note: but only in cases which are compatible with Rule (1).)

[32] Note that the Dagstuhl paper refers to them as problems of *segmentation*, but that, in line with our earlier reservation regarding the term *segment*, we prefer to avoid it here.

(4) Even when speakers collaborate in the completion of a unit, annotate at locations of speaker change (@).
(5) Optional: Annotate smaller units using weak boundaries (*) where the resulting sub-units serve the same illocutionary function.

In accordance with (3), Nakatani and Traum (1998) recommend treating discourse particles or cue phrases as separate utterance tokens, but note that this may not always be advisable for the former as they can sometimes be difficult to distinguish from other word classes, e.g. German *schon*, which may be used as either a discourse particle or an adverb. Some general remarks on the identification of utterances/dialogue acts are provided in Section 1.8.1.7 and on the coding of boundaries/utterances in Section 1.8.1.9.

1.8.1.6.2 Meso-level annotation

Meso-level annotation groups individual functional utterances into higher-order units directly above the micro-level of individual utterances/dialogue acts. There currently seem to exist two slightly different major approaches to treating meso-level structures: those exemplified by the HCRC Map Task Corpus (see Carletta and Taylor 1996) and by the Draft Coding Manual (see Nakatani and Traum 1998)[33] that is to serve as a basis for discussion at the third DRI conference.
The HCRC approach starts by identifying specific *initiating* dialogue acts, called *moves*, such as instructions, explanations, etc., taking them as the starting point for *(conversational) games*. Those games, in turn, then encompass all functional utterances up to the point where the purpose specified by the initiating act has either been fulfilled or is abandoned (see Carletta et al. 1995, p. 3).
In contrast to this, the approach suggested by Nakatani and Traum (1998) groups functional utterances according to *Common Ground Units (CGUs)*, which, at a more abstract level, represent all those units that are relevant to developing mutual understanding of the participants. CGUs may be cancelled, modified or corrected in retrospect.
Both schemes are based on initiating elements and responses to them and allow for nesting of games/CGUs within other units continued at a later stage. However, the main difference, and potential danger, in the latter scheme is that it also allows for explicit exclusion of functional utterances like 'self-talk' which are deemed as being irrelevant for the dialogue. We suggest, however, that no such elements be excluded until a later stage of the analysis: elements can always be 'flagged' or tagged as being irrelevant and consequently be ignored, but only when it has been firmly established that they actually are irrelevant.

1.8.1.6.3 Macro-level annotation

Macro-level annotation is concerned with identifying higher-order structures immediately below the level of the actual dialogue. In order to illustrate it, we shall be referring to the same two approaches as for meso-level annotation.

[33] "http://www.cs.umd.edu/users/traum/DSD/mtman.ps"

After having established games at meso-level, the Map Task approach groups those games into *transactions*, encompassing sub-dialogues that represent the achievement of one major step in the task.
The Nakatani and Traum scheme, again, seeks to capture relations between CGUs at a more abstract level by grouping them into *I-Units*. The 'I' in this term may stand for either '*informational*' or '*intentional*'. However, there seems to have been some controversy at the Chiba workshop as to how to encode CGUs in general and especially as to the usefulness of I-Units.
The VERBMOBIL scheme of functional annotation for negotiative telephone calls (Alexandersson et al. 1997) does not include a meso-level, but has a macro-level consisting of the following phases of the dialogue:

1. H – Hello
2. O – Opening
3. N – Negotiation
4. C – Closing
5. G – Goodbye

This is the canonical ordering of the phases, but some variation is allowed for.

1.8.1.7 Techniques for identifying dialogue acts or topics

In this section, we give a brief outline of some of the techniques that may be used for the automatic or manual segmentation of dialogues into dialogue acts or the identification of topics. As above, this cannot be an exhaustive account of all the possibilities, as some may depend heavily on the nature of individual tasks.

1.8.1.7.1 Techniques for identifying dialogue acts

As already indicated above (see Section 1.8.1.3), segmentation of dialogues into individual utterances is mainly performed by looking at a combination of syntactic clues, pauses and intonational information. It is clear that syntax is important, e.g. in signalling questions, but often in dialogues such syntactic clues are absent, and reliance has to be placed on lexical and prosodic information. Below, we shall give a short, very tentative list of items that may signal certain functional categories, and (where possible) point out how they may be disambiguated by taking intonational clues into account. Note that our reference to certain intonational features here only represents a set of rough-and-ready guidelines that may help identifying functional categories, mainly when there is no additional prosodic information available.

(1) Discourse particles /conjunctions/(linking) adverbs:

- *okay, alright, yes, yeah, right, good*, etc. in sentence-initial position,
- *but, however*, etc. as adversatives introducing possible disagreement,
- *well, maybe*, etc. as unclear or expressing doubt or reservation.

(2) Syntactically unmarked questions[34] used as:

[34] Note that even though questions in RP (Received Pronunciation) and many other dialects and languages are generally intuitively assumed to end in a rise, this does not always have to be the case and may depend on the speaker's intentions and status. For further detail, see Knowles (1987).

- request for information + rise; e.g. *Next Monday?*
- suggestions + rise; e.g. *Tuesday okay?*

(3) Commands & instructions: the former may be indicated by a strong falling intonation and added stress, and the latter will also, in many cases, exhibit a fall.[35]
(4) Negation (in responses): *no, not, don't*, etc. as indicators of possible rejection or disagreement or emphatic agreement, i.e. *not at all*, etc.
(5) Repetition ('uptake') of previous speaker's wording as:

- stating agreement,
- a request for clarification in questions,
- an expression of possible disagreement/incredulity in 'echo' questions with a strong rise.

(6) Backchanneling: *hm, yes, right, I see*, etc. + rise-fall + lengthening.
(7) Openers: *hello, hi*, etc. + (usually) rising intonation.
(8) Closing 'tags': *bye, goodbye*, etc. + fall-rise or fall on monosyllabic words.

While these examples may work for some varieties of English (and possibly for some other European languages as well), one has to bear in mind that they would probably need to be adapted for many other languages and indeed for other accents of English.
Techniques of this kind are used extensively in the VERBMOBIL project, especially with regard to discourse particles and sentence boundaries that are automatically disambiguated prosodically before the actual analysis of dialogue acts is undertaken (see Alexandersson et al. (1997), p. 71 ff, and Niemann et al. (1997b); Batliner et al. (1997).[36]

1.8.1.7.2 Techniques for identifying topics

Identification of dialogue acts may depend considerably on recognising the topic or domain being discussed in a particular part of the dialogue. Techniques for identifying topics, sometimes also referred to as *topic spotting*, rely heavily upon word-spotting, as well as knowledge about both the task and the domain (see Section 1.2). Once sufficient information is available about individual dialogue acts that may form the building blocks for task-specific interactions (games) and frequently occurring or topic-specific words are identified, a list of 'closed-class' items can be created. Based on this list, the dialogue can be analysed and (probable) functional utterance tags can be assigned automatically.
One possible way of arriving at such a list is creating a concordance of keywords and listing them according to their frequency after having eliminated non-topic-specific high-frequency words like articles, etc. by means of a stop-list. In the domain of travel arrangements, for example, likely candidates for such a topic list are place-names, means of transport, references to dates, time adverbials, etc.

[35] However, expression, especially of the latter, may be highly dependent on the domain. For example, instructions that take the form of long lists as in the Map Task corpus may well end on a high tone as signals of non-finality.
[36] For more information see our Section 1.7.1.5.

In fully computer-based systems like the VERBMOBIL system, topic spotting may be performed at either the word or the sub-word level (see Niemann et al. 1997a, for more detail).

1.8.1.8 Evaluation of coding schemes

In order to assess the validity of coding schemes for dialogue annotation, researchers have in the past looked at inter-rater consistency. However, the notion of being able to evaluate such schemes in this way, without taking the amount of chance agreement into account, is being increasingly challenged:

> Research was judged according to whether or not the reader found the explanation plausible. Now, researchers are beginning to require evidence that people besides the authors themselves can understand and make the judgements undelying the research reliably. This is a reasonable requirement because if researchers can't even show that different people can agree about the judgements on which their research is based, then there is no chance of replicating the research results. (Carletta 1996, p. 1)

Based on experiences in research in experimental psychology, there is now an increasing tendency amongst researchers to try and take the element of chance agreement into account by computing the *kappa coefficient* (see Carletta 1996; Flammia and Zue 1995, for inter-rater agreement):

> The kappa coefficient (K) measures pairwise agreement among a set of coders making category judgements, correcting for expected chance agreement.
>
> $$K = \frac{P(A) - P(E)}{1 - P(E)}$$
>
> where $P\ (A)$ is the proportion of times that the coders agree and $P\ (E)$ is the proportion of times that we would expect them to agree by chance, ... When there is no agreement other than that which would be expected by chance K is zero. When there is total agreement, K is one. (Carletta 1996, p. 4)

For detailed information on how to compute the K coefficient, see Siegel and Castellan Jr. (1988), pp. 284–291.
But even if computing the K coefficient somewhat 'objectifies' determining the validity of individual coding schemes, it still remains difficult to compare the efficiency and reliability of different sets of schemes:

> ..., although kappa addresses many of the problems we have been struggling with as a field, in order to compare K across studies, the underlying assumptions governing the calculation of chance expected agreement still require the units over which coding is performed to be chosen sensibly and comparably. (Carletta 1996, p. 4)

1.8.1.9 Annotation tools and general coding recommendations

As already indicated above, most projects in dialogue annotation make use of some form of annotation tool. Below, we shall give a brief list of some of the existing tools, some of which are freely available and can be downloaded from the respective web-sites. One thing that nearly all of them have in common is that they can produce fully, or at least partly, SGML-conformant output files.

(1) Flammia's Nb

- available from: "`http://sls-www.lcs.mit.edu/flammia/Nb.html`"
- status: free
- output format: similar to SGML; can be converted with supplied Perl script nb2sgml.pl
- platform(s): Unix and Windows95/NT
- other requirements: tcl/tk; Perl; platform must support long filenames

(2) dat (TRAINS)

- available from: "`http://www.cs.rochester.edu/research/trains/annotation/`"
- status: free
- output format: SGML
- platform(s): Windows and Unix
- other requirements: Perl 5.003, Perl Tk package, Perl FileDialog widget

(3) Python-based tools (Map Task)

- available from: "`http://www.ltg.ed.ac.uk/software/`"[37]
- status: free
- output format: SGML
- platform(s): Windows and Unix
- other requirements: Python

(4) Alembic Workbench:

- available from: "`http://www.mitre.org/cgi-bin/get_alembic/`"
- status: free
- output format: SGML
- platform(s): SunOS/Solaris, Linux, Windows95/NT, Macintosh[38]
- other requirements: none

(5) LT XML:

- available from: "`http://www.ltg.ed.ac.uk/software/xml/`"
- status: free
- output format: XML
- platform(s): Windows95/NT, Linux, various other flavours of Unix, Macintosh
- other requirements: none

(6) XED:

- available from: "`http://www.cogsci.ed.ac.uk/~ht/xed.html`"
- status: free for educational/research purposes; commercial licence also available

[37] At the time of writing, the tools were not actually available yet, but supposed to become available in the near future.

[38] Not yet available at the time of writing.

- output format: XML
- platform(s): Windows95/NT, Solaris
- other requirements: based on LT XML

(7) Speech Analyser/Speech Manager :

- available from: "`http://www.jaars.org/icts/software.html`"
- status: free
- output format: Wave file/MS Access^{TM} database
- platform(s): Windows 3.11/95/NT
- other requirements: none

While items (1)–(4) are graphical user interfaces, LT XML is a set of pre- and post-processing tools for handling XML documents. However, XED can be used to set up and manipulate those documents interactively, although it is more of a text editor than (1)–(4).
Item (7) does not fall into any of the above categories. It differs from the rest of the tools described here in at least two respects: for one thing, it was originally designed as a tool for phonetic analysis, rather than specifically for the annotation of corpora.
For the other, it does not actually produce any SGML or XML compatible output, but annotations are first written into the wave file by the Speech Analyser and may then be extracted by the Speech Manager and stored in a relational database. However, a relational database presents a highly efficient mechanism for storing and analysing data, and it would easily be possible to create SGML or XML annotated output from within the database.
The above selection of available tools shows that nowadays it should be no problem to create annotated dialogue material that is SGML- or even XML-encoded. The major obvious advantage of such an approach is that markup languages make it easy to separate form from content during the annotation. In other words, it should be(come) possible to annotate one's data according to functional criteria and then leave it up to the software to group and display categories according to the requirements of the (research) purpose. One added advantage is that additional items of information can easily be incorporated by making use of hyperlinking facilities. A very good example of how such an approach can be put to good use is the HCRC web-interface to the Map Task Corpus,[39] which allows the user to look at individual turns produced by each speaker and to play them back across the network.
As far as tools are concerned, though, one thing does remain a problem. Even though some of the tools already allow one to play back parts of dialogues associated with individual utterances, there are still only very few publicly available tools (apart from tools such as the above mentioned Speech Analyser) that actually allow the transcriber/annotator to look at prosodic information from within the annotation tool. Therefore we still have no way of making use of all the available parameters needed to extract information relevant to the interpretation of the dialogue.

1.8.2 Recommendations

(1) Always try to separate form from content, at least conceptually.

[39] "`http://www.hcrc.ed.ac.uk/dialogue/public_maptask/`"

(2) Try to integrate as many levels of analysis as possible, preferably all.
(3) Do not exclude any information at an early stage as it might prove relevant later.
(4) Code your dialogues so that they are exchangeable and allow different users to create different views of them, preferably in SGML or, better yet, XML. XML is on its way to becoming a standard, as already major software producers such as Microsoft and IBM have committed themselves to providing support for it in the future.

Appendix A: TEI paralinguistic features

Tempo

- fast
- very fast
- getting faster
- slow
- very slow
- getting slower

Loudness

- loud
- very loud
- getting louder
- soft
- very soft
- getting softer

Pitch range

- high pitch range
- low pitch range
- wide pitch range
- narrow pitch range
- ascending
- descending
- monotonous
- scandent (each successive syllable higher than the last, generally ending in a falling tone)

Tension

- slurred
- lax, a little slurred
- tense
- very precise
- staccato, every stressed syllable doubly stressed
- legato, every stressed syllable more-or-less equally stressed

Rhythm

- beatable rhythm
- arhythmic, particularly halting
- spiky rising, with markedly higher unstressed syllables

- spiky falling, with markedly lower unstressed syllables
- glissando rising, like spiky rising but the unstressed syllables also rise in pitch relative to each other
- glissando falling, like spiky falling but the unstressed syllables also fall in pitch relative to each other

Voice quality

- whisper
- breathy
- husky
- creaky
- falsetto
- resonant
- unvoiced laugh or giggle
- voiced laugh
- tremulous
- sobbing
- yawning
- sighing

Appendix B: TEI P3 DTD: base tag set for transcribed speech

```
<!-- teispok2.dtd:  written by OddDTD 1994-09-09             -->

<!-- 11:  Base tag set for Transcribed Speech                -->
<!-- Text Encoding Initiative: Guidelines for Electronic     -->
<!-- Text Encoding and Interchange. Document TEI P3, 1994.   -->
<!-- Copyright (c) 1994 ACH, ACL, ALLC. Permission to copy   -->
<!-- in any form is granted, provided this notice is         -->
<!-- included in all copies.                                 -->
<!-- These materials may not be altered; modifications to    -->
<!-- these DTDs should be performed as specified in the      -->
<!-- Guidelines in chapter "Modifying the TEI DTD."          -->
<!-- These materials subject to revision. Current versions   -->
<!-- are available from the Text Encoding Initiative.        -->
<!-- 11.2.7:  Components of Transcribed Speech               -->
<!ENTITY % u 'INCLUDE' >
<![ %u; [
<!ELEMENT %n.u;          - -  ((%phrase | %m.comp.spoken)+)      >
<!ATTLIST %n.u;               %a.global;
                              %a.timed;
          trans               (smooth | latching | overlap |
                              pause)              smooth
          who                 IDREF               %INHERITED
          TEIform             CDATA               'u'           >
]]>

<!ENTITY % pause 'INCLUDE' >
<![%pause; [
<!ELEMENT %n.pause;      - O  EMPTY                             >
<!ATTLIST %n.pause;           %a.global;
                             %a.timed;
```

```
          type               CDATA              #IMPLIED
          who                IDREF              #IMPLIED
          TEIform            CDATA              'pause'        >
]]>

<!ENTITY % vocal 'INCLUDE' >
<![ %vocal; [
<!ELEMENT %n.vocal;     - O  EMPTY                             >
<!ATTLIST %n.vocal;          %a.global;
                             %a.timed;
          who                IDREF              %INHERITED
          iterated           (y | n | u)        n
          desc               CDATA              #IMPLIED
          TEIform            CDATA              'vocal'        >
]]>

<!ENTITY \% kinesic 'INCLUDE' >
<![ %kinesic; [
<!ELEMENT %n.kinesic;   - O  EMPTY                             >
<!ATTLIST %n.kinesic;        %a.global;
                             %a.timed;
          who                IDREF              %INHERITED
          iterated           (y | n | u)        n
          desc               CDATA              #IMPLIED
          TEIform            CDATA              'kinesic'      >
]]>

<!ENTITY % event 'INCLUDE' >
<![ %event; [
<!ELEMENT %n.event;     - O  EMPTY                             >
<!ATTLIST %n.event;          %a.global;
                             %a.timed;
          who                IDREF              %INHERITED
          iterated           (y | n | u)        n
          desc               CDATA              #IMPLIED
          TEIform            CDATA              'event'        >
]]>

<!ENTITY % writing 'INCLUDE' >
<![ %writing; [
<!ELEMENT %n.writing;   - -  (%paraContent;)                   >
<!ATTLIST %n.writing;        %a.global;
          who                IDREF              %INHERITED
          type               CDATA              #IMPLIED
          script             IDREF              #IMPLIED
          gradual            (y | n | u)        #IMPLIED
          TEIform            CDATA              'writing'      >
]]>

<!ENTITY % shift 'INCLUDE' >
<![ %shift; [
<!ELEMENT %n.shift;     - O  EMPTY                             >
```

```
<!ATTLIST %n.shift;          %a.global;
          who                IDREF               #IMPLIED
          feature            (tempo | loud | pitch | tension |
                             rhythm | voice)     #REQUIRED
          new                CDATA               normal
          TEIform            CDATA               'shift'        >
]]>

<!-- (end of 11.2.7)                                            -->
<!-- The base tag set for transcriptions of speech uses the     -->
<!-- standard default text-structure elements, which are        -->
<!-- embedded here:                                             -->
<![ %TEI.singleBase [
<!ENTITY % TEI.structure.dtd system 'teistr2.dtd'                 >
%TEI.structure.dtd;
]]>
<!-- (end of 11)                                                -->
```

Appendix C: A few relevant web links

WP4 pages at Lancaster:
"http://www.ling.lancs.ac.uk/eagles/"
EAGLES SLWG telecooperation facilities:
"http://coral.lili.uni-bielefeld.de/EAGLES/SLWG/"
VERBMOBIL:
"http://www.phonetik.uni-muenchen.de/"
"http://www.phonetik.uni-muenchen.de/VMtrlex2d.html"
MAPTASK:
"http://www.cogsci.ed.ac.uk/hcrc/wgs/dialogue/dialog/maptask.html"
NFS Interactive Systems Grantees' Workshop:
"http://www.cse.ogi.edu/CSLU/isgw97/reports.html"
CHRISTINE project:
"http://www.cogs.susx.ac.uk/users/geoffs/RChristine.html"
Discourse Resource Initiative:
"http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html"
A corpus of Swedish dialogues:
"http://www.ida.liu.se/~nlplab/dialogues/corpora.html"
TRAINS:
"http://www.cs.rochester.edu/research/speech/dialogues.html"
Spoken language project at Gothenburg:
"http://www.ling.gu.se/~sylvana/SLSA/"

Appendix D: Specimen Annotated Dialogue

The selection of a dialogue extract suitable for illustrating the annotation guidelines had to meet a rather strict set of criteria, namely:

(a) One dialogue only could be attempted, because of the amount of work involved,

and the limited time constraint. It would have clearly been an advantage to provide samples in a number of European languages, but this was not feasible.
(b) The sound files should be available, and capable of being consulted by users.
(c) The standard of recording should be good.
(d) As (b) implies, the dialogue extract should be in the public domain, and should not suffer from copyright or confidentiality restrictions.
(e) Being a single illustrative extract, it should be in a language generally understood throughout the European Union.
(f) It should be a task-defined applications-oriented dialogue, of a kind directly relevant to the development of speech systems applications in the EU.

The possibility of using bilingual dialogue was investigated, but was not found to be practicable, bearing in the mind the above requirements.
In the end, a piece of dialogue was found which met all these criteria, and which emanated from a European project (Verbmobil) concerned with multilingual dialogue (speech to speech translation support). However, the recording was made in the USA and the dialogue was in American English. The details of the dialogue are given in the present draft of the document.
The following illustrations show various levels of dialogue representation and annotation, using a single specimen dialogue in English originating from the German VERBMOBIL programme (Dialogue r148c).[40] The illustrations which follow the aim, in the first instance, to be 'human friendly': the simplest mark-up is used, in order to demonstrate the kinds of information associated with each level. In the interests of clarity, no attempt is made at this stage to represent SGML or TEI (Text Encoding Initiative) standard encoding for all of the levels. The five levels are: A. Orthographic transcription; B. Morphosyntactic annotation; C. Syntactic annotation; D. Prosodic annotation; E. Pragmatic (dialogue act) annotation. Section F., however, is more advanced and complex, in showing (a) the combination of different levels of annotation, and (b) the use of SGML/XML as an encoding standard. In each version, the same dialogue is used, although not to the same degree of completeness. Also, each version is preceded and/or followed by explanatory notes and/or lists of symbols.
Note that while these transcriptions and annotations will hopefully provide useful illustrations, they are not intended as a general model to be followed by dialogue corpus compilers. At the levels both of linguistic categorisation and of encoding, there are many decisions to be made which cannot be pre- empted here, depending on such factors as the language represented and the purpose for which the annotation is required.

D.1: Orthographic Transcription

```
<dialog>
<A> so . we should meet again . how 'bout . how 'bout next week .
    what day are good for you . what days are good for you .
<B> actually next week I am on vacation .
<A> gosh . I guess we will have to meet the week after that .
    how 'bout Monday .
<B> Monday the tenth .
<A> aha .
```

[40] This dialogue can be obtained via anonymous ftp at the following address: "ftp.cs.rochester.edu/pub/packages/dialog-annotation/r148c.tar.gz".

```
<B> well unfortunately my vacation runs through the fourteenth
    and I have nonrefundable plane tickets . I was planning on
    being on a beach in Acapulco about that point .
<A> well . when are you getting back .
<B> I get back on the fifteenth rest up on the sixteenth . which is a
    Sunday . and I am back at work on the seventeenth . but I have
    a seminar all day . I think the first day that is really good for me .
    is the eighteenth that is a Tuesday .
<A> okay . want to have lunch .
<B> that sounds pretty good . are you available just before noon .
<A> we can meet at noon .
<B> sounds good . on campus or off .
<A> your choice .
<B> I say if I have got enough money to go to one of those silly places
    on Craig Street . how about Great Scott .
<A> sounds great except they have been out of business for a while .
    how about some other place . let us just wander up Craig . and pick
    one we like that day .
<B> that sounds pretty good . okay . I will meet you outside Cyert Hall .
    at noon . does that sound alright for you .
<A> see you then .
<last three turns omitted>
</dialog>
```

This is the simplest possible orthographic transcription, showing turns and minimal punctuation into 'orthographic sentences'. Contractions are represented as full, uncontracted forms: e.g. *let* us for *let's*, *I will* for *I'll*.

D.2: Morphosyntactic annotation

The morphosyntactic (POS) tags are here shown attached by the underline symbol _ to the words that they label. In SGML, the tag can be represented as follows:

```
<w AVC>so <w Ppp1N>we <w VM>should <w VVI>meet <w AV>again.
```

The tagset used for this annotation is very closely modelled on the EAGLES reduced illustrative tagset for English as given in Leech and Wilson (1994) and Monachini and Calzolari (1996). In Table 1.6 brief definitions of the tags used above are given. Among the tags listed in the Table, AJC, IJR and IJX are new tags introduced to handle phenomena of spoken language.

```
<A>  <utt1> so_AVC
     <utt2> we_PPp1N should_VM meet_VVI again_AV
     <utt3> how_AVWQ 'bout_APR
     <utt4> how_AVWQ 'bout_APR next_AJ week_NCs
     <utt5> what_DWQ day_NCs are_VVR good_AJ for_APR you_PP2
     <utt6> what_DWQ days_NCp are_VVR good_AJ for_APR you_PP2
<B>  <utt7> actually_AV next_AJ week_NCs I_PPs1N am_VVM on_APR vacation_NCs
<A>  <utt8> gosh_IJX
     <utt9> I_PPs1N guess_VVB we_PPp1N will_VM have_VVI to_UI meet_VVI the_ATD
         next_AJ week_NCs after_APR that_PDs
     <utt10> how_AVWQ 'bout_APR Monday_NPs
<B>  <utt11> Monday_NPs the_ATD tenth_NUOs
<A>  <utt12> aha_IJR
<B>  <utt13> well_AVC unfortunately_AV my_DVs1 vacation_NCs runs_VVZ
         through_APR the_ATD fourteenth_NUOs and_CC I_PPs1N have_VVB
         nonrefunadable_AJ plane_NCs tickets_NCp
     <utt14> I_PPs1N was_VPDZ planning_VVG on_APR being_VVG on_APR a_ATIs
         beach_NCs in_APR Acapulco_NPs about_APR that_DDs point_NCs
```

```
<A>  <utt15> well_AVC
     <utt16> when_AVWQ are_VPR you_PP2 getting_VVG back_AVP
<B>  <utt17> I_PPs1N get_VVB back_AVP on_APR the_ATD fifteenth_NUOs
         rest_VVB up_AVP on_APR the_ATD sixteenth_NUOs
     <utt18> which_PWR is_VVZ a_AT1s Sunday_NPs
         and_CC I_PPs1N am_VVM back_AVP at_APR work_NCs on_APR the_ATD
         seventeenth_NUOs
     <utt19> but_CC I_PPs1N have_VVB a_ATIs seminar_NCs all_DI day_NCs
     <utt20> I_PPs1N think_VVB the_ATD first_NUOs day_NCs that_PWR is_VVZ
         really_AV good_AJ for_APR me_PP1sO
     <utt21> is_VVZ the_ATD eighteenth_NUOs that_PDs is_VVZ a_ATIs Tuesday_NPs
<A>  <utt22> okay_IJR
     <utt23> want_VVI to_UI have_VVI lunch_NCs
<B>  <utt24> that_DDs sounds_VVZ pretty_AVD good_AJ
     <utt25> are_VVR you_PP2 available_AJ just_AV before_APR noon_NCs
<A>  <utt26> we_PPp1N can_VM meet_VVI at_APR noon_NCs
<B>  <utt27> sounds_VVZ good_AJ
     <utt28> on_APR campus_NCs or_CC off_APR
<A>  <utt29> your_DV2 choice_NCs
<B>  <utt30> I_PPs1N say_VVB if_CSF I_PPs1N have_VPB got_VVN enough_DI
         money_NCs to_UI go_VVI to_APR one_NUCs of_APR those_DDp silly_AJ
         places_NCp on_APR Craig_NPs Street_NCs
     <utt31> how_AVWQ about_APR Great_AJ Scott_NPs
<A>  <utt32> sounds_VVZ great_AJ except_CSF they_PPp3N have_VPB been_VVN out
         of_APR business_NCs for_APR a_ATIs while_NCs
     <utt33> how_AVWQ about_APR some_DI other_AJ place_NCs
     <utt34> let_VVB us_PPp1O just_AV wander_VVI up_APR Craig_NPs
     <utt35> and_CC pick_VVI one_PIs we_PPp1N like_VVB that_DDs day_NCs
<B>  <utt36> that_PDs sounds_VVZ pretty_AVD good_AJ
     <utt37> okay_IJR
     <utt38> I_PPs1N will_VM meet_VVI you_PP2 outside_APR Cyert_NPs Hall_NCs
     <utt39> at_APR noon_NCs
     <utt40> does_VPZ that_PDs sound_VVI alright_AV for_APR you_PP2
<A>  <utt41> see_VVI you_PP2 then_AV
     <utt42-utt44 omitted>
</dialog>
```

D.3: Syntactic annotation

The following is a sample of syntactic annotation, using the simple form of labelled bracketing mark-up according to the EAGLES illustrative scheme in Leech et al. (1996). Although morphosyntactic annotation is usually included with syntactic annotation, in this example, for clarity, the morphosyntactic tags have been omitted, since they have already been shown in B. above.

```
<A> [S so . [NP we NP][VP should meet again VP] . S]
    [S[ADVP how ADVP][PP 'bout # PP]S]
    [S[ADVP how ADVP][PP 'bout [NP next week NP]PP] . S]
    [S[NP what day NP][VP are [ADJP good [PP for [NP you NP]PP]ADJP]VP] . S]
    [S[NP what days NP][VP are [ADJP good [PP for [NP you NP]PP]ADJP]VP] . S]
<B> [S[ADVP actually ADVP][NP next week NP][NP I NP][VP am [PP on [NP vacation
    NP]PP]VP] . S]
<A> [S gosh . S]
    [S[NP I NP][VP guess [CL-Nom[NP we NP][VP will have to meet [NP the next
    week [PP after [NP that NP]PP]NP]VP]CL-Nom]VP] . S]
    [S[ADVP how ADVP][PP 'bout [NP Monday NP]PP] . S]
<B> [S[NP Monday [NP the tenth NP]NP] . S]
<A> [S aha . S]
<B> [S well [ADVP unfortunately ADVP][S[S-Co[NP my vacation NP][VP runs [PP
```

Table 1.6: Tag definitions

AJ	(Pos.) adj., general	PDs	Sing. demonstr. pron.
APR	Prep.	PDp	Pl. demonstr. pron.
ATD	Def. article	PIs	Indef. pron. singular
ATIs	Indef. article	PPs1N	Pers. pron. 1 pers. sg. nom.
AV	(Pos.) adv., general	PPs1O	Pers. pron. 1 pers. sg. obl.
AVC	Discoursal adv.	PP2	Pers. pron. 2 pers.
AVD	Adv. of degree	PPp1N	Pers. pron. 1 pers. pl. nom.
AVP	Adv. particle	PPp1O	Pers. pron. 1 pers. pl. obl.
AVWQ	General adv., interrog. wh-type	PPp3N	Pers. pron. 3 pers. pl. nom.
CC	Coord. conj.	PWR	Wh-pronoun, rel.
CSF	Subord. conj., finite	UI	Inf. marker
DDs	Sing. demonstr. det.	VM	Modal aux. v.
DDp	Pl. demonstr. det.	VPB	Fin. base form primary aux.
DI	Indef. determiner	VPDZ	Past -s form, primary aux.
DVs1	Poss. det, 1st pers. sg.	VPI	Inf. primary aux. v.
DV2	Poss. det., 2nd person	VPR	Pres. -re form, prim. aux.
DWQ	Interrog. wh-det.	VPZ	Pres. -s form, prim. aux.
IJR	Interj.: response form	VVB	Fin. base form, main v.
IJX	Interj.: exclamatory	VVG	-ing form, main v.
NCs	Sing. common noun	VVI	Inf. main v.
NCp	Pl. common noun	VVM	Pres. 1 pers. sg. main v.
NPs	Sing. proper noun	VVN	Past part. main v.
NUCs	Sing. card. numeral	VVR	Pres. -re form main v.
NUOs	Sing. ord. numeral	VVZ	Pres. -s form main v.

```
    through [NP the fourteenth NP]PP]VP]S-Co] and [S-Co[NP I NP][VP have [NP
    nonrefundable plane tickets NP]VP]S-Co]S] . S]
    [S[NP I NP][VP was planning [PP on [CL-Nom[VP being [PP on [NP a beach [PP
    in [NP Acapulco NP]PP]NP]PP][PP about [NP that point NP]PP]VP]CL-
    Nom]PP]VP] . S]
<A> [S well . [ADVP when ADVP][*1][VP are [NP you NP*1] getting back VP] . S]
<B> [S[NP I NP][VP get back [PP on [NP the fifteenth NP]PP]VP] . S]
    [S[VP rest up [PP on [NP the sixteenth . [CL-Rel[NP which NP][VP is [NP a
    Sunday NP]VP]CL-Rel]NP]PP]VP] . S]
    [S and [NP I NP][VP am back [PP at [NP work NP]PP][PP on [NP the
    seventeenth NP]PP]VP] . S]
    [S but [NP I NP][VP have [NP a seminar NP][NP all day NP]VP] . S]
    [S[NP I NP][VP think [CL-Nom[NP the first day [CL-Rel[NP that NP][VP is
    [ADJP really good [PP for [NP me NP]PP]ADJP]VP]CL-Rel]NP] . [VP is [NP the
    eighteenth NP]VP]CL-Nom]VP]S]
    [S[NP that NP][VP is [NP a Tuesday NP]VP] . S]
<last 11 turns omitted>
```

The symbols used for higher (non-terminal) constituents are given in Table 1.7. Other symbols used are *1 (the index representing the location of discontinuity and trace phenomena) and # (representing the location of a missing constituent in the case of dysfluency).

Table 1.7: Symbols for higher (non-terminal) constituents

ADJP	Adjective phrase
ADVP	Adverb phrase
CL	Clause
CL-Co	Coordinated clause
CL-Nom	Nominal complement clause
CL-Rel	Relative clause
NP	Noun phrase
PP	Prepositional phrase
S	This represents a sentence in written texts, but in spoken language, as in the present case, it represents a C-unit (or maximally parsable unit).
VP	Verb phrase

D.4: Prosodic Annotation

In this section, the first five exchanges of the specimen dialogue have been selected for prosodic annotation. The annotation methods chosen were ToBI, in this case English ToBI (E_ToBI,) and Tonetic Stress Marks (TSM). The E_ToBI transcriptions consist of speech waveform and F_0 contours, along with time-aligned labels on four tiers. These are provided in Figures 1.2–1.9. These are, from top to bottom in the figures: tonal, orthographic, junctural (break index), and 'miscellaneous' where, amongst other things, dysfluencies are recorded.
The TSM transcriptions involve diacritics in the line of text itself and are provided in text format below.

```
<dialog>
<A> \so || we should \meet ∘again || ↑how _bout || how ∘bout next
\week || what ∘day are \good for ∘you || what \days are ∘good for
/you ||
<B> \/actually | next ∘week I am ∘on\vacation||
<A> \gosh || I ↑\guess we will ∘have to ∘meet the ∘week after \that
|| ⁻how bout \Monday ||
<B> ∘Monday the /tenth ||
<A> ⁻aha ||
</dialog>
```

ToBI transcriptions were provided by Laura Dilley of the Speech Communications Group at MIT and Tonetic transcriptions were performed by Gerry Knowles of the University of Lancaster.

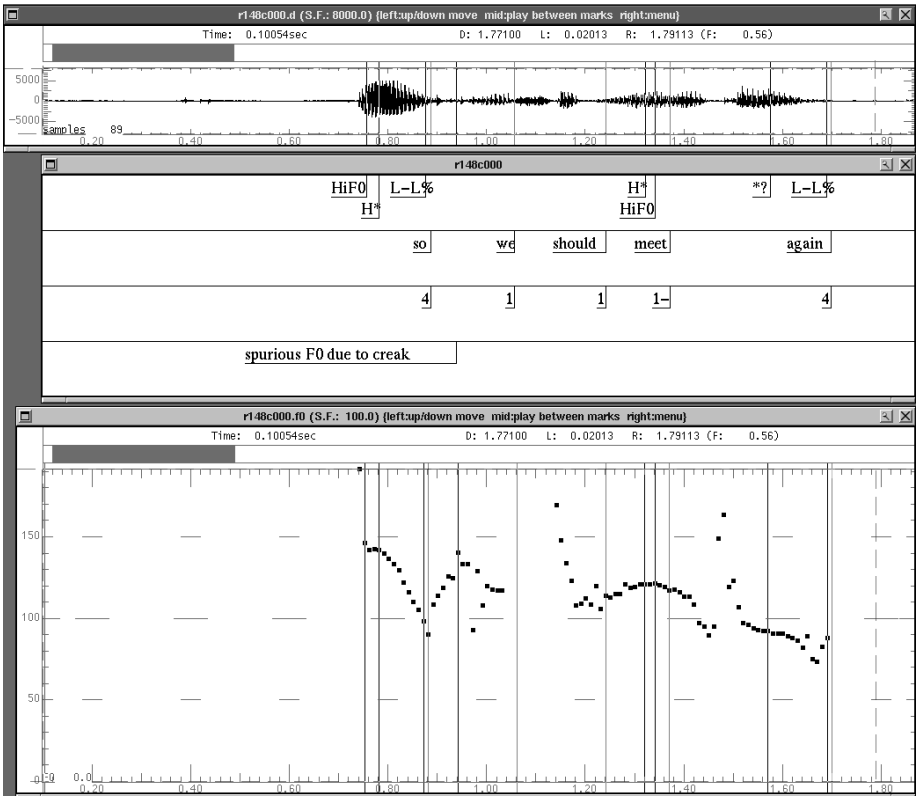

Figure 1.2: utt1 & utt2

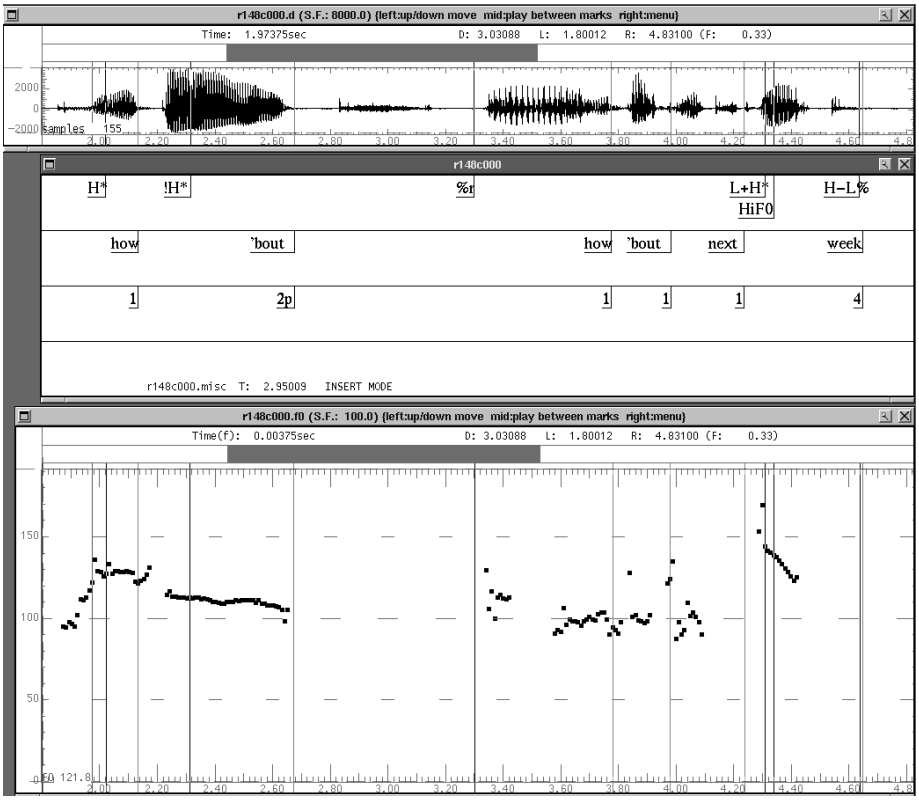

Figure 1.3: utt3 & utt4

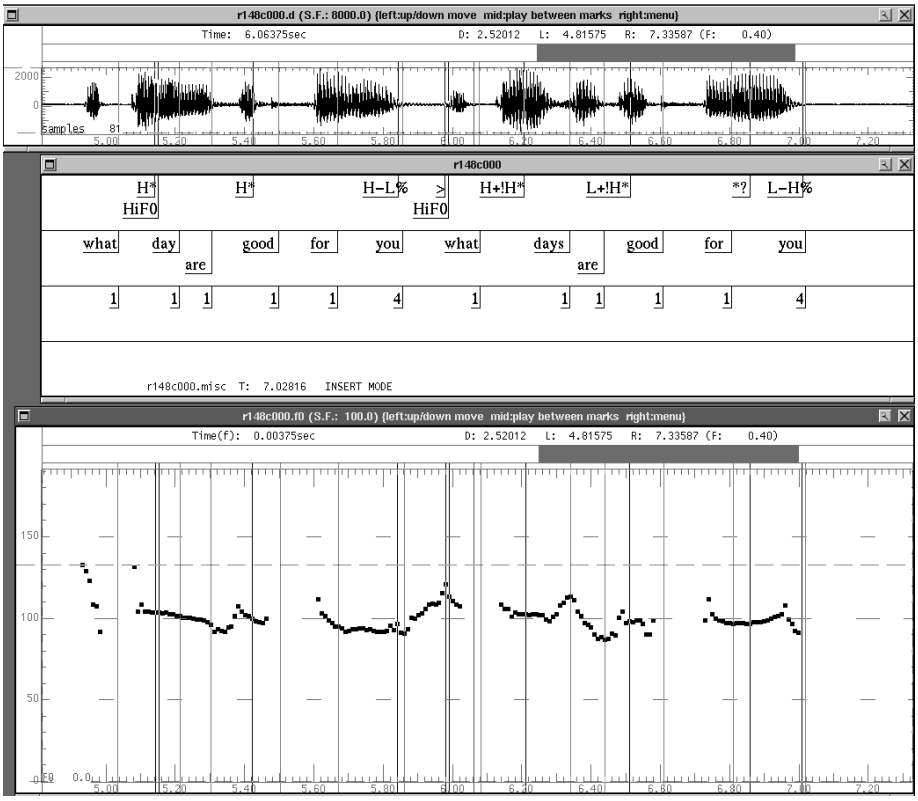

Figure 1.4: utt5 & utt6

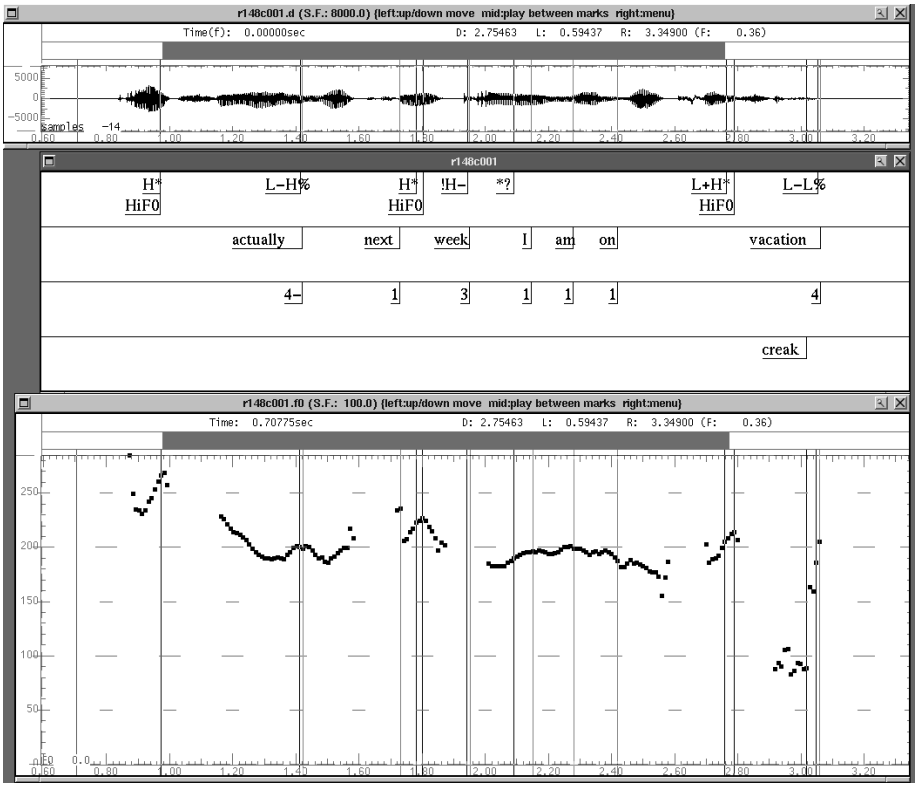

Figure 1.5: utt7

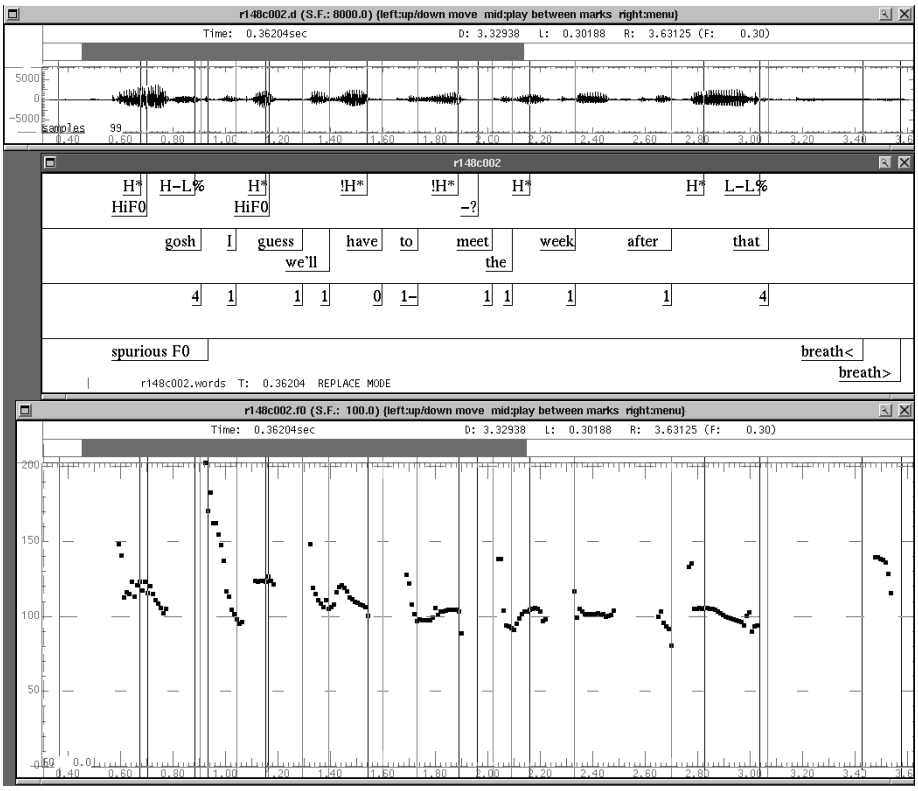

Figure 1.6: utt8 & utt9

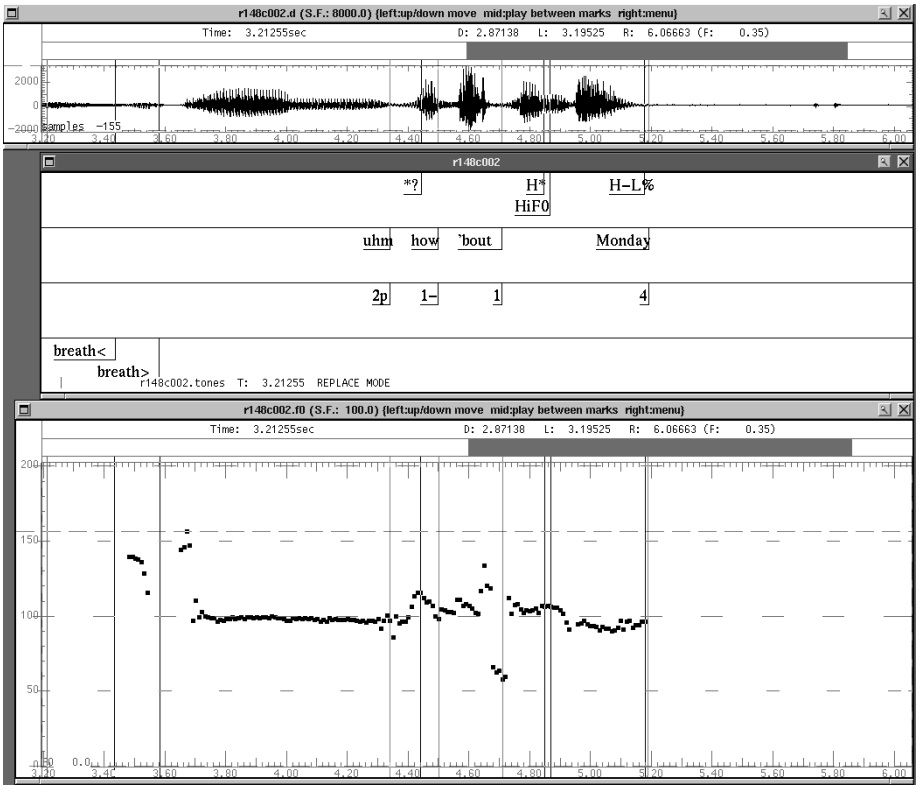

Figure 1.7: utt10

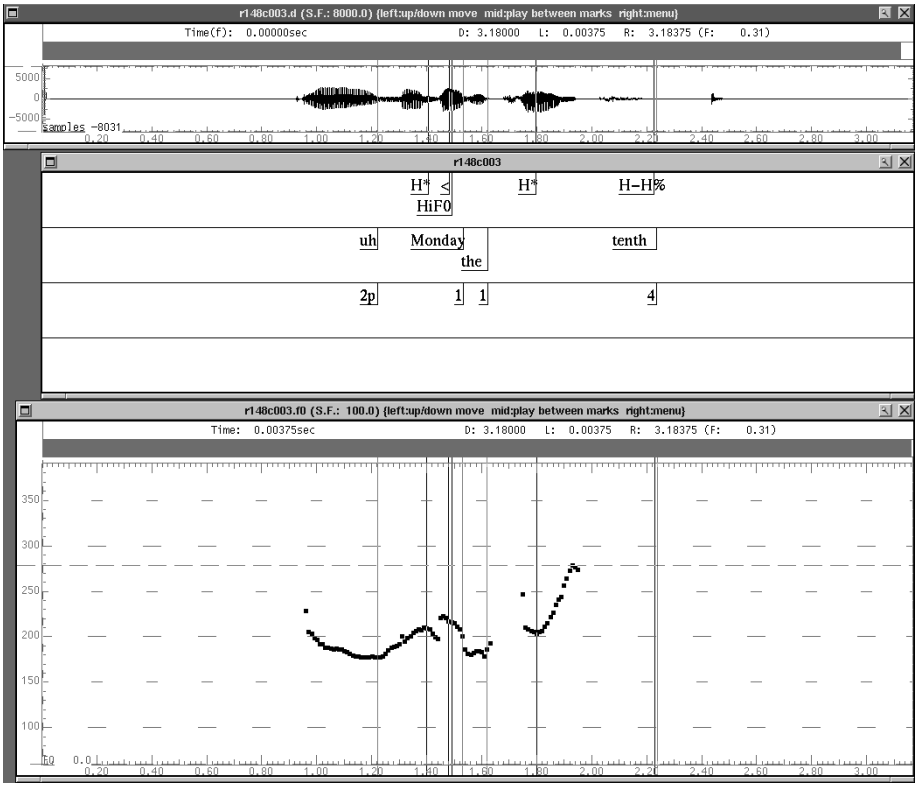

Figure 1.8: utt11

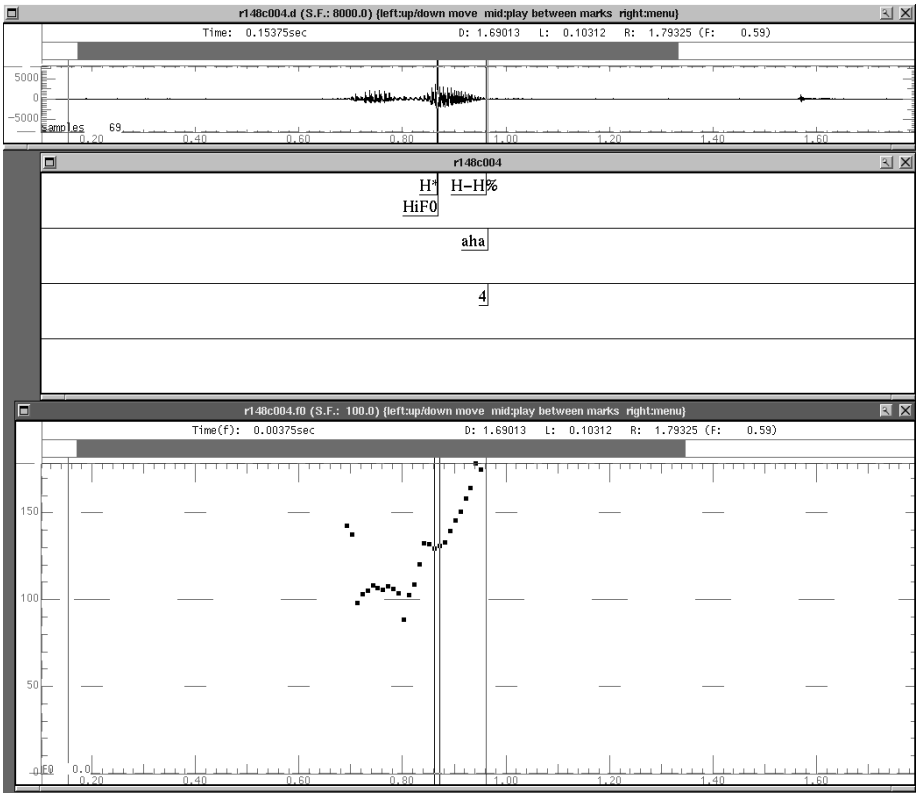

Figure 1.9: utt12

D.5: Pragmatic (Dialogue Act) Annotation

The following constitutes an example of two different ways in which pragmatic information in dialogues can be represented.
The first, and more comprehensive one, is loosely modelled on the dialogue act annotation scheme developed by the participants in the VERBMOBIL project. However, neither the tagset used for our annotation, nor the SGML/XML-conformant type of markup are actually used by any of the partners involved in the VERBMOBIL project data collection, annotation or processing. The second, abridged, example represents the kind of coding scheme developed by the DRI and actually constitutes part of a presentation that was made at the 3rd annual DRI workshop in Chiba.

```
<DIALOG ID="r148c" STATUS="preprocessed">
<TURN ID="t1" SPEAKER="A">
<UTT ID="utt1" SPEECH="soundfiles/r148c000.au">
<DISCOURSE_PARTICLE ID="1">so</DISCOURSE_PARTICLE>
      </UTT>
<UTT ID="utt2" SPEECH="soundfiles/r148c000.au">
<INIT>
<SUGGEST ID="1">we should meet again</SUGGEST>
</INIT>
      </UTT>
<UTT ID="utt3" SPEECH="soundfiles/r148c000.au">
<REQUEST_SUGGEST ID="1">
<FALSE_START ID="1">how 'bout</FALSE_START>
         </REQUEST_SUGGEST>
      </UTT>
<UTT ID="utt4" SPEECH="soundfiles/r148c000.au">
<REQUEST_SUGGEST ID="1">how 'bout next week</REQUEST_SUGGEST>
      </UTT>
<UTT ID="utt5" SPEECH="soundfiles/r148c000.au">
<REQUEST_COMMENT ID="1">
<FALSE_START ID="2">what day are good for you</FALSE_START>
         </REQUEST_COMMENT>
      </UTT>
<UTT ID="utt6" SPEECH="soundfiles/r148c000.au">
<REQUEST_COMMENT ID="1">what days are good for you</REQUEST_COMMENT>
      </UTT>
   </TURN>
<TURN ID="t2" SPEAKER="B">
<UTT ID="utt7" SPEECH="soundfiles/r148c001.au">
<EXPLAINED_REJECT ID="1">actually next week I am on vacation</EXPLAINED_REJECT>
      </UTT>
   </TURN>
<TURN ID="t3" SPEAKER="A">
<UTT ID="utt8" SPEECH="soundfiles/r148c002.au">
<FEEDBACK_INTERJECT ID="1">gosh</FEEDBACK_INTERJECT>
      </UTT>
<UTT ID="utt9" SPEECH="soundfiles/r148c002.au">
<SUGGEST ID="2">I guess we will have to meet the week after that</SUGGEST>
      </UTT>
<UTT ID="utt10" SPEECH="soundfiles/r148c002.au">
<SUGGEST ID="2">how 'bout Monday</SUGGEST>
      </UTT>
   </TURN>
<TURN ID="t4" SPEAKER="B">
<UTT ID="utt11" SPEECH="soundfiles/r148c003.au">
<CLARIFY ID="1">Monday the tenth</CLARIFY>
```

```
      </UTT>
   </TURN>
<TURN ID="t5" SPEAKER="A">
<UTT ID="utt12" SPEECH="soundfiles/r148c004.au">
<CONFIRM ID="1">aha</CONFIRM>
      </UTT>
   </TURN>
<TURN ID="t6" SPEAKER="B">
<UTT ID="utt13" SPEECH="soundfiles/r148c005.au">
<EXPLAINED_REJECT ID="2">well unfortunately my vacation runs through the
fourteenth and I have nonrefundable plane tickets</EXPLAINED_REJECT>
      </UTT>
<UTT ID="utt14" SPEECH="soundfiles/r148c005.au">
<CLARIFY ID="2">I was planning on being on a beach in Acapulco about
that point</CLARIFY>
      </UTT>
   </TURN>
<TURN ID="t7" SPEAKER="A">
<UTT ID="utt15" SPEECH="soundfiles/r148c006.au">
<DISCOURSE_PARTICLE ID="2">well</DISCOURSE_PARTICLE>
      </UTT>
<UTT ID="utt16" SPEECH="soundfiles/r148c006.au">
<REQUEST_CLARIFY ID="1">when are you getting back</REQUEST_CLARIFY>
      </UTT>
   </TURN>
<TURN ID="t8" SPEAKER="B">
<UTT ID="utt17" SPEECH="soundfiles/r148c007.au">
<CLARIFY_ANSWER ID="1">I get back on the fifteenth
rest up on the sixteenth</CLARIFY_ANSWER>
      </UTT>
<UTT ID="utt18" SPEECH="soundfiles/r148c007.au">
<CLARIFY_ANSWER ID="1">which is a Sunday and I am back at work
on the seventeenth</CLARIFY_ANSWER>
      </UTT>
<UTT ID="utt19" SPEECH="soundfiles/r148c007.au">
<CLARIFY_ANSWER ID="1">but I have a seminar all day</CLARIFY_ANSWER>
      </UTT>
<UTT ID="utt20" SPEECH="soundfiles/r148c007.au">
<SUGGEST ID="3">I think the first day that is really good for me</SUGGEST>
      </UTT>
<UTT ID="utt21" SPEECH="soundfiles/r148c007.au">
<SUGGEST ID="3">is the eighteenth</SUGGEST>
<CLARIFY ID="3">that is a Tuesday</CLARIFY>
      </UTT>
   </TURN>
<TURN ID="t9" SPEAKER="A">
<UTT ID="utt22" SPEECH="soundfiles/r148c008.au">
<ACCEPT ID="1">okay</ACCEPT>
      </UTT>
<UTT ID="utt23" SPEECH="soundfiles/r148c008.au">
<SUGGEST ID="4">want to have lunch</SUGGEST>
      </UTT>
   </TURN>
<TURN ID="t10" SPEAKER="B">
<UTT ID="utt24" SPEECH="soundfiles/r148c009.au">
<ACCEPT ID="2">that sounds pretty good</ACCEPT>
      </UTT>
<UTT ID="utt25" SPEECH="soundfiles/r148c009.au">
<SUGGEST ID="5">are you available just before noon</SUGGEST>
      </UTT>
```

```
   </TURN>
<TURN ID="t11" SPEAKER="A">
<UTT ID="utt26" SPEECH="soundfiles/r148c010.au">
<REJECT ID="1">
<SUGGEST ID="6">we can meet at noon</SUGGEST>
         </REJECT>
      </UTT>
   </TURN>
<TURN ID="t12" SPEAKER="B">
<UTT ID="utt27" SPEECH="soundfiles/r148c011.au">
<ACCEPT ID="3">sounds good</ACCEPT>
      </UTT>
<UTT ID="utt28" SPEECH="soundfiles/r148c011.au">
<REQUEST_SUGGEST ID="1">on campus or off</REQUEST_SUGGEST>
      </UTT>
   </TURN>
<TURN ID="t13" SPEAKER="A">
<UTT ID="utt29" SPEECH="soundfiles/r148c012.au">
<REQUEST_SUGGEST ID="2">your choice</REQUEST_SUGGEST>
      </UTT>
   </TURN>
<TURN ID="t14" SPEAKER="B">
<UTT ID="utt30" SPEECH="soundfiles/r148c013.au">
<DIGRESS ID="1">I say if I have got enough money to go to Acapulco
I have got enough money to go to one of those silly places on
Craig street</DIGRESS>
      </UTT>
<UTT ID="utt31" SPEECH="soundfiles/r148c013.au">
<SUGGEST ID="7">how about Great Scott</SUGGEST>
      </UTT>
   </TURN>
<TURN ID="t15" SPEAKER="A">
<UTT ID="utt32" SPEECH="soundfiles/r148c014.au">
<FEEDBACK_NEGATIVE ID="1">
            sounds great except
<GIVE_REASON ID="1">they have been out of business for a while</GIVE_REASON>
         </FEEDBACK_NEGATIVE>
      </UTT>
<UTT ID="utt33" SPEECH="soundfiles/r148c014.au">
<REQUEST_SUGGEST ID="3">how about some other place</REQUEST_SUGGEST>
      </UTT>
<UTT ID="utt34" SPEECH="soundfiles/r148c014.au">
<SUGGEST ID="8">let us just wander up Craig</SUGGEST>
      </UTT>
<UTT ID="utt35" SPEECH="soundfiles/r148c014.au">
<SUGGEST ID="8">and pick one we like that day</SUGGEST>
      </UTT>
   </TURN>
<TURN ID="t16" SPEAKER="B">
<UTT ID="utt36" SPEECH="soundfiles/r148c015.au">
<ACCEPT ID="4">that sounds pretty good</ACCEPT>
      </UTT>
<UTT ID="utt37" SPEECH="soundfiles/r148c015.au">
<DISCOURSE_PARTICLE ID="3">okay</DISCOURSE_PARTICLE>
      </UTT>
<UTT ID="utt38" SPEECH="soundfiles/r148c015.au">
<SUGGEST ID="9">I will meet you outside Cyert Hall</SUGGEST>
      </UTT>
<UTT ID="utt39" SPEECH="soundfiles/r148c015.au">
<SUGGEST ID="9">at noon</SUGGEST>
```

```
      </UTT>
<UTT ID="utt40" SPEECH="soundfiles/r148c015.au">
<REQUEST_COMMENT ID="2">does that sound alright for you</REQUEST_COMMENT>
      </UTT>
   </TURN>
<TURN ID="t17" SPEAKER="A">
<UTT ID="utt41" SPEECH="soundfiles/r148c016.au">
<CONFIRM ID="1">see you then</CONFIRM>
      </UTT>
   </TURN>
<TURN ID="t18" SPEAKER="B">
<UTT ID="utt42" SPEECH="soundfiles/r148c017.au">
<CONFIRM ID="2">roger</CONFIRM>
<GREETING_END ID="1">over and out</GREETING_END>
      </UTT>
   </TURN>
<TURN ID="t19" SPEAKER="A">
<UTT ID="utt43" SPEECH="soundfiles/r148c018.au">
<DEVIATE_SCENARIO ID="1">thought it was roger wilco</DEVIATE_SCENARIO>
      </UTT>
   </TURN>
<TURN ID="t20" SPEAKER="B">
<UTT ID="utt44" SPEECH="soundfiles/r148c019.au">
<REFER_TO_SETTING ID="1">oh no it is what we always say when we are
talking on screen</REFER_TO_SETTING>
      </UTT>
   </TURN>
</DIALOG>
```

The indentations of some of the end tags in the previous sample are there purely to improve readability of the text and are not necessarily meant to imply any hierarchical ordering of the data.

1	A: so	ASSERT(?), DIRECTIVE, COMMISSIVE
2	we should meet again	ASSERT(?), DIRECTIVE, COMMISSIVE
3	how 'bout	DIRECTIVE, COMMISSIVE
4	how 'bout next week	DIRECTIVE, COMMISSIVE
5	what day are good for you	ABANDONED, INFO-REQ
6	what days are good for you	ABANDONED, INFO-REQ
7	B: actually next week I am on vacation	ASSERT, REJECT(3,4), ANSWER(5.6)
8	A: gosh	ACKNOWLEDGE(7), EXCL
9	I guess we will have to meet the week after that	ASSERT, DIRECTIVE, COMMISSIVE, ACCEPT(7)
10	how 'bout Monday	DIRECTIVE, COMMISSIVE
11	B: Monday the tenth	INFO-REQ(?)
12	A: uh-huh	ANSWER(11)
13	B: well unfortunately my vacation runs through the fourteenth and I have plane tickets	ASSERT, REJECT(10-11)
14	I was planning on being on a beach in Acapulco about that point	ASSERT, REJECT(10-11), EXPLANATION(13)
15	A: well	? ACKNOWLEDGE(13-14)
16	when are you getting back	INFO-REQ

D.6: Combined Multi-level Annotation

The following short sample of a combined multi-level annotation is an attempt to incorporate information from all the different levels of annotation into one

computer-readable dialogue file. As such, is not meant to be 'human-friendly', but rather to represent something that may be produced using parsers and annotation tools, and to illustrate the complexity that results from incorporating multi-level annotation. This level of complexity is also responsible for the fact that many of the tags used to describe different features of the dialogue will overlap. Therefore, in order to produce a displayable output, use of a DTD defining possible levels of nesting will be necessary.

```
<DIALOG ID="r148c" STATUS="preprocessed"><TURN ID="t1" SPEAKER="A">
<UTT ID="utt1" SPEECH="soundfiles/r148c000.au"><DISCOURSE_PARTICLE ID="1">
<S ID="1"><W ID="1" AVC>so</W>.</DISCOURSE_PARTICLE></UTT>
<UTT ID="utt2" SPEECH="soundfiles/r148c000.au"><INIT><SUGGEST ID="1">
<NP ID="1"><W ID="2" PPp1N>we</W></NP><VP ID="1"><W ID="3" VM>should</W>
<W ID="4" VVI>meet</W><W ID="5" AV>again</W></VP>.</S></SUGGEST></INIT></UTT>
<UTT ID="utt3" SPEECH="soundfiles/r148c000.au"><REQUEST_SUGGEST ID="1">
<FALSE_START ID="1"><S ID="2"><ADVP ID="1"><W ID="6" AVWQ>how</W></ADVP>
<PP ID="1"><W ID="7" APR>'bout</W></PP></S></FALSE_START></REQUEST_SUGGEST>
</UTT><UTT ID="utt4" SPEECH="soundfiles/r148c000.au"><REQUEST_SUGGEST ID="1">
<S ID="3"><ADVP ID="2"><W ID="8" AVWQ>how</W></ADVP><PP ID="2"><W ID="9" APR>
'bout</W><NP ID="2"><W ID="10" AJ>next</W> <W ID="11" NCs>week</W></NP></PP>
</S>.</REQUEST_SUGGEST></UTT>
```

Appendix E: Morphosyntactic annotation of corpora

Appendix E.1: English tagset, with intermediate tags

Tables 1.8 and 1.9 are taken from Leech and Wilson (1994).

Table 1.8: English tagset, with intermediate tags

Tag	Description of word category	Example(s)	Intermediate Tag
AJ	(Positive) adjective, general	big	AJ10000000
AJR	Comparative adjective	bigger	AJ20000000
AJT	Superlative adjective	biggest	AJ30000000
APR	Preposition	at, of	AP1
APO	Postposition	's	AP3
ATD	Definite article	the	AT1000
ATIs	Indefinite article, singular	a, an	AT2010
AV	(Positive) adv., general	soon	AV1120
AVD	(Positive) adv. of degree	very, so	AV1220
AVDR	Comparative adverb of degree	more, less	AV2220
AVDT	Superlative adv. of degree	most, least	AV3220
AVDWQ	Adv. of degree, other wh-type	how	AV021[1|3]
AVR	Comparative adv., general	sooner	AV2120
AVT	Superlative adv., general	soonest	AV3120
AVWQ	General adv., other wh-type	when, why	AV011-2
AVWR	General adv., relative	where, why	AV0112
CC	Coord. conj., simple	and	C110
CCI	Coord. conj., initial	both (...and)	C130
CCM	Coord. conj., medial	(neither ...) nor	C140
CSC	Subord. conj., comparative	than	C203
CSF	Subord. conj., with finite	if, while	C201
CSN	Subord. conj., with nonfin.	in order (to)	C202

Table 1.8 (cont): English tagset, with intermediate tags

Tag	Description of word category	Example(s)	Intermediate Tag
DDs	Sing. demonstr. det.	this, that	PD001002010000
DDp	Pl. demonstr. det.	these	PD002002010000
DI	Indef. det., neutral for number	no, some	PD000002020000
DIs	Indef. determiner, sg.	every, much	PD001002020000
DIp	Indef. determiner, pl.	both, many	PD002002020000
DVs1	Poss. det., 1st pers. sg.	my	PD100102030000
DV2	Poss. det., 2nd person	your	PD200002030000
DV3sF	Poss. det., 3rd pers. sg. fem.	her	PD320102030000
DV3sM	Poss. det., 3rd pers. sg. masc.	his	PD310102030000
DV3sU	Poss. det., 3rd pers. sg. neut.	its	PD330102030000
DVp1	Poss. det., 1st pers. pl.	our	PD100202030000
DVp3	Poss. det., 3rd pers. pl.	their	PD300202030000
DWR	Relative det.	which	PD000002040200
DWQ	Other wh-det.	what	PD000002040-200
IJ	Interjection	Oh, Yes	I
NCs	Sing. common noun	book, man	N101000
NCp	Pl. common noun	books, men	N102000
NPs	Sing. proper noun	Mary	N201000
NPp	Pl. proper noun	Rockies	N202000
NUC	Cardinal numeral, neutral for number	two	NU10000
NUCs	Sing. cardinal numeral	one	NU10100
NUCp	Pl. cardinal numeral	fifties	NU10200
NUOs	Sing. ordinal numeral	seventh	NU20100
NUOp	Pl. ordinal numeral	sevenths	NU20200
PDs	Sg. demonstr. pron.	this	PD001001100000
PDp	Pl. demonstr. pron.	those	PD002001100000
PI	Indef. pron., neutral for number	all, none	PD000001200000
PIs	Sg. indef. pron.	anyone	PD001001200000
PIp	Pl. indef. pron.	few, many	PD002001200000
PPs1N	Pers. pron., 1st pers. sg. nom.	I	PD101011501000

Table 1.8 (cont): English tagset, with intermediate tags

Tag	Description of word category	Example(s)	Intermediate Tag
PPs1O	Pers. pron., 1st pers. sg. obl.	me	PD101061501000
PP2	2nd pers. pers. pron.	you	PD2000[1|6]1501000
PPs3NF	Pers. pron., 3rd pers.sg.nom.fem.	she	PD321011501000
PPs3NM	Pers. pron., 3rd pers.sg.nom.masc.	he	PD311011501000
PPs3U	Pers. pron., 3rd pers.sing.neuter	it	PD3310[1|6]1501000
PPs3OF	Pers. pron., 3rd pers.sg.obl.fem.	her	PD321061501000
PPs3OM	Pers. pron., 3rd pers.sg.obl.masc.	him	PD311061501000
PPp1N	Pers. pron., 1st pers. pl. nom.	we	PD102011501000
PPp1O	Pers. pron., 1st pers. pl. oblique	us	PD102061501000
PPp3N	Pers. pron., 3rd pers. pl. nom.	they	PD302011501000
PPp3O	Pers. pron., 3rd pers. pl. oblique	them	PD302061501000
PRs1	Reflexive pron., 1st pers. sg.	myself	PD101001502000
PRs2	Reflexive pron., 2nd pers. sg.	yourself	PD201001502000
PRs3F	Reflexive pron., 3rd pers. sg. fem.	herself	PD321001502000
PRs3M	Reflexive pron., 3rd pers. sg. masc.	himself	PD311001502000
PRs3U	Reflex. pron., 3rd pers. sg. neut.	itself	PD331001502000
PRp1	Reflex. pron., 1st pers. pl.	ourselves	PD102001502000
PRp2	Reflex. pron., 2nd pers. pl.	yourselves	PD202001502000
PRp3	Reflex. pron., 3rd pers. pl.	themselves	PD302001502000
PVs1	Poss. pron., 1st pers. sg.	mine	PD100101300000
PV2	Poss. pron., 2nd pers.	yours	PD200001300000

Table 1.8 (cont): English tagset, with intermediate tags

Tag	Description of word category	Example(s)	Intermediate Tag
PVs3F	Poss. pron., 3rd pers. fem.	hers	PD320101300000
PVs3M	Poss. pron., 3rd pers. masc.	his	PD310101300000
PVs3U	Poss. pron., 3rd pers. neut.	its	PD330101300000
PVp1	Poss. pron., 1st pers. pl.	ours	PD100201300000
PVp3	Poss. pron., 3rd pers. pl.	theirs	PD300201300000
PWQ	Other wh-pron., neutral for case	what, which	PD000001400-200
PWQN	Other wh-pron., nominative	who	PD000011400-200
PWQO	Other wh-pron., oblique	whom	PD000061400-200
PWR	Rel. pron., neutral for case	which	PD000001400200
PWRN	Relative pron., nominative	who	PD000011400200
PWRO	Relative pron., oblique	whom	PD000061400200
RFO	Formula	X/21	R200
RFW	Foreign word	mawashi	R100
RSY	Symbol, neutral for number	£, '	R300
RSYs	Symbol, singular	A, b	R310
RSYp	Symbol, plural	As, b's	R320
RUN	Unclassified	er, um	R600
UI	infinitive marker	to (eat)	U1
UN	negative particle	not, -n't	U2
UX	existential *there*	there	U3

Table 1.8 (cont): English tagset, with intermediate tags

Tag	Description of word category	Example(s)	Intermediate Tag
VM	Modal aux. verb	can, will	V0001100200002
VPB	Finite base form, primary aux.	be, do, have	V[[-301|002]111|000121|000130]0200001
VPD	Past tense, primary aux.	did, had	V0001140200001
VPDR	Past tense -re form, primary aux.	were	V[[201|002]11|00012]40200001
VPDZ	Past tense -s form, primary aux.	was	V-2011140200001
VPG	-Ing form, primary aux.	being, having	V0002900200001
VPI	Infinitive, primary aux.	(to) be/have	0002500200001
VPM	Pres. tense 1st pers. sg, primary aux.	am	V1011110200001
VPN	Past participle, primary aux.	been	V0002640200001
VPR	Pres. tense -re form, primary auxiliary	are	V[201|002]1110200001
VPZ	Pres. tense -s form, primary auxiliary	is, does, has	V3011110200001
VVB	Finite base form, main verb	eat, have	V[[-301|002]111|000121|000130]0100000

Table 1.8 (cont): English tagset, with intermediate tags

Tag	Description of word category	Example(s)	Intermediate Tag
VVD	Past tense, main verb	ate, had	V0001140100000
VVDR	Past tense -re form, main verb	were	V[[201|002]11|00012]40100000
VVDZ	Past tense -s form, main verb	was	V-2011140100000
VVG	-Ing form, main verb	leaving, being	V0002900100000
VVI	Infinitive, main verb	(to) leave/do	V0002500100000
VVM	Pres. tense 1st pers. sing, main verb	am	V1011110100000
VVN	Past part., main verb	eaten, left	V0002640100000
VVR	Pres. tense -re form, main verb	are	V[201|002]1110100000
VVZ	Pres. tense -s form, main verb	is	V3011110100000

Appendix E.2: Italian DMI codes, with intermediate tags

Table 1.9: Italian DMI codes, with intermediate tags

Code	Description of word category	Example(s)	Intermediate Tag
AFN	Adj.pos.femm.inv.	carta/e valore	AJ12[1|2]0
ANS	Adj.pos.comm.sing.	dolce	AJ1410
ANP	Adj.pos.comm.plur.	dolci	AJ1420
AMN	Adj.pos.masc.inv.	complemento/i oggetto	AJ11[1|2]0
AFSS	Adj.sup.femm.sing.	grandissima, massima	AJ3210
AFPS	Adj.sup.femm.plur.	grandissime, massime	AJ3220
AMPS	Adj.sup.masc.plur.	grandissimi, massimi	AJ3120
AMSS	Adj.sup.masc.sing.	grandissimo, massimo	AJ3110
ANSC	Adj.com.comm.sing.	maggiore	AJ2410
ANPC	Adj.pos.comm.plur.	maggiori	AJ2420
ANNC	Adj.pos.comm.inv.	meglio, peggio	AJ24[1|2]0
ANN	Adj.pos.comm.inv.	pari, dappoco	AJ14[1|2]0
AFS	Adj.pos.femm.sing.	vera	AJ1210
AFP	Adj.pos.femm.plur.	vere	AJ1220
AMP	Adj.pos.masc.plur.	veri	AJ1120
AMS	Adj.pos.masc.sing.	vero	AJ1110
B	Adv.pos.	forte	AV1000
BC	Adv.com.	maggiormente	AV2000
BS	Adv.pos.mann.	fortemente	AV1600
BSS	Adv.sup.mann.	fortissimamente	AV3600
C	Conj.subord.	perché	C200
CC	Conj.coord.	e	C100
DDMS	PrAdj.dem.masc.sing.	quello, quel	PD01100201
DDMP	PrAdj.dem.masc.plur.	quelli	PD01200201
DDFS	PrAdj.dem.femm.sing.	quella	PD02100201
DDFP	PrAdj.dem.femm.plur.	quelle	PD02200201
DDNS	PrAdj.dem.comm.sing.	ciò	PD04100201
DDNP	PrAdj.dem.comm.plur.	costoro	PD04200201
DIMS	PrAdj.ind.masc.sing.	alcuno, alcun	PD01100202
DIMP	PrAdj.ind.masc.plur.	alcuni	PD01200202
DIFS	PrAdj.ind.femm.sing.	qualcuna	PD02100202
DIFP	PrAdj.ind.femm.plur.	poche	PD02200202
DINS	PrAdj.ind.comm.sing.	ogni	PD04100202
DINP	PrAdj.ind.comm.plur.	tali, altrui	PD04200202

Table 1.9 (cont): Italian DMI codes, with intermediate tags

Code	Description of word category	Example(s)	Intermediate Tag
DEMS	PrAdj.escl.masc.sing.	quanto!	PD0110020003
DEMP	PrAdj.escl.masc.plur.	quanti!	PD0120020003
DEFS	PrAdj.escl.femm.sing.	quanta!	PD0210020003
DEFP	PrAdj.escl.femm.plur.	quante!	PD0220020003
DENS	PrAdj.escl.comm.sing.	quale!	PD0410020003
DENP	PrAdj.escl.comm.plur.	quali!	PD0420020003
DENN	PrAdj.escl.comm.inv.	che!	PD04[1|2]0020003
DPMS1	PrAdj.poss.1p.masc.sing.	mio	PD11100201
DPMP1	PrAdj.poss.1p.masc.plur.	miei	PD11200201
DPFS1	PrAdj.poss.1p.femm.sing.	mia	PD12100201
DPFP1	PrAdj.poss.1p.femm.plur.	mie	PD12200201
DPMS2	PrAdj.poss.2p.masc.sing.	tuo	PD21100201
DPMP2	PrAdj.poss.2p.masc.plur.	tuoi	PD21200201
DPFS2	PrAdj.poss.2p.femm.sing.	tua	PD22100201
DPFP2	PrAdj.poss.2p.femm.plur.	tue	PD22200201
DPMS3	PrAdj.poss.3p.masc.sing.	suo	PD31100201
DPMP3	PrAdj.poss.3p.masc.plur.	suoi	PD31200201
DPFS3	PrAdj.poss.3p.femm.sing.	sua	PD32100201
DPFP3	PrAdj.poss.3p.femm.plur.	sue	PD32200201
DPMS1	PrAdj.poss.1p.masc.sing.	nostro	PD11100201
DPMP1	PrAdj.poss.1p.masc.plur.	nostri	PD11200201
DPFS1	PrAdj.poss.1p.femm.sing.	nostra	PD12100201
DPFP1	PrAdj.poss.1p.femm.plur.	nostre	PD12200201
DPMS2	PrAdj.poss.2p.masc.sing.	vostro	PD21100201
DPMP2	PrAdj.poss.2p.masc.plur.	vostri	PD21200201
DPFS2	PrAdj.poss.2p.femm.sing.	vostra	PD22100201
DPFP2	PrAdj.poss.2p.femm.plur.	vostre	PD22200201
DPNP3	PrAdj.poss.3p.comm.plur.	loro	PD34200201
DPNN	PrAdj.poss.comm.inv.	altrui	PD04[1|2]00201
DTMS	PrAdj.int.masc.sing.	quanto?	PD0110020001
DTMP	PrAdj.int.masc.plur.	quanti?	PD0120020001
DTFS	PrAdj.int.femm.sing.	quanta?	PD0210020001
DTFP	PrAdj.int.femm.plur.	quante?	PD0220020001
DTNN	PrAdj.int.comm.inv.	che?	PD04[1|2]0020001
DTNS	PrAdj.int.comm.sing.	quale?	PD0410020001
DTNP	PrAdj.int.comm.plur.	quali?	PD0420020001
DRNN	PrAdj.rel.comm.inv.	che	PD04[1|2]0020002
DRNS	PrAdj.rel.comm.sing.	quale	PD0410020002
DRNP	PrAdj.rel.comm.plur.	quali	PD0420020002
I		oh!	I

Table 1.9 (cont): Italian DMI codes, with intermediate tags

Code	Description of word category	Example(s)	Intermediate Tag
SFN	Noun comm.femm.inv.	attività (la/le)	N12[1|2]
SFP	Noun comm.femm.plur.	case	N122
SFS	Noun comm.femm.sing.	casa	N121
SMN	Noun comm.masc.inv.	re, caffè (il/i)	N11[1|2]
SMP	Noun comm.masc.plur.	libri	N112
SMS	Noun comm.masc.sing.	libro	N111
SNN	Noun comm.comm.inv.	sosia (il/la, i/le)	N14[1|2]
SNP	Noun comm.comm.plur.	insegnanti (gli/le)	N142
SNS	Noun comm.comm.sing.	insegnante (un/una)	N141
SPFP	Noun prop.femm.plur.	Marie	N222
SPFS	Noun prop.femm.sing.	Maria	N221
SPMP	Noun prop.masc.plur.	Borboni	N212
SPMS	Noun prop.masc.sing.	Mario	N211
PDMS3	Pron.dem.masc.sing.3	costui	PD31100110
PDMS	Pron.dem.masc.sing.	quello	PD01100110
PDMP	Pron.dem.masc.sing.	quelli	PD01200110
PDFS	Pron.dem.femm.sing.	quella	PD02100110
PDFP	Pron.dem.femm.plur.	quelle	PD02200110
PDNS	Pron.dem.comm.sing.	ciò	PD04100110
PDNP	Pron.dem.comm.plur.	tali	PD04200110
PEMS	Pron.escl.masc.sing.	quanto!	PD0110010003
PEMP	Pron.escl.masc.plur.	quanti!	PD0120010003
PEFS	Pron.escl.femm.sing.	quanta!	PD0210010003
PEFP	Pron.escl.femm.plur.	quante!	PD0220010003
PENS	Pron.escl.comm.sing.	che (vedo!)	PD0410010003
PENN	Pron.escl.comm.inv.	chi!	PD04[1|2]0010003
PIMS	Pron.ind.masc.sing.	uno	PD01100120
PIMP	Pron.ind.masc.plur.	alcuni	PD01200120
PIFS	Pron.ind.femm.sing.	una	PD02100120
PIFP	Pron.ind.femm.plur.	alcune	PD02200120
PINS	Pron.ind.comm.sing.	chiunque	PD04100120
PINP	Pron.ind.comm.plur.	tali, quali	PD04200120
PPMS1	Pron.poss.1p.masc.sing.	mio	PD11100130
PPMP1	Pron.poss.1p.masc.plur.	miei	PD11200130
PPFS1	Pron.poss.1p.femm.sing.	mia	PD12100130
PPFP2	Pron.poss.1p.femm.plur.	mie	PD12200130
PPMS2	Pron.poss.2p.masc.sing.	tuo	PD21100130
PPMP2	Pron.poss.2p.masc.plur.	tuoi	PD21200130
PPFS2	Pron.poss.2p.femm.sing.	tua	PD22100130

Table 1.9 (cont): Italian DMI codes, with intermediate tags

Code	Description of word category	Example(s)	Intermediate Tag
PPFP2	Pron.poss.2p.femm.plur.	tue	PD22200130
PPMS3	Pron.poss.3p.masc.sing.	suo	PD31100130
PPMP3	Pron.poss.3p.masc.plur.	suoi	PD31200130
PPFS3	Pron.poss.3p.femm.sing.	sua	PD32100130
PPFP3	Pron.poss.3p.femm.plur.	sue	PD32200130
PPMS1	Pron.poss.1p.masc.sing.	nostro	PD11100130
PPMP1	Pron.poss.1p.masc.plur.	nostri	PD11200130
PPFS1	Pron.poss.1p.femm.sing.	nostra	PD12100130
PPFP1	Pron.poss.1p.femm.plur.	nostre	PD12200130
PPMS2	Pron.poss.2p.masc.sing.	vostro	PD21100130
PPMP2	Pron.poss.2p.masc.plur.	vostri	PD21200130
PPFS2	Pron.poss.2p.femm.sing.	vostra	PD22100130
PPFP2	Pron.poss.2p.femm.plur.	vostre	PD22200130
PPNP3	Pron.poss.3p.comm.plur.	loro	PD34200130
PTNS	Pron.int.comm.sing.	chi?	PD0410010001
PTNN	Pron.int.comm.inv.	che?	PD04[1|2]0010001
PTMS	Pron.int.masc.sing.	quanto?	PD0110010001
PTMP	Pron.int.masc.plur.	quanti?	PD0120010001
PTFS	Pron.int.femm.sing.	quanta?	PD0210010001
PTFP	Pron.int.femm.plur.	quante?	PD0220010001
PRNN	Pron.rel.comm.inv.	che, chi, cui	PD04[1|2]0010002
PRNS	Pron.rel.comm.sing.	quanto	PD0410010002
PRMS	Pron.rel.masc.sing.	quanto	PD0110010002
PRMP	Pron.rel.masc.plur.	quanti	PD0120010002
PRFP	Pron.rel.femm.plur.	quante	PD0220010002
PQNS1	Pron.pers.comm.sing.1	io	PD141001001
PQNS2	Pron.pers.comm.plur.2	tu	PD241001001
PQMS3	Pron.pers.masc.sing.3	egli, lui, esso	PD311001001
PQFS3	Pron.pers.femm.sing.3	ella, lei, essa	PD321001001
PQNP1	Pron.pers.comm.plur.1	noi	PD142001001
PQNP2	Pron.pers.comm.plur.2	voi	PD242001001
PQNP3	Pron.pers.comm.plur.3	loro	PD342001001
PQMP3	Pron.pers.masc.plur.3	essi	PD312001001
PQFP3	Pron.pers.femm.plur.3	esse	PD322001001
PQNS1	Pron.pers.comm.sing.1	me	PD141001001
PQNS2	Pron.pers.comm.sing.2	te	PD241001001
PQMS3	Pron.pers.masc.sing.3	lui, esso	PD311001001
PQFS3	Pron.pers.femm.sing.3	lei, essa	PD321001001
PQNP1	Pron.pers.comm.plur.1	noi	PD142001001
PQNP2	Pron.pers.comm.plur.2	voi	PD242001001
PQNP3	Pron.pers.comm.plur.3	loro	PD342001001

Table 1.9 (cont): Italian DMI codes, with intermediate tags

Code	Description of word category	Example(s)	Intermediate Tag
PQMP3	Pron.pers.masc.plur.3	essi	PD312001001
PQFP3	Pron.pers.femm.plur.3	esse	PD322001001
PQNS1	Pron.pers.comm.sing.1	mi	PD141001001
PQNS2	Pron.pers.comm.sing.2	ti	PD241001001
PQMS3	Pron.pers.masc.sing.3	gli	PD311001001
PQNP1	Pron.pers.comm.plur.1	ci	PD142001001
PQNP2	Pron.pers.comm.plur.2	vi	PD242001001
PQNP3	Pron.pers.comm.plur.3	loro	PD342001001
PQMP3	Pron.pers.masc.plur.3	li	PD312001001
PQFP3	Pron.pers.femm.plur.3	le	PD322001001
PFNS1	Pron.refl.comm.sing.1	mi (me stesso)	PD141001002
PFNS2	Pron.refl.comm.sing.1	ti (te stesso)	PD241001002
PFNN3	Pron.refl.comm.inv. 3	sè, si	PD311001002
PFNP1	Pron.refl.comm.plur.1	ci	PD142001002
PFNP2	Pron.refl.comm.plur.2	vi	PD242001002
PFNP3	Pron.refl.comm.plur.3	loro	PD342001002
VFY	Verb aux. inf.pres.	avere	V00025101
VGY	Verb aux. ger.pres.	avendo	V00027102
VF	Verb main inf.pres.	amare	V00025101
VG	Verb main ger.pres.	amando	V00027102
VP1IFY	Verb aux. 1pl.ind.fut.	avremo	V10211302
VP2IFY	Verb aux. 2pl.ind.fut.	avrete	V20211302
VP3IFY	Verb aux. 3pl.ind.fut.	avranno	V30211302
VS1IFY	Verb aux. 1sg.ind.fut.	avrò	V10111302
VS2IFY	Verb aux. 2sg.ind.fut.	avrai	V20111302
VS3IFY	Verb aux. 3sg.ind.fut.	avrà	V30111302
VP1IF	Verb main 1pl.ind.fut.	ameremo	V10211301
VP2IF	Verb main 2pl.ind.fut.	amerete	V20211301
VP3IF	Verb main 3pl.ind.fut.	ameranno	V30211301
VS1IF	Verb main 1sg.ind.fut.	amerò	V10111301
VS2IF	Verb main 2sg.ind.fut.	amerai	V20111301
VS3IF	Verb main 3sg.ind.fut.	amerà	V30111301
VP1CIY	Verb aux. 1pl.subj.impf.	avessimo	V10212202
VP2CIY	Verb aux. 2pl.subj.impf.	aveste	V20212202
VP3CIY	Verb aux. 3pl.subj.impf.	avessero	V30212202
VS1CIY	Verb aux. 1sg.subj.impf.	avessi	V10112202
VS2CIY	Verb aux. 2sg.subj.impf.	avessi	V20112202
VS3CIY	Verb aux. 3sg.subj.impf.	avesse	V30112202
VP1CI	Verb main 1pl.subj.impf.	amassimo	V10212201
VP2CI	Verb main 2pl.subj.impf.	amaste	V20212201
VP3CI	Verb main 3pl.subj.impf.	amassero	V30212201

Table 1.9 (cont): Italian DMI codes, with intermediate tags

Code	Description of word category	Example(s)	Intermediate Tag
VS1CI	Verb main 1sg.subj.impf.	amassi	V10112201
VS2CI	Verb main 2sg.subj.impf.	amassi	V20112201
VS3CI	Verb main 3sg.subj.impf.	amasse	V30112201
VP1IIY	Verb aux. 1pl.ind.impf.	avevamo	V10211202
VP2IIY	Verb aux. 2pl.ind.impf.	avevate	V20211202
VP3IIY	Verb aux. 3pl.ind.impf.	avevano	V30211202
VS1IIY	Verb aux. 1sg.ind.impf.	avevo	V10111202
VS2IIY	Verb aux. 2sg.ind.impf.	avevi	V20111202
VS3IIY	Verb aux. 3sg.ind.impf.	aveva	V30111202
VP1II	Verb main 1pl.ind.impf.	amavamo	V10211201
VP2II	Verb main 2pl.ind.impf.	amavate	V20211201
VP3II	Verb main 3pl.ind.impf.	amavano	V30211201
VS1II	Verb main 1sg.ind.impf.	amavo	V10111201
VS2II	Verb main 2sg.ind.impf.	amavi	V20111201
VS3II	Verb main 3sg.ind.impf.	amava	V30111201
VP1CPY	Verb aux. 1pl.subj.pres.	abbiamo	V10212102
VP2CPY	Verb aux. 2pl.subj.pres.	abbiate	V20212102
VP3CPY	Verb aux. 3pl.subj.pres.	abbiano	V30212102
VS1CPY	Verb aux. 1sg.subj.pres.	abbia	V10112102
VS2CPY	Verb aux. 2sg.subj.pres.	abbia	V20112102
VS3CPY	Verb aux. 3sg.subj.pres.	abbia	V30112102
VP1CP	Verb main 1pl.subj.pres.	amiamo	V10212101
VP2CP	Verb main 2pl.subj.pres.	amiate	V20212101
VP3CP	Verb main 3pl.subj.pres.	amino	V30212101
VS1CP	Verb main 1sg.subj.pres.	ami	V10112101
VS2CP	Verb main 2sg.subj.pres.	ami	V20112101
VS3CP	Verb main 3sg.subj.pres.	ami	V30112101
VP1DPY	Verb aux. 1pl.cond.pres.	avremmo	V10214102
VP2DPY	Verb aux. 2pl.cond.pres.	avreste	V20214102
VP3DPY	Verb aux. 3pl.cond.pres.	avrebbero	V30214102
VS1DPY	Verb aux. 1sg.cond.pres.	avrei	V10114102
VS2DPY	Verb aux. 2sg.cond.pres.	avresti	V20114102
VS3DPY	Verb aux. 3sg.cond.pres.	avrebbe	V30114102
VP1DP	Verb main 1pl.cond.pres.	ameremmo	V10214101
VP2DP	Verb main 2pl.cond.pres.	amereste	V20214101
VP3DP	Verb main 3pl.cond.pres.	amerebbero	V30214101
VS1DP	Verb main 1sg.cond.pres.	amerei	V10114101
VS2DP	Verb main 2sg.cond.pres.	ameresti	V20114101
VS3DP	Verb main 3sg.cond.pres.	amerebbe	V30114101
VP1IPY	Verb aux. 1pl.ind.pres.	abbiamo	V10211102
VP2IPY	Verb aux. 2pl.ind.pres.	avete	V20211102

Table 1.9 (cont): Italian DMI codes, with intermediate tags

Code	Description of word category	Example(s)	Intermediate Tag
VP3IPY	Verb aux. 3pl.ind.pres.	hanno	V30211102
VS1IPY	Verb aux. 1sg.ind.pres.	ho	V10111102
VS2IPY	Verb aux. 2sg.ind.pres.	hai	V20111102
VS3IPY	Verb aux. 3sg.ind.pres.	ha	V30111102
VP1IP	Verb main 1pl.ind.pres.	amiamo	V10211101
VP2IP	Verb main 2pl.ind.pres.	amate	V20211101
VP3IP	Verb main 3pl.ind.pres.	amano	V30211101
VS1IP	Verb main 1sg.ind.pres.	amo	V10111101
VS2IP	Verb main 2sg.ind.pres.	ami	V20111101
VS3IP	Verb main 3sg.ind.pres.	ama	V30111101
VP2MPY	Verb aux. 2pl.imp.pres.	abbiate	V20213102
VS2MPY	Verb aux. 2sg.imp.pres.	abbi	V20113102
VP2MP	Verb main 2pl.imp.pres.	amate	V20213101
VS2MP	Verb main 2sg.imp.pres.	ama	V20113101
VNPPPY	Verb aux. comm.pl.part.pres.	aventi	V04226102
VNSPPY	Verb aux. comm.sg.part.pres.	avente	V04126102
VNPPP	Verb main comm.pl.part.pres.	amanti	V04226101
VNSPP	Verb main comm.sg.part.pres.	amante	V04126101
VP1IRY	Verb aux. 1pl.ind.past	avemmo	V10211402
VP2IRY	Verb aux. 2pl.ind.past	aveste	V20211402
VP3IRY	Verb aux. 3pl.ind.past	ebbe	V30211402
VS1IRY	Verb aux. 1sg.ind.past	ebbi	V10111402
VS2IRY	Verb aux. 2sg.ind.past	avesti	V20111402
VS3IRY	Verb aux. 3sg.ind.past	ebbe	V30111402
VP1IR	Verb main 1pl.ind.past	amammo	V10211401
VP2IR	Verb main 2pl.ind.past	amaste	V20211401
VP3IR	Verb main 3pl.ind.past	amarono	V30211401
VS1IR	Verb main 1sg.ind.past	amai	V10111401
VS2IR	Verb main 2sg.ind.past	amasti	V20111401
VS3IR	Verb main 3sg.ind.past	amò	V30111401
VFPPRY	Verb aux. femm.pl.part.past	avute	V02226402
VFSPRY	Verb aux. femm.sg.part.past	avuta	V02126402
VMPPRY	Verb aux. masc.pl.part.past	avuti	V01226402
VMSPRY	Verb aux. masc.sg.part.past	avuto	V01126402
VFPPR	Verb main femm.pl.part.past	amate	V02226401
VFSPR	Verb main femm.sg.part.past	amata	V02126401
VMPPR	Verb main masc.pl.part.past	amati	V01226401
VMSPR	Verb main masc.sg.part.past	amato	V01126401

2 Audio-visual and multimodal speech-based systems

Communication between humans uses many modalities. We communicate not only via verbal language, but also through our use of intonation, gaze, hand gestures, body gestures, and facial expressions. Using these modalities, we can add, modify, and substitute information in spoken conversations. Complementary use of several modalities in human-to-human communication ensures high accuracy, and only few communication problems occur. When communication problems do occur, conversation partners can easily recover, using the redundancy and complementarity of modalities. The goal of research on multimodal systems is to investigate how human–computer interaction can benefit from multiple modalities in similar ways.
It is not easy to define the notion of modality, or of multimodality, and the popularity of the term 'multimedia' complicates the issue somewhat. This chapter presents an overview of current standards and common resources for multimodal speech systems. While multimodality may occur in many contexts, the scope of the EAGLES handbook is limited to spoken language systems. This chapter is therefore limited to multimodal systems that use speech as either an input or output modality. Although the value of such systems for the user often depends on the level of system competence (in some contexts, such as conversational agents, "intelligence" may be more appropriate), current research on multimodal systems, and also this chapter, focuses on multimodality in the user interface.
The following four categories of multimodal systems are covered:

1. systems that combine speech with visual input,
2. speech with visual output,
3. speech with other input modalities, and
4. speech with other output modalities.

Examples of visual modalities are lip movements and whole faces, and other modalities include handwriting and gestures. The chapter also clarifies basic terminology, surveys current multimodal systems, and provides recommendations on the different system components. Section 2.9 describes the few standards that are already established in this comparatively recent research field and the details of multimodal technology.

2.1 Introduction

The ease and robustness of human–human communication is due to extremely high recognition accuracy (using multiple input channels) and the redundant and complementary use of several modalities. Research in multimodal systems is based on the expectation that human–computer interaction can benefit from modelling several modalities in analogous ways.
This chapter focuses on multimodal systems that have speech either as input or output modality. The introductory section is organised as follows. First, Section 2.1.1 clarifies the basic terminology for multimodal systems and Sec-

tion 2.1.2 gives a brief outline of the whole chapter. Then, Section 2.1.3 motivates multimodal systems by summarising what benefits of multimodal systems other researchers have identified. The following two sections (2.1.4 and 2.1.5) enumerate and define input and output modalities that have been actively researched in the field. In Section 2.1.6, we present two taxonomies of multimodal applications, one based on the available input modalities and output media, and the other based on the task categories that are supported by the application.

2.1.1 Terminology

The term ultimodal interface has recently become a buzzword by which different researchers mean different things. Overall, there is a lack of agreement, probably due to the interdisciplinary nature of the field. Here, we describe the most important terms, their varying usage in the field, and a model of human–human and human–computer interaction that helps to clarify the different terms and the definitions. Then we introduce the definitions that are used in this chapter. There has been discussion in the field on the meaning of and distinction between *multimodal* and *multimedia* applications (e.g. Cohen and Massaro 1990; Balbo et al. 1993; Schomaker et al. 1995a; Dugast 1998), and on related terminology. Several expressions are used: types of information, channel, media, modality, communication means, communication code, mode, multimedia, multimode, multimodal. These expressions are sometimes qualified with "sensory" (i.e. sensory channel) or "intelligent" (i.e. intelligent multimedia system).
However, in spite of the existing confusion in terminologies, a number of general concepts appear to be established in this field. Existing multimodal systems will usually possess a subset of the following features:

- The user communicates with the computer using several physical input devices (keyboard, mouse, microphone).
- In order to achieve this communication, several muscles are activated by the user (e.g. vocal cords, hand).
- The information sensed by the computer input devices can be processed at different levels of abstraction, providing different levels of the understanding of the intention of the user.
- The computer communicates with the user using several output devices (e.g. screen, loudspeaker).
- On these output devices, the computer may send statically predefined raw data (static images, recorded audio files, video clips ...) or data generated dynamically from more abstract representations (such as generation of text, graphics, images, speech synthesis).
- Thus, several senses of the user may be stimulated by computer output (e.g. vision, hearing).

Figure 2.1 shows a model of human–computer interaction that illustrates the situation on an abstract level. Humans employ several output modalities (or channels in the figure) to communicate with each other, and also with computers. The latter, called computer input modalities, are obviously constrained by what current computer technology can process. The computer system represents output to the human user, choosing one or more computer output media. These human input modalities (or channels) are interpreted based on human cognitive capabilities. The loop from human output channels to human input

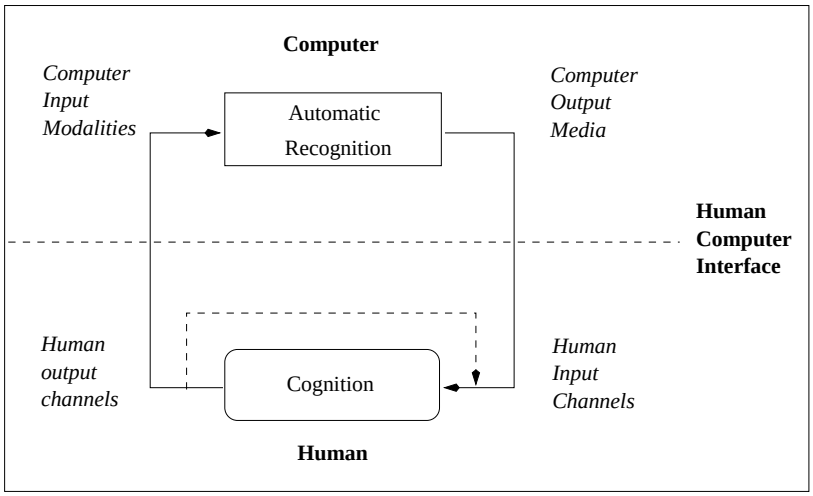

Figure 2.1: Model of human–computer interaction (from Schomaker et al. 1995a)

channels in the figure alludes to human–human communication. The human intrinsic autofeedback loop plays an important role. It is continuously active in normal daily life, where muscle activity will have a direct and perceptible effect on the senses. Movement of the head or body causes optic flow, for instance, and the user will see his hands moving in the work area. In speech, the intrinsic autofeedback loop has a great part. This becomes quickly evident if we use an audio system with an imperfect transfer function overruling the immediate feedback channel. Examples are the severe disturbances of speech in case of delayed feedback. However, delays in the visual effects of pointer-device movement are also disturbing. Using a computer system means that some proportion of the feedback will remain physical, whereas a part of the feedback channel bandwidth will be taken over by the computer system. In the asymptotic case of Virtual Reality applications and, e.g. flight simulators, the goal is to maximise the bandwidth of the computer-based feedback. We will consider "input" as concerning interaction from the user to the computer and "output" as concerning interaction from the computer to the user.

2.1.1.1 How are some of the basic terms used in the field?

There are several human senses (sight, hearing, touch, smell, taste, and balance) (Stein and Meredith 1993; Howard Hughes medical institute 1995), and psychologists use the term modality explicitly in the context of sensory modalities. Some researchers use the term media for "physical devices" and the term modality for "a way to use a media"[1] (Coutaz 1992; Bellik and Teil 1992; Martin 1995). For example, with the pen (input) medium, you can use several

[1] Like the term *data* (originally a plural form with singular *datum*, but now frequently found as a singular), the term *media* is originally a plural form (singular *medium*), but is gradually coming to be used as a singular. – Ed.

modalities such as drawing, writing, and gestures to provide input to a computer system, and with the "screen", the computer can use several modalities such as text, graphics, picture, and video to present output to the user. Some consider that media relate to machines while modalities relate to human beings. In Maybury (1997), media are both the physical entity and the way to use it, and modalities refer to the human senses, whereas in Bordegoni et al. (1997), a medium is a type of information and representation format (e.g. pixmap graphics or video frames), and a modality is a way of encoding the data to present it to the user (e.g. graphics, textual language, spoken language, and video). "Natural language" has been counted among the media (Arens et al. 1993), a mode (André et al. 1993), and a modality (Burger and Marshall 1993). Others claim that the difference between multimedia and multimodal is the use of semantic representations and understanding processes, and use the term channel interchangeably with modality (e.g. Schomaker et al. 1995a). But the term *intelligent multi-media* is also being increasingly used. Obviously, there is no overall agreement in the field on terminology. Therefore, researchers are forced to work around this confusion. Some authors explicitly provide their own definitions in each paper, while others provide examples from which the reader can infer implicit definitions. Below we introduce the definitions used in the present document. They were arrived at as a consensus between the authors and other experts involved in the EAGLES project.

- *Multimodal systems.* For present purposes we define multimodal systems as systems which *represent and manipulate information from different human communication channels at multiple levels of abstraction.* Multimodal systems can automatically extract meaning from raw multimodal input data, and conversely they produce perceivable multimodal information from symbolic abstract representations. We will assume that a multimodal system is either a multimodal interface or a multimodal speech system.
- *Multimedia systems.* We characterise a multimedia system as *a system which offers more than one device for user input to the system, and for system feedback to the user.* Such devices include microphone, speaker, keyboard, mouse, touch screen, camera. In contrast to multimodal systems, multimedia systems do not generate abstract concepts automatically (which are typically encoded manually as meta-information instead), and they do not transform the information. This chapter focuses on multimodality in input and output, and not on multimedia. However, a brief outline of multimedia in the context of enhancing speech output is included (see Section 2.7). More information on the design of multimedia user interfaces can be found in 14915 (1998) and in Vossen et al. (1998).
- *Multimodal interface.* Since in this chapter we focus on multimodal systems that include speech either as input or output, we define a multimodal interface as *an interface that combines speech input or output with other input and output modalities.* The overall goal is to facilitate human–computer interaction. In part, that can be achieved by using the same communication channels that people naturally employ when they communicate, but trade-offs are necessary to make such interaction feasible with current automatic recognition technology.
- *Multimodal speech system.* Since speech is multimodal in nature, we also discuss multimodal aspects of speech recognition. People accompany speech naturally with non-verbal cues, including facial expression, eye/gaze, and lip

movements. All cues interpreted together ensure fluent human-to-human communication. We therefore define multimodal speech systems (or audio-visual speech systems) as *systems which attempt to utilise the same multiple channels as human communication by integrating automatic speech recognition with other non-verbal cues, and by integrating non-verbal cues with speech synthesis to improve the output side of a multimodal application* (e.g. in talking heads).

Obviously, there are forms of multimodality where speech does not play a role at all, for example in conventional keyboard and mouse input in most current desktop applications, pen and keyboard input in pen-based computers such as PDAs (Personal Digital Assistants), and camera and keyboard in some advanced security systems. In the EAGLES project, however, we have consciously decided to focus on multimodality in the context of speech.

2.1.2 Chapter outline

This chapter on multimodal interfaces and multimodal speech systems (as defined above) is structured as follows. After this introductory section, Section 2.2 presents results from a literature review and survey of multimodal systems conducted by the authors of this chapter. Section 2.3 discusses evaluation of multimodal systems. Challenges in evaluating multimodal systems are identified, known evaluation methodologies are reviewed, and issues in evaluating certain kinds of multimodal systems are discussed, including talking heads and synthetic conversational agents. The next four sections discuss aspects of the various types of multimodal systems, according to which speech- and non-speech modalities speech input and output is associated with. Section 2.4 describes systems that combine speech input with information from the visual channel (face detection, face recognition, tracking of facial features, and lip-reading), Section 2.5 describes systems that combine speech with visual output (e.g. talking heads), Section 2.6 describes systems that combine speech input with other input modalities (defined as multimodal interface above), and Section 2.7 describes systems that combine speech output with other modalities (defined as multimedia systems above). These sections focus on concepts and issues; the details of the technology necessary to implement such systems is reviewed in Section 2.8. Finally, Section 2.9 presents established standards and common resources for multimodal systems.

2.1.3 Benefits of multimodal systems

2.1.3.1 Multimodal interfaces

What are the advantages of multimodal interfaces? A workshop on multimedia and multimodal interface design (Blattner and Dannenberg 1990) identified areas where user interface design can benefit from the use of multiple modalities:

- *Modality synergy*: Interfaces can benefit from modality synergy on both the input and output sides of the system. On the input side, interpreting input which is conveyed redundantly and/or complementary in several modalities can increase interpretation accuracy, e.g. combining speech recognition and lipreading in noisy environments.
- *Different modalities, different benefits*: Having several modalities available enables the system to get the specific benefit of each modality. Combined use

of speech and pointing can facilitate interaction, compared to speech-only interaction. For instance, deictic references to graphical objects are easier to express in pointing than in speech, and it is easier to speak commands than to choose from a menu using a pointing device. On the output side of a computer system, multimedia output is inherently more expressive than single-medium output (but one should take precautions to avoid cognitive overload of the user by stimulation with too many media).

- *New applications*: Some tasks are cumbersome or even impossible to perform if constrained to a single modality. For instance, interactive TV is much more compelling in spoken natural language dialogue with the system than pushing buttons on a remote control, or interaction with a WIMP (Windows, Icons, Menus, Pointing) interface. WIMP is attractive in other applications and for certain user populations, e.g. in word processors for Western languages and skilled typists.
- *Freedom of choice*: Although the same task may be achieved with equal efficiency using different modalities, users may differ in their modality preferences, and therefore there is value in being able to choose among different modalities. User populations may have different needs; for instance, handicapped users or users with other debilitating illnesses may not be able to use traditional input devices (keyboard and mouse).
- *Naturalness*: Offering multiple modalities to interact with a computer can be more natural to the human user if habits and strategies learned in human–human communication can be transferred to human–computer interaction. The mapping of user intention to input can also be more direct (Rhyne and Wolf 1993, p. 206). But 'natural' remains a rather vague term in this context.
- *Adaptation to either several possible environmental settings or evolving environments*: The possibility of switching from one modality to another (or combination of modalities) depending on external conditions (noise, light...) is also a benefit of multimodal systems.

These very broad claims need substantiation by empirical evidence, and explanation why and in what situation multimodal interfaces are superior. Both are on-going processes in the field. From our survey of multimodal interfaces (presented later in Section 2.2), we identified different domains where multimodal interfaces have been explored. The following summarises evidence gleaned from our survey on relative strengths and weaknesses that are inherent in modalities across tasks.

- *Speech input*: Speech input is preferable over traditional input modalities (keyboard and mouse) in tasks where either hands or eyes are occupied (e.g. car navigation), where mobility is necessary (e.g. equipment check-ups), or simply where speech input is more convenient (e.g. automated telephone services). Speech input is not preferable in intrinsically visual tasks (e.g. navigation, drawing tasks, object references).
- *Gesture input*: Gesture input is preferable in resolving deictic object references, and indicating the scope of commands.
- *Handwriting input*: Handwriting input can be more efficient for numerical data. Handwriting is preferable for note-taking and form-filling. Recently, small handheld devices which offer handwriting input, the most well-known of which is the PalmPilot, have become popular for note-taking, as electronic address books, for mobile fax, email, and world wide web communication, and for intranet communication. The difficulty of recognising handwriting input

varies greatly, and generally increases from isolated characters to words and sentences. The latter input type poses similar problems as the recognition of continuous speech.

In addition to these modality inherent factors, capabilities of current recognition technology and human factors of a particular interface have a great impact on modality preferences. Further research is necessary to establish relative strengths and weaknesses of modalities over a much wider range of tasks.
Unfortunately some disadvantages are also created with the use of multimodal systems. Some obvious ones are:

- increased system complexity (recognition and interpretation of several input streams)
- lack of general framework for multimodal systems, explaining where and how multimodality helps (i.e., in which application to use multimodality, what modalities to use, and how).

2.1.3.2 Benefits of multimodal speech

As R.H. Stetson stated in 1928 (Stetson 1928), "Speech is rather a set of movements made audible than a set of sounds produced by movements". Speech is the product of several activities: the configuration of the vocal cords, larynx and lungs, the movement of the lips and tongue. Its generation involves biomechanical commands to control organs and to contract muscles. The visual and audio channels are also associated with speech. The ear hears the sound while the eye sees the lip and tongue movements. These channels are the most common ones. But the tactile channel is also a speech medium. Blind people use their touch sense to understand spoken or written language using for example the braille or the Tadoma methods (users feel with their hands the speaker's articulators; see Section 2.1.4). In research and development, attempts are under way to integrate these various human senses in order to enhance the understanding and production of speech. The two other human senses, taste and smell, are not involved in speech.
Different studies have shown that considering visual signals as well as audio signals can improve speech intelligibility and speech perception (Risberg and Lubker 1978; Schwippert and Benoît 1997). The redundancy of audio and visual signals is exploited (Magno-Caldognetto and Poggi 1997; Hadar et al. 1983; Bolinger 1989). For example, an accent can be marked by any one of the following signals: the pitch of the voice, eyebrow raising, a head movement, a gesture, or a combination of these signals. At the same time, the interpretation of a signal from one modality can be modulated by other co-occurring signals. In American English, a raised eyebrow coinciding with a high utterance-final tone may tend to be interpreted as a question signal, rather than as the emotional signal of surprise (Ekman 1979).
Signals from visual and audio channels complement each other. The complementary relation between audio and visual cues helps in ambiguous situations. Indeed, some phonemes can be easily confused auditively (e.g. /m/ and /n/) but can be easily differentiated visually (/m/ is done by lip closure while /n/ is not). Looking at a face while talking improves human perception (Massaro

and Cohen 1990; Benoît 1990; Summerfield 1992). People, especially those hard of hearing, make use of gesture information to perceive speech. Similarly, speech recognition performance when combining the audio and visual channels is higher than when only one channel is used (Risberg and Lubker 1978). Some ASR systems increase their recognition performance rate and system robustness by considering both visual and audio signals (see Section 2.4.4).

2.1.4 Input modalities associated with speech

This section describes input modalities that are currently available for multimodal applications, and recognition devices that have been developed for these modalities. Future research may make additional modalities available, and this taxonomy may then have to be extended. We first describe multimodal aspects of speech input, followed by non-speech input modalities.

2.1.4.1 Multimodal speech input

Speech input allows the user to interact with an application using spoken words and utterances. Speech can be classified in terms of four categories:

1. continuous speech,
2. speech isolated by brief pauses between words (discrete speech),
3. isolated words, and
4. spelling.

Furthermore, different speaking styles can be distinguished, including

- read speech (e.g. a radio anchor reading out the news),
- spontaneous speech (e.g. two people talking on the phone), and
- hyperarticulated speech.

Different speaking styles are known to have a large impact on the accuracy of an automatic speech recogniser. As noted in Section 2.1.3, speech involves signals from various channels. Audio, visual and tactile channels each contribute to the transmission of speech signals.

- Audio channel:
  Common speech recognition systems can be classified as:[2]
  - *Discrete speech recognition systems*: the speaker has to separate each word by a pause which makes it easier for the system to recognise the enunciated word.
  - *Continuous speech recognition systems*: Fluent speech is more difficult to recognise; the end of a word is not easily distinguished from the beginning of the next word.

  Early ASR was done using only audio information. Background noise, tongue clicks, lip smacks, grunts, and (though less so for modern systems) particles like 'uhuh' or 'er', accompany speech and can disturb its recognition by machines. Speech recognisers work well in lab conditions but their performance drops dramatically in 'everyday' environments such as offices and public places. Speech recognisers have to recognise these background signals and disregard them. Low quality microphones or poor recording environments with much

[2] See also the COTS product evaluation Chapter of the Handbook. – Ed.

reverberation or low speech transmission quality can also distort the audio signal, resulting in deterioration of speech recognition performance.
Speech recognition is not limited to recognising what is being said but also how it is being said. Voice quality and intonation are important features of speech:

- Voice quality parameters co-occur with speech. Pitch, pitch range, loudness, timbre and tempo are examples of voice quality parameters. Pitch is the subjective property of a sound. It fluctuates as one speaks. Depending on the context, pitch range is a principal determinant of emotion and of social status (Scherer 1979; Ladd et al. 1985). It depends on the force of the air that is expelled from the lungs. Intensity and loudness are directly related to the magnitude of the effort involved in phonation. Timbre refers to sound quality, and tempo refers to the rate of articulation. It is often measured by the number of syllables per second, though languages differ systematically in their choice of timing units and syllables may not be equally relevant in different languages.
- A sequence of pitch accents constitutes the basis for an intonation contour. Intonation contours are ended with a phrasal tone that is eventually followed by a boundary tone (Scherer 1980; Pierrehumbert and Hirschberg 1990).[3] Intonation plays a crucial role in speech. It may convey information on the syntactic structure and semantics of an utterance. It is also related to the speaker's attitude and emotion.

The reader is referred to Gibbon et al. (1997) for detailed information on the techniques related to speech recogniser systems.

- Visual channel:
Methods that use visual signals to improve speech perception include:
  - *Lipreading*: It corresponds to the perception of speech when only visual cues are available. This is the case of hearing-impaired persons that rely only on visual signals. Lip shape, tongue position and teeth visibility are the cues distinguishing elementary visual speech units called *visemes*. Lipreading is also useful for the normal hearing (Risberg and Lubker 1978; Schwippert and Benoît 1997; Cerrato et al. 1997; Magno-Caldognetto and Poggi 1997). Visual cues can compensate for the loss of audio information in noisy environments. Noise up to 4–6 dB can be tolerated in speech understanding if one can see the speaker's lips (Summerfield 1992).
  - *Cued speech*: Hand shapes and hand placements produce cued speech. One hand is placed close to the lips, and changes shape in synchronisation with speech. The shape distinguishes the consonant while the hand placement serves for recognising vowels. The combination of speechreading, hand shape, and hand placements provide a unique representation of each phoneme. Contrary to sign-language based on words, cued speech is based only on phonemes. Speechreading communication is enhanced using cued speech as a supplement of information to distinguish visually confusable consonants and vowels. Eight distinctive hand shapes differentiate consonants and four hand positions near the mouth characterise the vowels that are confused auditively.
  - *Sign language*: Hand shapes, hand placement, hand orientation, and hand movement are the elements of sign language (Battison 1975).

[3] See also the Chapter on Dialogue Annotation in this Handbook. – Ed.

Body movement and placement of all other body parts in relation with each others, and also the environment setting (gestures towards, or representing people / objects) are also considered in the interpretation of sign language communication. Sign language has its own grammar and linguistic structure (Liddell 1980). Facial expressions, gaze and body movements also play an important role. Facial expressions are used for negation of a lexical item. Eye and head movements express agreement (Bahan 1996). Researchers in the field agree that facial expressions demarcate grammatical structures, specify the dependency of relative clauses, and determine the topic of a sentence (Liddell 1980). For example, an syntactically unmarked yes/no question is accompanied by a leaning forward of the head and shoulders, chin raising and raised eyebrows; a relative clause is distinguished by raised eyebrows, head tilted back and upper lip raised (Liddell 1980). Head nods can be a signal of topicalisation of a verb phrase. Some facial expressions and body positions co-occur with adverbs in American Sign Language (ASL); they modulate the sense of the adverbs. For example duration specified by an adverb can vary with the use of non-manual movements. Facial expressions are also used to communicate emotions. Facial expressions of surprise, anger, or happiness are superimposed to the hand signs. Facial movements may not necessarily have a semantic interpretaion but a pragmatic communicative function, for instance head nods. A head nod can be a signal of making a decision, or can be used to insert items in parentheses within the sentence, or can designate a first-person subject (Stokoe et al. 1965). Analysis of facial actions with linguistic significance during a signing session enhances the performance of sign language translation systems (Ichikawa et al. 1997).

- Tactile channel: Research has shown that deaf-and-blind people can perceive language through a tactile process, for example the Tadoma method. This method is based on tactile perception of the articulatory elements of humans. The user places a hand on the talker's face and neck to feel the facial movements associated with speech. Tadoma users are able to understand language at almost normal speaking rate. Braille and Optacon are alternative methods of perceiving speech via tactile perception, but there, the fingers are used to scan special keys to gather information. Pressure variation is another method involving the tactile sense and using the fingers: different variations in pressure are applied to the fingers. The variations form patterns that convey information.

2.1.4.2 Non-speech input modalities

Human senses as defined in the psychology field include sight, hearing, touch, smell, taste, and balance. Current (multimodal) computer input technologies model (in part) hearing, sight, and touch. Some believe balance will be useful in future systems, while smell and taste are probably useless in the context of human–computer interaction (HCI) (e.g. Schomaker et al. 1995a; Dix et al. 1998; et al. 1994). For instance, some electronic systems do exist for smell recognition (Technology n.d.), but they have not been used for HCI. We enumerate currently available non-speech input modalities that are available for human–computer interaction.

- *Pointing and gestures*: In this chapter we consider the following three kinds of gesture input: pointing, 2D gestures, and 3D gestures. We do not consider hand motions as a complete substitution of verbal communication (in the sense of sign language), nor body motions as means of emphasis in conversation. We define the different types of gesture as follows, from camera input. Pointing refers to using a pointing device, or a finger/pen on a touch-sensitive screen. 2D gestures (graphical marks, or simply gestures) refer to movements on a flat surface, for example marks drawn with a pen on a touch-sensitive display. 3D gestures refer to movements of fingers, hand, or head in the three dimensional space. Gesture input is captured using dedicated input devices, e.g. for 2D gestures and pointing: mouse and stylus, for 3D gestures: data glove, position trackers, or cameras. Pattern classification and computer vision algorithms have been developed to automatically recognise gestures.
- *Characters and handwriting*: Recognition systems for handwritten script and character input are classified in two categories according to the type of input: off-line systems recognise script that is presented visually as a whole (e.g. scanned text), and on-line systems recognise the trace of a character or word written on a digitising tablet or touch-sensitive display. Different categories or styles of writing can be distinguished: cursive and printed writing. Cursive writing is usually more difficult to recognise than printed writing, since cursive writing aggravates the character segmentation problem. Character segmentation in on-line recognition of printed handwriting is solved by means of the user interface.
- *Eye/gaze*: There have been three main approaches to the use of eye movement and gaze information in multimodal applications:

  1. improving speech recognition performance,
  2. employing gaze as an alternative pointing device (to refer to or select objects, or to control the mouse),
  3. and disambiguating references.

- *Lip movement*: People move their lips while speaking. Lip movements assist speech recognition: both in human–human communication (e.g. people hard of hearing can understand people just using lip movements) and human–computer interaction (augmenting speech recognition with lipreading increases recognition accuracy, especially in noisy environments).
- *Keyboard and mouse*: Keyboard and mouse input are still the most widespread input modalities for interaction with computer systems, particularly in direct-manipulation interfaces. Keyboard input ranges from a few dedicated buttons (e.g. on a remote control) to standard QWERTY keyboards. Mouse input ranges from "pressing" buttons to trajectories on the display (2D gestures).

2.1.5 Output modalities associated with speech

The main modalities explored for output in multimodal applications are:

- speech synthesis,
- face synthesis,
- talking heads (combination of speech and face synthesis),
- synthetic agents,
- force feedback, and
- traditional multimedia output (text, graphics, video, sound).

Talking faces and synthetic agents[4] in multimodal speech systems can be generated using a combination of the following output modalities:

- *Vocal tract*: Articulatory speech synthesis investigates the modelling of the vocal tract apparatus. Some speech synthesis systems are based on the relationship between the articulatory gestures required to produce a sound and the acoustic output of the speech (Rahim et al. 1993; Fang 1992; Salztman and Munhall 1989). The problem is to find a mapping between the acoustic parameters and the geometric parameters representing the vocal tract. Some examples of geometric parameters are:
  - tongue body center,
  - jaw angle,
  - lip height, and
  - lip protrusion (Riegelsberger 1997).

  On the other hand, given an articulatory model capable of output speech, some attempts have been made to teach a robot to speak (Bailly 1996; Badin and Abry 1996). This research involves two major approaches. The first one computes the vocal tract parameters using computer vision techniques as well as X-ray measurements; the second computes the behaviour patterns of the different speech articulators using inverse kinematics and/or dynamics techniques.
- *Acoustic generation*: Text-to-speech (TTS) systems generate speech from text. One approach to TTS is to store speech samples. A second approach to TTS is to decompose text into phonemes and compute the various parameters associated with them: fundamental frequency, formant duration, and the inherent stress of a word. Different algorithms have been proposed to generate the appropriate intonation for a given text (Prevost 1996; Davis and Hirschberg 1988; Hirschberg 1990; Monaghan 1991; Zacharski et al. 1993), generating natural sounding speech.
- *Optical generation*: As mentioned above, visual speech improves speech understanding and permits the hard of hearing to perceive speech. Interest in optical generation has increased in recent years. Different procedures compute the lip shape associated with speech (LeGoff and Benoît 1997; Cohen et al. 1996; Brooke 1996; McAllister et al. 1997; Yamamoto et al. 1997; Ezzat and Poggio 1997; Meier et al. 1997; Petajan 1984; Guiard-Marigny et al. 1996; Stork et al. 1992). Talking heads with facial expressions linked to the emotions, intonation, and personality have been developed (Beskow 1997b; Pelachaud et al. 1996; Takeuchi and Naito 1995; Waters et al. 1996; Koda and Maes 1996; Thórisson 1997). Attempts have been made to develop automatic cued speech systems (Cornett et al. 1977; Uschanski et al. 1992), and to simulate sign language (Loomis et al. 1983; Holden and Roy 1992). An interactive language training system is being developed for profoundly deaf children (Cole et al. 1998). Some systems use a speech-to-graphics conversion in a computer-game context with the purpose of training speech parameters such as voicing, loudness, pitch, nasality, and individual vowels and consonants, to hearing-impaired children (Arends 1993).
- *Mechanical generation*: Few attempts have been made to create synthetic mechanical faces that simulate the Tadoma system. The mechanical faces were built to reproduce the tactile communication scheme (Reed et al. 1992; Tan

[4] We define talking face/talking head as a synthetic face, while synthetic agents represent a whole persona including the whole body.

et al. 1989). They simulate lip movements (in/out, up/down) and jaw movement (up/down) (Tan et al. 1989). Oral airflow, tongue position and laryngeal vibration can also be incorporated (Reed et al. 1985, 1992). A text editor for blind people, MEDITOR (Bellik 1996), has been built. It is able to perform the main actions as a normal text editor. The system uses four input devices: a speech recognition system, a braille keyboard, a normal keyboard, and a mouse. A braille display, a speech synthesis module and a screen are the output devices of the system.

Multimodal speech production and perception requires synchronisation of audio and visual information, adding another challenge to multimodal speech systems. Even slight delays between both channels are detected by the listener and can give rise to confusion. More will be said on this in Section 2.5.3.
Non-speech output modalities include:

- *Text*: Text can serve as references. When a lot of information is given it is easier to be able to read the text again than to hear the same speech over and over again. The reason is that speech represents a high short-term memory load for the listener, while read text is persistently present in the visual field.
- *Image data*: Images, films, and animation may serve to illustrate a text and/or speech. In some applications, showing visual information can be more convincing and can make the speech content clearer.
- *Sound*: As in movies, sound and music can serve to accompany speech and/or visual data. Other non-verbal sounds, such as a beep sound or jingles can be used to attract the attention of the user.
- *Ambient sounds*: Soundscapes or ambient sounds, which last for a prolonged period of time, may be distinguished from brief sounds with a signalling function:
  - *Earcons*: Earcons are defined as abstract sounds for signalling. Short-lasting sound samples of a stylised or caricatural nature (to everyday sound events) are sometimes called earcons, by analogy with (visual) icons.
  - *Auditory icons*: Auditory icons are defined as natural or natural sounding sounds.

2.1.6 Taxonomies of multimodal applications

Since the range of possible applications that can benefit from multimodal input and output is still being explored, there is no agreement on categorising multimodal applications in the field. However, a preliminary taxonomy of multimodal applications is useful for a number of reasons: it provides the larger context for multimodal technology, and serves as a conceptual framework for discussing application-oriented issues. We present two preliminary taxonomies of multimodal applications:

1. first, a taxonomy based on the modalities available for input and output (as described above),
2. second, a taxonomy based on tasks supported by the application.

Other taxonomies of multimodal systems that focus on modality integration are presented in the context of modality integration technology in Section 2.6.2.

Since this chapter focuses on multimodal aspects of speech recognition technology, and since speech recognition applications tend to be interactive, our discussion is mainly concerned with interactive multimodal applications.

2.1.6.1 A modality-oriented taxonomy of multimodal applications

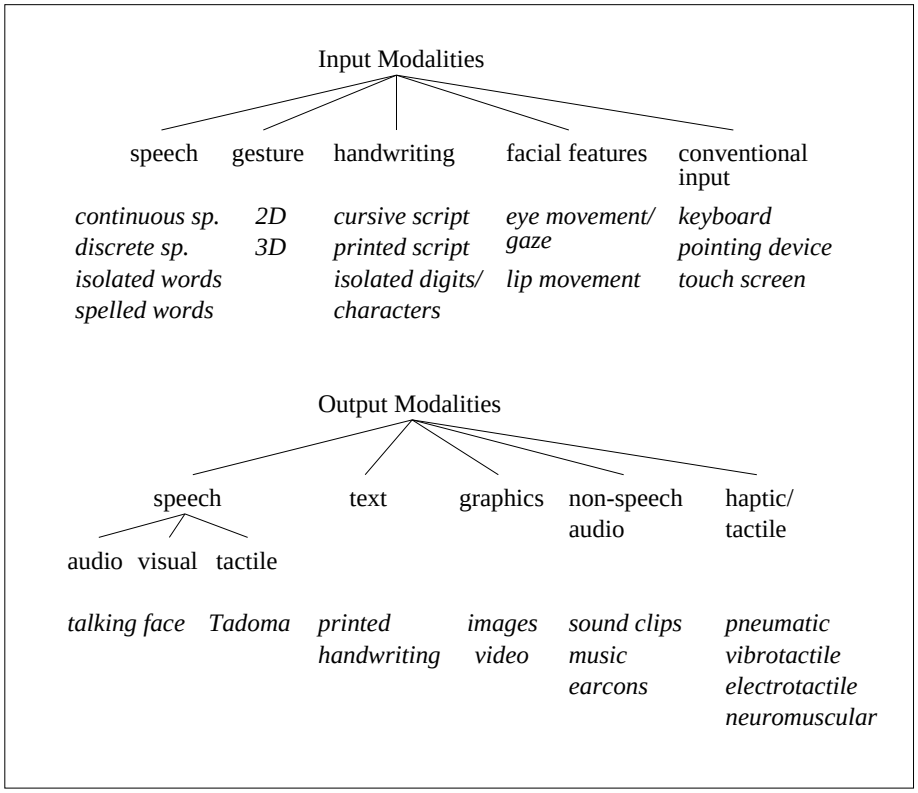

Figure 2.2: Modality-oriented classification of multimodal systems

A modality-oriented taxonomy categorises multimodal applications according to the kind of modalities that are offered for input and output, respectively. Since multimodal applications are likely to integrate multimodal input with multimodal output, each multimodal application is likely to cover several areas in the taxonomy which we present.
Different input and output modalities that are associated with speech input and output have been described in the previous subsections. Figure 2.2 illustrates the taxonomy of multimodal applications according to these modalities. Note that multimodal applications are typically characterised by a set of input and output modalities, rather than by a single leaf in this taxonomy tree.

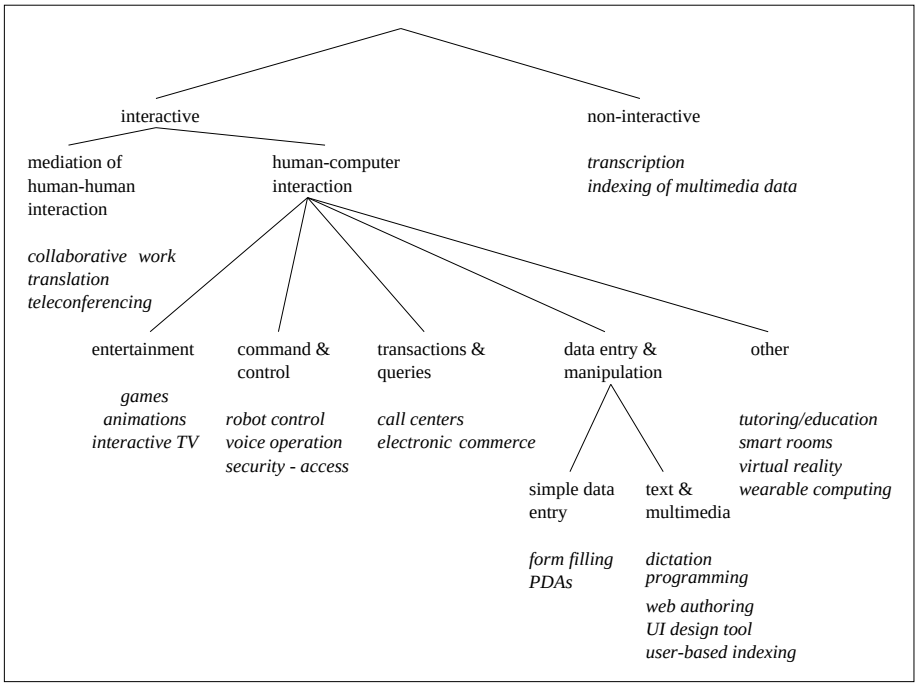

Figure 2.3: Task-oriented taxonomy of multimodal applications

2.1.6.2 A task-oriented taxonomy of multimodal applications

Tasks play a central role in the field of human–computer interaction. This section therefore proposes a taxonomy of multimodal applications that focuses on tasks supported by multimodal applications. Figure 2.3 illustrates the taxonomy and assigns published multimodal applications to the appropriate task category. Since an application may involve more than one task category, assigning applications to task categories may be ambiguous. In order to avoid ambiguities in cases where an application involves more than one task, the category is chosen according to an application's main task.

The taxonomy divides tasks, in speech recognition applications on the top-level, into *interactive* and *non-interactive* tasks. In non-interactive tasks, input that did not originate in direct user interaction is processed. In this case, the user does not want the application to accomplish a task as a result of the input, but to process input that was obtained elsewhere. Examples of such tasks include automatic transcription (e.g. of court proceedings) and automatic indexing of multimedia data (e.g. of radio and TV broadcasts). By contrast, interactive tasks process input that originates from direct user interaction. In this case, the user expects the application to accomplish some task as a result of the input. Examples include multimodal control of a robot, interactive TV, and directory assistance systems.

Interactive tasks can be further subdivided into *mediation of human–human communication* (e.g. a translation aid for foreign travel, teleconferencing tools,

support for collaborative work) and *human–computer interaction*, to access a service or some functionality offered by a computer system. There are many goals a user can pursue in interacting with a computer, such as entertainment, to have the system perform some task (command and control), to perform a transaction or retrieve information (transactions and queries), to create and manipulate data (data entry and manipulation), and other tasks. We provide typical examples of each of these categories.
Examples of entertainment applications include new interactive games, animation of artificial characters (for example, the recent popular movie "Toy Story"), and interactive TV. In command and control tasks, the user needs to initiate some action or control some process. In command tasks, the user issues concise commands to the system, typically single words or short phrases. Examples include controlling a robot by voice, or applications offering voice equivalents to menu and button interactions. Novel security systems will control access to buildings or services using multiple channels. In transaction and query tasks, the user engages in a spoken natural language dialogue with a dedicated device to access some service. For instance, standard telephone services such as directory assistance and call routing, as well as the call centers of many companies, are increasingly being automated using speech recognition technology. Another important future application domain in this category includes services related to travel, such as scheduling inquiries for different means of transportation (rental cars, trains, flights), booking of accommodation, and navigational support in foreign locales. In data-entry and manipulation tasks, the user creates and manipulates data that is stored in machine readable form. According to the complexity of the data, two subcategories are defined as follows:

1. 'simple data entry' deals with isolated words, digits or short phrases. Example applications are form filling, and personal assistants for addresses and note-taking.
2. 'Text and Multimedia entry' tasks support the production (or composition) of text, and multimedia material in general.

Dictation systems fall into the second category, as well as web authoring tools, and, in a more general sense, user interface design tools. Other tasks where multimodal interfaces are being actively researched include smart rooms (Maes et al. 1995, e.g. in MIT's media lab), education (Mostow et al. 1994, e.g. CMU's Listen project), and wearable computing (Rudnicky et al. 1996, e.g. at CMU). This taxonomy will be further developed as multimodal applications continue to emerge. Since it is a task-oriented taxonomy, some applications do not fit into a single category. For example, automatic processing of car rental requests involves filling out a form (to specify the type of car, rental period, rates), but may be accessed via telephone in a natural language dialogue. Therefore, it matches both the 'transactions & queries' and the 'simple data entry' category. Such ambiguities however do not diminish the usefulness of the proposed taxonomy in providing a conceptual framework for the discussion of application issues in multimodal systems.

2.2 Survey of multimodal systems

In addition to performing a survey of the field of multimodal systems based on research publications, the authors designed a survey questionnaire for email distribution in order to solicit feedback from the field. The questionnaire covered the following four main areas:

1. *Design and software implementation*: Specification methods, category of cooperation of modalities (see Section 2.6.2.1), architecture of the distributed multimodal system, methods of signal coding and storage, usability of engineering methods employed during the design process, and finally, toolkits and publicly available software.
2. *Hardware for multimodal systems*: Issues in and methods of sampling various modalities, transmission of input streams (audio input, pen/gesture input, and image input), issues in synchronisation of modalities, and methods of creating logs of multimodal interactions.
3. *Evaluation*: Basic evaluation methodology (evaluation criteria, measures, and measurement methods), types of evaluation (benchmark, informal user test in iterative design, user study with simulated or real system), experiment design issues, and methods of assessing qualitative issues.
4. *Talking face design*: Category of modelling technique (software, 3D measurement, scans, etc.), method of data acquisition, mathematical modelling of face, facial control for animation, and performance issues.

The low turn-out did not permit a formal analysis of the survey. Instead, results of the survey are worked into the main sections of this chapter.
From our survey of multimodal systems, we identified what modalities have been associated with speech input in published research systems. The following enumerates task domains and references for each modality combination that we identified from our survey.

- *Combining speech with pointing and 2D/3D gestures*: Combining speech with gestures is motivated by the fact that deictic references to objects are much easier to express using gestures than speech (Bolt 1980). Furthermore, gestures may be helpful in indicating the scope of operations. Research systems which combine speech with pointing or 2D/3D gestures include interaction with maps: city maps (Cheyer and Julia 1995), real estate maps (Oviatt et al. 1997), geographic maps (Koons et al. 1993), and calendars (Vo and Wood 1996). Other systems combining speech with pointing or 2D/3D gestures have been developed for graphical document manipulation (Faure and Julia 1993; Hauptmann 1989), analysis of video and image data (Cheyer 1997; Waibel et al. 1997), and flight control (Salisbury et al. 1990).
- *Pen-based interfaces*: Handwriting input in multimodal interfaces basically imitates editing using a pen and paper. Instead of pen and paper, the user writes with a stylus on a writable display (e.g. a touch-sensitive display). Handwriting input has long been considered as an alternative to keyboard input – without necessarily combining it with voice input. Pen-computing (or pen-based interfaces) has emerged as a field devoted to developing useful computer devices which are based on handwriting and gesture input. Despite the fact that, until now, handwriting is inherently a slow input modality (i.e. writing versus speech input for a specific task) and that the performance of current handwriting technology is considered too inaccurate, all studies exploring pen-based interfaces conclude that pen computing is promising for

future use (Briggs et al. 1992; Frankish et al. 1995; Rhyne 1987; Thomas 1987; Wexelblat 1995). Some recent successful commercialisations such as 3Com's PalmPilot, a personal digital assistant (PDA) with networking capabilities, provide further evidence of the attractiveness of pen-based interfaces. In addition to PDAs, applications conducive to pen-based interfaces include text editing ("electronic paper"), spreadsheets, and graphics. A recent overview can be found in Schomaker (1998).

- *Combining speech and pen input*: Combining speech with handwriting and gesture input has so far been explored for visual programming (Leopold and Ambler 1997), multimodal maps (Cheyer and Julia 1995; Waibel et al. 1997), and interactive correction of recognition errors (Oviatt and VanGent 1996; Suhm 1997; Suhm et al. 1996). Schomaker et al. (1995a) contains a section that discusses combinations of speech and pen-input both on the input and output side of a human–computer interface, but no prototype system has been implemented. It appears to be difficult to exploit the speech and handwriting multimodality in the case of text input, as opposed to the case of command input (Faure and Julia 1993). The natural "chunk size" of the input seems to be longer than an isolated word – which is easily handled from a technical point of view – but shorter than a complete sentence, in both modalities. Searching for correspondances becomes a difficult task, and usually the user is asked to solve mutual reference problems by clicking on menus, and so on. These extra actions make the interaction less natural and counteract the potential improvements in user-to-system bandwidth. The user performs a subjective cost evaluation with regard to the choice of modality. This evaluation can be modelled, taking into account recogniser performance and time (speed of production plus speed of processing) (Rudnicky and Hauptmann 1991).
- *Combining multimodal speech cues*: Lipreading has been successfully combined with speech recognition to improve word accuracy, especially under noise conditions (e.g. Bregler et al. 1993; Duchnowski et al. 1994). The technology of combining multimodal speech cues will be described in Section 2.4.4.
- *Combining speech with eye movement and gaze*: Three main approaches have combined speech with eye movement or gaze information. First, gaze information can be used to improve speech recognition performance. Since there is a natural tendency to look at objects while referring to them (provided the object is visible to the speaker), eye fixations may correlate with deictic object references during human–computer interactions. Therefore, eye fixations can provide hints as to what a user is likely to point out. For instance, when looking at a map, the user is likely to refer to objects that are within the local range of fixations (Sarukkai and Hunter 1997). Furthermore, information on the position of a speaker (gained from a face tracker) can improve headset-free speech input by means of microphone arrays: acoustic beamforming to localise a speaker in a room and to remove noise is more accurate if aided by visual information from the face tracker (Bub et al. 1995). Second, gaze information has been used for the selection and manipulation of objects, equivalent to mouse click and dragging operations (Flanagan 1997; Jacob 1993; Wang 1995). Third, eye fixations have been used to resolve object references, for example, in multimodal interactions with a map (Koons et al. 1993).
- *Combining speech input with speech synthesis or face synthesis*: Speech input has been combined with speech synthesis in two domains: first, a multitude of so-called dialogue systems that offer human–computer interaction in a natural language dialogue similar to human–human communication, and second in spoken language translation systems. More recently, due to significant progress in

the area of synthesising talking faces, research on integrating speech input with speech and face synthesis in talking faces has received increased interest. The ultimate goal of such research is animated intelligent humanoid agents that can both understand and express themselves in speech. Applications include animations for the film industry, and for telephone and other interactive services, and education. For example, a recent ARPA funded project has the goal of helping in the education of deaf children with synthesised lipreading tutors (TM 1998). The Olga project (Beskow et al. 1996) integrates conversational spoken dialogue (i.e., speech recognition and natural language understanding), 3D animated facial expressions, gestures, lip-synchronised audio-visual speech synthesis, and a direct manipulation interface. There is a multitude of literature on dialogue systems which integrate speech recognition with speech synthesis that is beyond the scope of this chapter and is reviewed elsewhere (see Cole et al. 1995; Gibbon et al. 1997). The following discussion will therefore not specifically mention dialogue systems. Examples of such dialogue systems include the various Air Travel Information Service (ATIS) systems developed as part of the ARPA Spoken Language Technology project (e.g. Levin and Pieraccini 1995; Pallett et al. 1994; Ward 1991). Other important multimodal applications that integrate speech recognition with speech synthesis include automatic interpreters, or speech-to-speech translation systems. They take conversational spoken language as input, translate it into another language, and use speech synthesis technology to speak the translated output (Morimoto et al. 1993; Roe et al. 1992; Waibel 1996). Enhancing the output with synthetic agents is expected to increase the realism and the usability of such translation tools.

In order to illustrate the discussion of technology and evaluation of multimodal applications that is to follow, Tables 2.1 and 2.2 summarise the results of our survey of the literature on multimodal applications relevant to the following aspects: input and output modalities, domain/application, types of modality cooperation, type of multimodal fusion (if applicable), multimodal architecture, qualitative and quantitative evaluation (if performed). In the "Fusion" column, the merging of the modalities can be done either at a signal, intermediate or semantic level. The abbreviations used in the tables are defined as follows:

- Input Modalities: ASR = Automatic Speech Recognition, GR = Gesture Recognition, H = Handwriting Recognition, P = Pointing, ET = Eye Tracker, GT = Gaze Tracking, K = Keyboard, 3D C = 3D Controller Device, OR = Object Recognition, SV = Speaker Verification, FR = Face Recognition
- Output modalities: GUI = Graphic User Interface, SS = Speech Synthesis, FS = Face Synthesis, S = Sounds, A = Audio Feedback, H = Haptic Feedback, VC = Video Conferencing
- Cooperation: E = Equivalence, C = Complementarity, R = Redundancy, CC = Concurrency, S = Specialisation, T = Transfer
- Fusion: sem = semantic, int = intermediate
- Architectures: MMI (Vo 1998), PAC-Amodeus (Nigay and Coutaz 1993), OAA (Moran et al. 1997)
- Measure: TCT = Task Completion Time

For illustration we briefly describe two multimodal applications for which images were available to the authors.
QuickDoc (Waibel et al. 1997) allows the user to view, manipulate, and summarise multimedia data using multimodal interaction techniques. A doctor

looks through a series of X-ray images or CT scans, identifies anomalous images, labels the appropriate regions in the image with the name of the disease or condition (using gesture to refer to the region), and attaches relevant comments through continuous speech dictation. The output is an HTML report that summarises the doctor's findings based on a listing of annotated images, the corresponding preliminary diagnoses, and automatically generated hotlinks to relevant Web sites. In terms of the taxonomies presented earlier in Section 2.6.2, QuickDoc is an example for concurrent complementary cooperation and synergistic fusion of modalities. Figure 2.4 shows a snapshot from user interaction with QuickDoc: the user has identified a region of interest with a circling gesture, and concurrently labels it with a voice annotation "This is Subdural Hematoma, confidence 90%".

Figure 2.4: QuickDoc application – User gesture with speech input "This is Subdural Hematoma, confidence 90%" (from Waibel et al. 1997)

The Multimodal Text Editor (Suhm 1997) allows the user to produce and edit text using several modalities, including continuous speech and handwriting. Modalities are used interchangeably to either produce new text, or to correct recognition errors and edit previously drafted text. By offering the possibility of switching the modality for corrections, strengths of one modality can compensate for the weaknesses of another modality to avoid repeated recognition errors. The multimodal text editor is an example of equivalent and specialised

modalities, and an alternate fusion of modalities. Figure 2.5 shows a snapshot from the multimodal text editor: the user inserts the word "handwriting" by using the handwriting modality.

Figure 2.5: Multimodal Text Editor – User inserting the word "handwriting" by handwritten input (from Suhm 1997)

2.3 Evaluation of multimodal systems

Multimodal systems are particularly challenging to evaluate. For that reason, and since the field is still emerging, there are few commonly accepted practices and standards. This section introduces the challenges and methodologies of evaluating multimodal systems by summarising what has been published. After a brief overview of different basic types of evaluation in Section 2.3.1, we discuss why the evaluation of multimodal systems is challenging. Then, Section 2.3.2 reviews evaluation methodologies that are suitable for multimodal systems. Basically, there are, as in other areas, two complementary approaches to evaluating multimodal systems: evaluation of the system components, and system-level evaluation. Evaluation issues in speech systems and recognisers for non-speech modalities are discussed in other publications, for example in Gibbon et al. (1997) and the Survey of Human Language Technology (Cole et al. 1995).

2.3.1 Types of evaluation

An early basic question must be asked before starting any evaluation process: "What is it that is evaluated?". Is it the system or one of its components? Is it the man–machine interaction process? Or is it the user (satisfaction, opinion etc.)? Depending on what goal an evaluation pursues, we can distinguish three broad categories of evaluation. The material presented in this subsection is based on Cole et al. (1995).

1. *Adequacy evaluations* determine the fitness of a system for a purpose: does it meet the requirements, and if so how well, and at what cost? The requirements are mainly determined by user needs. Therefore user needs have to be identified, which may require considerable effort in itself. Consumer reports are a typical example of adequacy evaluation.
2. *Diagnostic evaluations* obtain a profile of system performance with respect to some taxonomy of possible uses of a system. It requires the specification of an appropriate test suite. It is typically used by system developers.
3. *Performance evaluations* measure system performance in specific areas. Performance evaluation is only meaningful if a well-defined baseline performance exists, typically a previous version of the system, or a different technology that supports the same functionality. Performance evaluation is typically used by system developers and program managers.

As multimodal systems are still limited to research, adequacy and diagnostic evaluations to date play only a marginal role. If evaluation is performed at all, performance evaluations dominate. The following paragraph will therefore briefly summarise the basic methodology of performance evaluation.
Three basic components of a performance evaluation have to be defined prior to evaluating a system:

- *Criterion*: what characteristic or quality are we interested in evaluating (e.g. speed, error rate, accuracy, learning)?
- *Measure*: by which specific system property do we report system performance for the chosen criterion?
- *Method*: how do we determine the appropriate value for a given measure and a given system?

For example, for speech recognition, the criterion is typically accuracy, the measure is word accuracy, and the method is to align the output of a recogniser with the true hypothesis, counting the number of substitutions, insertions, and deletion errors.
Some criteria have emerged as quasi-standards for performance evaluations of different components of multimodal systems:

- Interactive services, data input applications: Task completion time and success rate, rate of unimodal versus multimodal interactions, complexity of interactions.
- Talking heads, synthesised agents: Intelligibility of speech output, realism of animation.
- Tracking of faces and facial features (e.g. eyebrows, gaze): Tracking accuracy (percent deviation from true position) and tracking success (ratio of time when feature tracked and time when feature is lost).
- Intelligent devices (and "smart" rooms): Success rate (how frequently is the

action that was intended by the user actually triggered), accuracy of user modelling.

A taxonomy of evaluation methodologies for interactive systems that distinguishes the basic approaches taken to evaluation defines the following three categories (Sweeney et al. 1993):

1. The user-based approach involves one or more users completing one or more tasks. Task, user, and environment characteristics must match those for which the system is being designed. Data on how user and system behave are collected while the user performs experimental tasks.
2. The theory-based approach involves a designer or evaluator who models task and user (some modelling techniques are described in de Haan et al. (1991); John and Kieras (1994)), based on the system specification. This ultimately generates quantitative values for interaction times, learnability or usability of the evaluated system. The evaluation involves neither a user–computer interaction nor a system prototype.
3. The expert-based approach involves an expert using the system in a more or less structured way, to determine whether the system matches predefined criteria or guidelines. The evaluation yields the evaluator's subjective judgement on the system's conformity to general human factors principles and approved guidelines.

Why is the evaluation of multimodal systems challenging? The evaluation of component recognition technologies is well developed in most areas, for example, speech, handwriting, gesture, and face recognition, as well as audio-visual speech synthesis. In contrast, evaluation of multimodal systems is difficult because:

- Standard benchmark databases (available for most of the component technologies) are only of limited use. The point of a multimodal system lies in the combination of different modalities. Since multimodal interaction is by nature application specific, there are currently no benchmark databases for multimodal applications. However, the available benchmarks for the component technologies are useful in evaluating the performance of the components of a multimodal application.
- Multimodal interaction is difficult to record under normalised, easily reproduceable conditions. Multimodal interactions depend on the user's behaviour and current hardware/software.
- The evaluation criterion is frequently unclear, in part since qualitative aspects play a significant role. This is in sharp contrast to components of recognition technologies, where accuracy (in different variations) is a commonly accepted criterion. The lack of commonly accepted evaluation criteria makes it difficult to compare across different evaluations of multimodal systems.
- The evaluation of qualitative aspects is difficult: user studies are very costly, and user self-reports are unreliable.

2.3.2 Evaluation methodologies

The following principle applies to the evaluation of any multimodal system: low-level evaluation of components and their integration has to be combined with a task-level evaluation of the overall system. This section outlines what methodologies are available for each of these.

2.3.2.1 Component evaluation

For the evaluation of the components of a multimodal system, evaluation methodologies that are accepted in the various subfields can be reused, including evaluation of speech recognition, handwriting recognition, and gesture recognition, as well as the evaluation of talking heads. In addition, the quality of the integration of the components in a multimodal system may have to be evaluated, for example, the accuracy of automatically assigning multimodal input to the appropriate (specialised) recognisers. The reader can find more material on the evaluation of specific component recognisers in other parts of this handbook or in Cole et al. (1995).

2.3.2.2 System-level evaluation

Different evaluation methodologies are available for system-level evaluation. All of them involve some form of user testing, either informal or formal, or during data collection (to build a database of multimodal interactions). System-level evaluation is therefore generally costly.
Depending on the kind of fusion of multimodal input, different criteria apply to system-level evaluation of multimodal applications. If the fusion of multimodal input events occurs at the signal level, the evaluation criterion can be similar to the evaluation of the components. For example, in audio-visual speech recognition, the obvious criterion for the integrated system is the same as for each of the components, namely recognition accuracy. However, if the fusion of multimodal input occurred at the semantic level, an appropriate metric has to be defined. Possible metrics include task level metrics (task completion time or success rate), naturalness, user satisfaction, cost, and unobtrusiveness. Except for task level metrics and costs, these criteria tend to be difficult to proceduralise.
The field of human–computer interaction developed several evaluation methodologies for interactive systems (cf. Shneiderman 1997) that are applicable to system-level evaluation of multimodal applications. Figure 2.6 shows a useful taxonomy of evaluation techniques (adopted from Balbo et al. 1993). There are three main categories of interface evaluation techniques (cf. Sweeney et al. 1993): experimental techniques that deal with real data observed from real users accomplishing real tasks with an actual system (benchmark, user study, simulation study, and iterative design in description below), predictive models that predict user behaviour and performance variables based on a theory or an empirical model, and expert evaluations.

- Benchmark evaluation: A benchmark evaluation requires a test set on which the performance of the multimodal application is evaluated. Benchmark evaluations are adequate when a criterion suitable for end-to-end evaluation can be defined. An example of end-to-end evaluation is the evaluation of speech-to-speech translation systems: they consist of speech recognition, natural language understanding, and (multimodal) speech synthesis modules; and these modules are connected as a processing pipeline. End-to-end evaluation therefore is particularly appropriate for speech-to-speech translation. As the criterion, the overall rate of acceptable (and intelligible) translations on a set of input utterances has been used (Waibel 1996). Since benchmark evaluations require a test set, some form of data collection must precede any benchmark

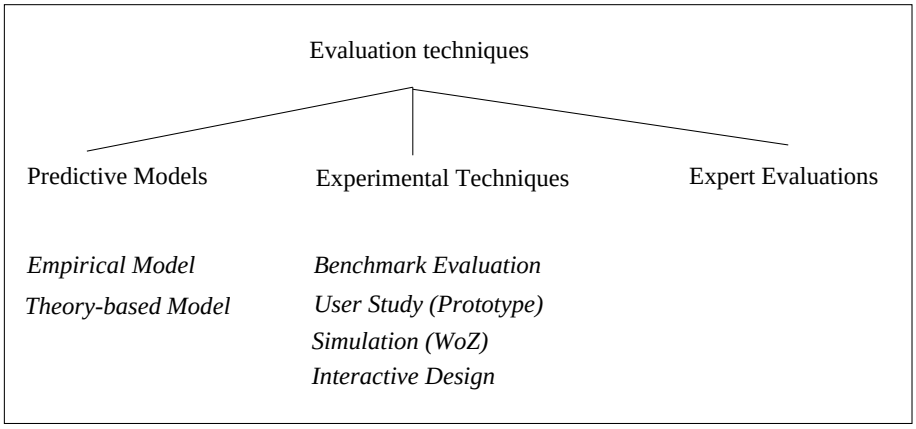

Figure 2.6: Taxonomy of system-level evaluation techniques, adopted from Balbo et al. (1993)

evaluation. Such data collection can occur during informal user tests (that occur as part of the regular iterative design cycle), or a specific effort can be dedicated towards establishing a database of multimodal interactions. Any such effort is to date however internal to projects; there is no publicly available benchmark of multimodal applications.

- User study of system prototype: If a prototype of a multimodal application has been implemented, informal or formal studies of users performing real tasks using the system can be performed. User studies typically yield a rich set of data, ranging from quantitative measures to informal observations. Additionally, user studies can be employed to build up a database of multimodal interactions for later benchmark evaluations. User studies however are quite costly, and require a careful experimental design of the study. Refer to standard literature for issues in experimental design and statistical analyses (e.g. Hayes 1993; Rosenthal and Rosnow 1991), and to Vo and Wood (1996); Suhm (1997) for examples of user evaluations.
- Simulation studies: Multimodal applications have the inherent problem that a working system is necessary to investigate their potential usefulness, on the other hand, a working system typically requires a lot of development work. Simulation studies can break this vicious circle: system performance that is not yet feasible can be simulated, and thus multimodal systems and issues in multimodal human–computer interaction can be examined without having to first implement a multimodal system. The Wizard-of-Oz technique is widely accepted for simulation studies. For a review of the Wizard-of-Oz technique, see Dahlbäck et al. (1992) and Gibbon et al. (1997), and for an example of simulation studies see Oviatt and VanGent (1996); Oviatt et al. (1997).
- Iterative design: Iterative design (or rapid prototyping) has been widely adopted in the field of human–computer interaction, especially for product development. It is suitable for the development of multimodal applications, since many detailed implementation issues can be explored rather quickly. The iterative design cycle includes (re)design of the application, implementation, and (informal) user testing. Iterative design is highly desirable from the HCI

point of view but is difficult to reconcile with the pipeline or cascaded process organisation in software development which, mainly for reasons of cost control, is currently still predominant.

- Predictive models: Predictive models are either based on a theory or on an abstraction from empirical observations. They predict user behaviour or important performance variables based on assumptions and model parameters (e.g. see Mellor and Baber 1997; Suhm 1998). They are useful since they allow the evaluation of multimodal interfaces at the design stage. Thus, a design can be improved before implementation. On the other hand, specifying data in a predictive model may be as time consuming as the implementation. Furthermore, model prediction may be wrong.
- Expert evaluations: In expert evaluations, an experienced professional uses a prototype or evaluates a specification in a more or less structured way, in order to determine whether the system matches predefined design criteria, or whether it violates established design guidelines and heuristics. However, experts are difficult to obtain for such evaluations, and several evaluators are necessary to discover a reasonable number of problems in the system design (at least three evaluators to discover about half of the usability problems in a design).

An elaborate discussion of different usability evaluation techniques, along with a lot of practical advice, can be found in Sweeney et al. (1993).

2.3.2.3 Qualitative issues

As mentioned above, qualitative issues are difficult to evaluate. Additionally, a variety of task-level measures have been used in the few evaluations that are reported in the literature, which makes it difficult to compare the utility across systems and tasks. Recent research in evaluation of spoken language system applications begins to address the issue. For example, PARADISE (Walker et al. 1997) is a framework for evaluating dialogue systems from a user point of view. It assumes that the ultimate measure of success for a dialogue system is user satisfaction. Since many different factors influence user satisfaction, depending on the application, PARADISE proposes to use statistical methods of determining the most significant predictor(s) of cumulative user satisfaction for a specific application, out of a large set of potentially useful variables. Predictor variables are categorised either as *task-based success measures* or *cost measures*. To maximise user satisfaction, maximal success has to be achieved at minimal cost. Cost measures are subdivided into efficiency measures (e.g. number of utterances, dialogue time, task completion time) and qualitative measures (e.g. ratio of inappropriate or repair utterances). After collecting satisfaction data from user evaluations of the multimodal application under study, a set of good predictor variables can be determined using multivariate regression analysis.

2.3.3 Specific evaluation issues

In this section we will present evaluation issues related to multimodal interfaces, especially those using talking faces.

2.3.3.1 Evaluation of lip shapes for talking faces

For the evaluation of lipreadable movements, Benoît and Pols (1992) propose intelligibility, naturalness, pleasantness, acceptability of the lip movements as evaluation criteria. Special care should be given to the computation of the main visible articulators involved in speech (tongue, teeth and lips). Particular attention should be paid to the representation of labials, dentals and alveolars where no ambiguity should be noticeable (Cosi and Magno-Caldognetto 1996). Two types of evaluation test have been proposed (LeGoff 1997): the quantitative evaluation test, based on measurements to test if the movements produced are correct or not; and the qualitative evaluation test, an approach based on perceptual tests to check how visual information is perceived. Both have different evaluation procedures. Ideally both evaluation tests should be performed.

- *Quantitative evaluation* compares computed values with real values. For example, values of lip height and lip width parameters of the synthetic face can be compared with the same values obtained from the analysis of a human subject. Image analysis or FACS can be used to analyse and compare muscle contraction from real and synthetic images. The weighting of different parameters and the definition of equalness in real and synthesised parameters is still a problematic open issue (e.g. lip width could be more important than upper lip raising) (Cohen and Massaro 1993; Benoît et al. 1996)?
- *Qualitative evaluation* tests the intelligibility of the system. The amount of intelligibility a synthetic model adds during speech recognition tests is compared to the amount of intelligibility a human speaker adds during the same tests (Guiard-Marigny 1996). The test is performed in different audiovisual situations: audio alone (degraded or normal audio), visual alone (of the synthetic actor and of the human subject), and audio-visual combined (of the synthetic actor and of the human subject). Benoît and his team also included the following conditions (LeGoff et al. 1996): lip alone of the synthetic face, jaw and lip alone of the synthetic face, subject's lips. The audio stimuli can be degraded by adding noise. For each setting a confusion matrix is established. The comparison over these matrices gives the overall intelligibility of each phonemic item in each setting.

2.3.3.2 Evaluation of talking faces

The design of synthetic talking faces is a very fast developing research area but is still in an early stage. Very few evaluation procedures have been considered up to now. Comparing the system with systems using only plain text, audio and still images, is a first step toward establishing evaluation criteria. Evaluation criteria include task completion time, subjective liking of synthetic agent, efficiency of information exchange (between user and agent), and error rate.
Evaluation tests usually compare different conditions, e.g. single modality with combined modality conditions (Takeuchi and Naito 1995).
For multimodal interface systems, an evaluation procedure has to check the exposure and the utilisation of the system (Waters et al. 1996). Exposure refers to the number of participants that use the system; utilisation corresponds to the percentage of time the system was used.

2.3.3.3 Evaluation of multimodal interfaces

In evaluating multimodal interfaces, both component-level evaluation of the various recognisers used and system-level evaluation are required. Without knowledge of the recognition performance of the input modalities that are supported by the application, task-level measures from a system-level evaluation, such as task completion success rate and task completion time, cannot be interpreted meaningfully. However, the recognition performance obviously depends on the current state of the art, and thus, results from system-level evaluations may become obsolete with each new version of the system. This problem can be circumvented by developing predictive performance models that abstract from recognition performance and interface implementation, and allow one to extrapolate results from empirical user tests and simulation studies. For a description of such performance models and their application to evaluating multimodal user interfaces, see Mellor and Baber (1997) and Suhm (1998).

2.3.4 Recommendations

- Unless benchmarks indicate sufficient performance of the components, two types of evaluation are necessary: low-level evaluation of the components and their integration, and task-level evaluation of the overall system.
- A combination of user-based empirical evaluation and theory-based predictions from a performance model can effectively compensate for the weaknesses of these two major evaluation methodologies. Model-based predictions can generalise the results of user studies, and abstract from the performance of available recognisers and the interface implementation. User studies provide rich data, and offer the possibility of addressing qualitative issues (e.g. in post-experimental questionnaires).
- System-level evaluation almost always requires some form of user testing.

2.4 Speech input with facial information (audio-visual speech recognition)

In this section we are interested in multimodal systems whose input modalities are speech and images recorded with a camera, followed by analysis. We will concentrate on systems that record faces and analyse either facial expression and/or lip movements. More details of the technology of analysing the visual channel associated with speech input can be found in Section 2.8.1.

2.4.1 Face recognition

Face recognition is a very active area of research in the computer vision field, and several recent surveys of the literature on face recognition exists (Chellappa et al. 1995; Samal and Iyengar 1992; Valentin et al. 1994). Applications of face recognition include: identification of criminals (mugshot matching), authentication in secure systems (e.g. for credit/ATM cards), locating faces for lipreading, and advanced telephone services. Since this chapter is mainly concerned with multimodal systems that integrate speech input and output, the discussion will be limited to the recognition and tracking of faces and other facial features either as a necessary first step for lipreading and speechreading systems (which will be reported later on in this chapter in Section 2.4.4), or

as an additional input modality, for example, for eye-movement based human–computer interaction (Jacob 1993). This section first provides a cursory review of face recognition and tracking (based mainly on material from Said and Tan (1996); Samal and Iyengar (1992)), then of recognising and tracking other facial features, in particular lips and gaze.
Samal Samal and Iyengar (1992) identified five basic problems in face recognition: face representation, face detection (i.e. determining whether a scene has any faces and locating the face), recognition (or identification) of faces (i.e., matching a face in the image with one of the known faces in the database, also called *face recognition*), analysis of facial expressions (i.e., model human emotions and correlate them with the facial features), and classification based on physical features (e.g. male or female, and age or race).

- *Representation of faces*: Two types of representation are commonly employed in face recognition and identification algorithms: 2D intensity images as a 2D array of intensity values (not very compact, but robust), and feature vectors. Two types of feature are used: features derived from intensity images, and features derived from face profiles. See Samal and Iyengar (1992) for complete listings of features. To achieve real-time performance, it is often necessary to keep multiple representations of the image data at different levels of detail, and apply computationally intensive algorithms to simple representations first, backing-off to the more costly complex representations only when necessary.
- *Detection of faces*: The face detection problem has been approached in two ways: locating faces as whole units, and locating faces by important facial features (e.g. eyes) (see Section 2.8.1.1).
- *Face recognition*: The two main ways of representing faces correspond to two approaches to recognising faces: feature-based matching and template matching. The first method extracts a set of geometrical features such as the relative position and size of the nose, eyes, mouth and chin. The second method compares images (represented as a two dimensional array of intensity values) with an initial set of images, using adequate metric measurements (see Section 2.8.1.2).
- *Face tracking*: Face tracking is distinguished from face recognition in that local rather than global search techniques are sufficient: since the movement of a head is typically slow relative to the frame rate, a head moves only a small distance from one frame to the next, and simple tracking algorithms can follow a person's motion in a video sequence. However, to track faces outside close proximity to the camera, the tracking system has to control the camera, including panning, tilting, and zooming. Face tracking algorithms first apply a face recognition algorithm to locate a face, and then local search algorithms to follow face motion within a sequence of video images.

2.4.2 Locating and tracking of other facial features

Beyond locating and tracking faces, other facial features may be useful in multimodal human–computer interaction, including eyes, gaze, and lips. Such nonverbal cues are useful for lipreading and multimodal user modelling. Finding and tracking the precise location of a facial feature and its shape is obviously more challenging than just locating whole faces. First, the resolution of images is frequently such that facial features are only a few pixels wide, so that one pixel difference represents already a substantial inaccuracy. Second, for tracking facial features simple tracking algorithms are not sufficient since features

can move (and change shape) substantially from frame to frame.
Many approaches to face detection and tracking are suitable for the detection and tracking of other facial features as well. For example, eigenfaces have been successfully applied to the problem of recognition and tracking of lips (termed eigenlips, Bregler and Konig 1994).

2.4.3 Automatic lipreading systems

Lip movements of a subject are recorded and analysed (Brooke and Petajan 1986; Blake and Isard 1994; Kausic et al. 1996; Bregler and Konig 1994; Goldschen et al. 1996; Adjoudani 1996; Cosi and Magno-Caldognetto 1996; Yuille 1991; Stiefelhagen et al. 1997a). Parameters defining lip shape (for example, width and height of the lip or lip protrusion) are extracted and the phonemic items associated with particular lip shapes are recognised. The first step is to locate the lips on the image. This can be done manually by placing a window on the mouth region, or automatically by marking particular points (such as corner of the lip, mid-point of the lip and so on) using reflective paper, by drawing a contour around the lip shape, or by using the chroma-key technique and painting the lips in a given colour. When the lips are located in the image the next task is to follow their movements and to extract the desired significant parameters. All these techniques have in common that a set of parameters is computed from the extracted features, including the inner area of the lips, the height and width of the lip opening, upper and lower lip protrusion. Principal component analysis (PCA) (Bothe 1996), Hidden Markov Models (HMMs) (Brooke and Scott 1994), eigenlips (Turk and Pentland 1991; Bregler and Konig 1994), and optical flow (DeCarlo and Metaxas 1996; Yacoob and Davis 1994; Mase 1991; Essa 1995) are other techniques of lip movement and facial expression analysis. The output of these techniques is a particular lip shape and muscle parameters that control lip movement.
Reported recognition rates for lipreading systems range from 100% for distinguishing 3 vowels (Brooke and Petajan 1986), to 70% on 10 digits (Goldschen 1993) and 25% on whole sentences (Pentland and Mase 1989).

2.4.4 Integration of audio and visual signals

As already mentioned, using audio and visual signals increase speech recognition performance. Few attempts have been made to integrate both signals to recognise speech (the first two in the following list) and lip shapes (the last one):

- *Neural Networks* and *Time Delay Neural Network (TDNN)* have been used to recognise speech (Meier et al. 1997; Stork et al. 1992; Yuhas et al. 1989; Rahim et al. 1993; Morishima and Harashima 1991; Lavagetto and Lavagetto 1996; Vogt 1997; Yamamoto et al. 1997). Acoustic and visual TDNNs are trained separately using a set of phonemes and a set of visemes, respectively (Meier et al. 1997). The combination of phoneme and viseme activations is obtained from the weighted sum of the individual activations. Recognition performance is much better when both channels are integrated as we can see on Table 2.3 (the results are for spelled letters). A typical neural network architecture is shown in Figure 2.8.

- *Hidden Markov Model-based audio-visual ASR systems* show better speech recognition results than systems using only audio information (Su and Silsbee 1996; Adjoudani 1996; Potamianos et al. 1997). Integrating audio and visual information in HMMs can be done by either using early or late integration (Su and Silsbee 1996). Early integration means that recognition is done using the combination of both signals. Late integration makes a first decision based on the separate signals and then takes the final decision based on the combination of both results (see Figure 2.7). At 0dB SNR Level, the audio-visual recognition rate varies from 62.3% (if visual recognition is based on geometric features) to 83.5% (if discrete wavelet transforms are used for the visual analysis). Above 16dB SNR Level, the results are 100% successful (for both image analysis techniques).
- *Fast Fourier Transforms* are used to analyse the speech spectrum. Mouth shape and the basic shape of the spectrum are correlated (McAllister et al. 1997). Mouth shapes described by three parameters (jaw position, horizontal and vertical lip opening) are predicted by analysing the speech spectrum.

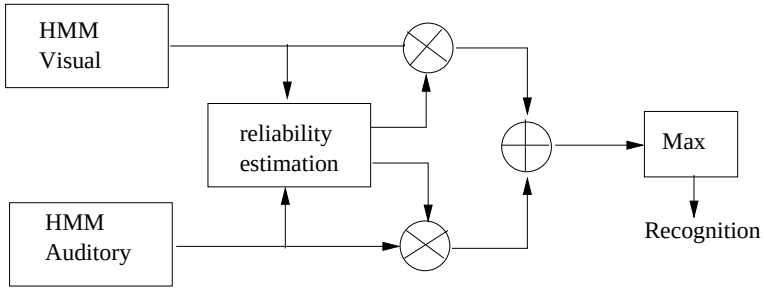

Figure 2.7: Late integration model (from Adjoudani et al. 1997)

2.5 Speech output with talking heads

Talking head systems model a synthetic agent with whom a user can communicate. Figure 2.9 (from Guiard-Marigny 1996) shows a general flow chart of a talking head system. Input models include camera, image, speech, keyboard input. First, appropriate recognition modules interpret the input. Then the module calculates the input parameters to be used for the computation of the control parameters. These control parameters are then used to drive the facial model and the voice synthesiser. Technical details of lip and face modelling can be found in Sections 2.8.2, 2.8.3, and 2.8.4.

2.5.1 Control techniques

Animating a facial model is a very difficult task. Human faces exhibit very subtle and complex motions that the face synthesis module has to imitate. Specifying each movement manually is feasible but time consuming and requires an experienced animator. Face synthesis techniques overcome this problem by automatically generating faces. Different approaches have been used to drive facial models in a multimodal speech system.

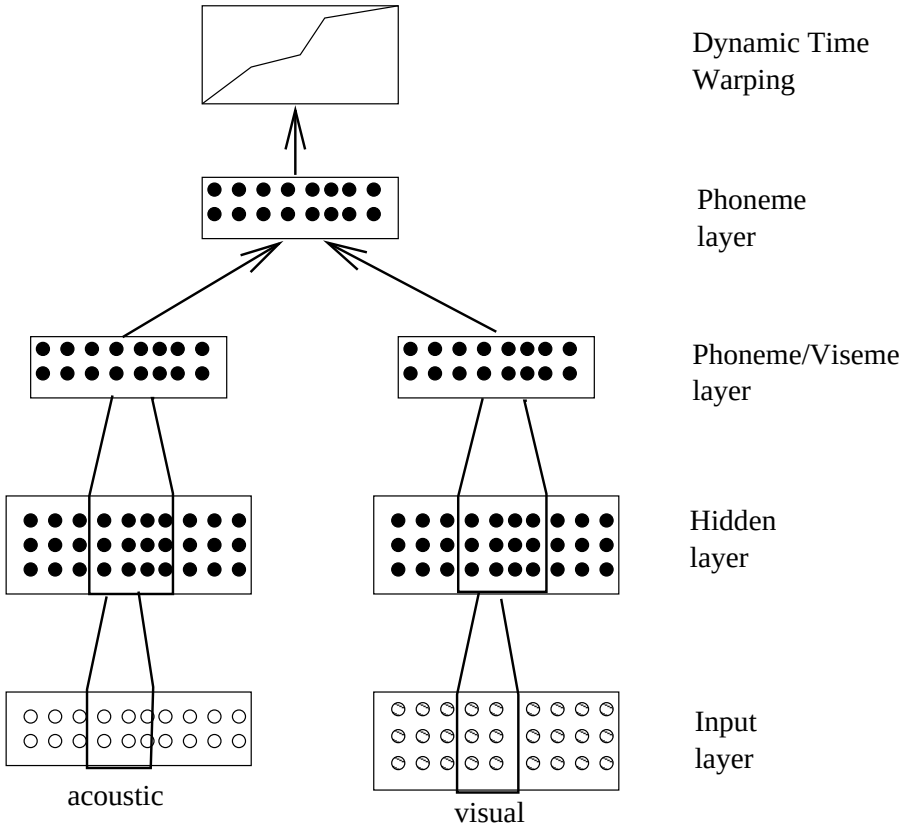

Figure 2.8: Audio-visual time delay neural network (from Meier et al. 1997)

- *Performance-driven system*: A person's movements are tracked and converted into parameters controlling the facial models (see Figure 2.10). Some techniques track reflective spots attached artificially to the person's face, others directly track the actor's facial features. A mapping is constructed from the extracted data and the facial model parameters. This method works well if the facial features or reflective spots are always visible. Using head-mounted cameras eliminates such a constraint since the reflective spots are always visible, but the display is even more obtrusive. Performance-driven face synthesis is well suited for reproducing one's actions. However, the facial model only knows how to mimic one's behaviour. No new animation can be done without having to record first the actor performing the actions, which can be a disadvantage for some applications (e.g. conversational systems). This technique is not easily adaptable to lip shape computation during speech when precise control of the lip movements is required. But replaying concatenated articulation sequences is less difficult and might be more appropriate in some applications (e.g. games).
- *Audio-driven*: Pre-recorded speech is analysed. Information about phonemes, pauses and their respective durations is extracted from speech. Additional

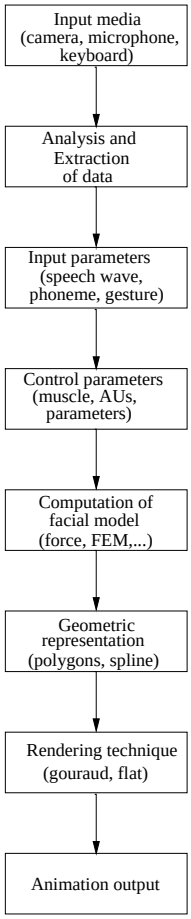

Figure 2.9: General flow chart of a talking head system (from Guiard-Marigny 1996)

paralinguistic vocal features (e.g. speech rhythm, intonation, loudness) can also be analysed. The reader is referred to Gibbon et al. (1997) for more information on speech analysis. When phonemes have been identified they are associated with facial control parameters to compute the appropriate mouth shape. Linear prediction analysis (Lewis and Parke 1987), sound segmentation (Nahas et al. 1988), TDNN (Lavagetto and Lavagetto 1996), HMM modelling and decoding (Yamamoto et al. 1997) techniques have been used to generate mouth shapes.

- *Puppeteer control*: A puppeteer moves input devices such as a data glove and joysticks, or uses a keyboard to drive a facial model (see Figure 2.11). Each input device control is associated with a facial parameter (Limantour 1994; Sturman 1998; Robertson 1988). For example, a key or a hand shape corresponds to a particular facial expression: raising the eyebrows or opening the mouth. As the puppeteer moves the hand or presses different keys, the facial model moves accordingly. This technique is often used for real-time

applications and movies.

- *Text-to-visual-speech*: The input of the system is plain text. The input text is first decomposed into its phonetic representation (Hill et al. 1988; Pelachaud et al. 1996; Beskow 1995; Kalra 1993; Nahas et al. 1988; Waters and Levergood 1993). Information about phonemes and their duration is automatically generated from the text. Formants and other speech parameters (frequency, pitch, pitch range and so on) are then computed. The text-to-visual-speech technique is suited when parametric facial models are used. Parameters defining facial animation are added to the set of speech parameters: lip shape, facial expressions, jaw rotation, etc. As a novel approach, speech synthesis systems have been extended to include facial parameters in their speech output parameters (Pearce et al. 1986; Hill et al. 1988). The parallel computation of the auditory and visual parameters ensures a perfect synchronisation of the two channels, which is an advantage of such a technique. But different sampling rates of the speech synthesiser and of the animation system have to be reconciled. While the animation system uses 25–30 frames/sec, an acceptable audio system requires at least 50–60 frames/sec. To avoid temporal aliasing effects of the visual images, motion blur between successive frames can be used. Parameter values driving the facial model are blurred with their neighbourhood parameters (corresponding to the precedent and successive frames), using a Gaussian filter (Hill et al. 1988). Text-to-visual-speech systems may be enhanced by adding markers describing intonation, speech rhythm, type of voice to the input text. Speech would be of better quality and such parameters could be used to get a more complex facial animation. For example, accents could be synchronised with raised eyebrows and head nods (Pelachaud et al. 1996; Beskow 1995). Different facial models corresponding to different types of voice have also been explored (Waters and Levergood 1993).
- *Conversational agent*: We also present another type of multimodal speech system that involves the development of a special agent capable of semi-autonomous actions such as taking decisions and conversing with a user. The agent can infer the user's state of mind and understand what he says. It is also able to make decisions, show emotions, and have a personality. To animate such an agent, one needs to select the appropriate verbal and non-verbal signals that accompany the agent's discourse: which words to utter with which intonation, what the gaze patterns and facial expressions are that emphasise the speech, how the agent's mental states and goals are derived and modulated from an understanding of the other conversant (here, the user), from the context in which the discourse takes place, and from the personality and relationship existing between both conversants. The variables that take part in determining behaviour patterns during a conversation are complex and their number is enormously large. A human–human relationship is extremely intricated. Simulating a conversational agent is therefore a big challenge. Several agents are being developed or have been developed (Thórisson 1997; Ball and Breese 1998; Churchill et al. 1998; Pelachaud and Poggi 1998; Badler et al. 1998; Binsted 1998; Cole et al. 1998; Rickel and Johnson 1998). A conversational synthetic agent consists of various modules: audio, visual and planning control. On the audio side, a speech recognition module recognises and understands what the user is saying, and a speech synthesis module generates the agent's spoken responses. In speech synthesis, intonation and paralinguistic elements enhance the agent's speech naturalness and intelligibility. Generating the appropriate intonation has to be based on a semantic analysis of the utterances. Finding which words are in focus, which parts of an utterance are

emphasised, is based on the syntactic information, but most of all on the cognitive/semantic information in what is being said. On the visual side, a facial recognition module analyses the user's nonverbal behaviour: emotional facial expression, conversational signal, etc. Other recognition systems may be incorporated in order to capture other non-verbal clues, including head movement, body motion, gaze pattern, and hand gesture (see Section 2.6.2). This data is used to emulate turn-taking protocols (Thórisson 1997), to call for the user's attention (Waters et al. 1996), and to indicate objects of interest in the conversation. Such an agent may also be used to teach language to hearing-impaired students (Cole et al. 1998). The conversational agent is able to interact with students. In particular, the agent will ask the student to repeat a word until he pronounces it correctly. In order to be convincing, an agent must be able to imitate complex and subtle human behaviour patterns: not only lip movements have to be synchronised with speech, but also gaze, head position and facial expressions. As for intonation, the computation of non-verbal signals is done at the syntactic level but is also based on semantic information and the inference of intentions (mutually, between system and user). For example, raising the eyebrows may mark an accented item or a question (Ekman 1979); looking away from the user and looking at the user are different actions of a turn taking act (Duncan 1974); hand gestures should finish with the agent's speaking turns (McNeill 1992) and lip movement should obviously be synchronised at the phonemic level. In order to achieve a realistic impression, agent behaviour has to be compatible with user behaviour. User modelling based on the recognition of speech, gesture and facial expression can be used (see Section 2.6.2). For all computations, close to real-time performance is necessary to ensure a natural dialogue between user and agent.

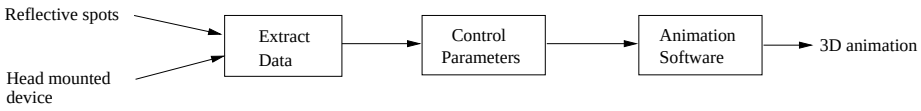

Figure 2.10: Performance-based animation control (from Parke and Waters 1996)

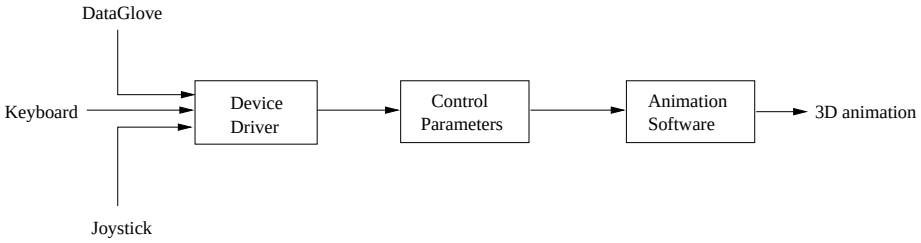

Figure 2.11: Puppeteer animation control (from Parke and Waters 1996)

Multimodal speech systems may include the following modules: speech recognition module (recognition and interpretation of what the user is saying), face recognition module (lip shape and expression), face synthesis module and speech

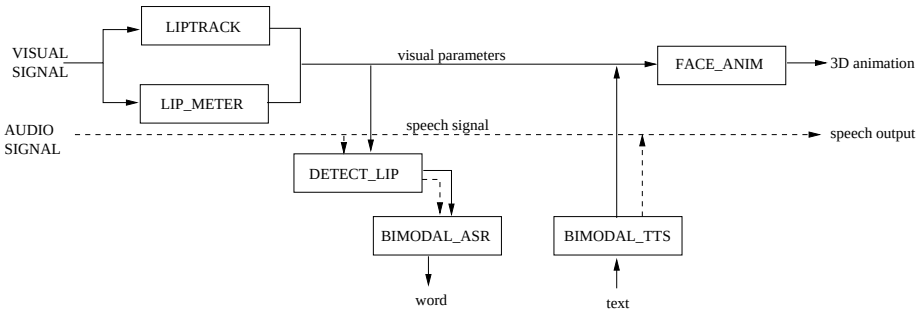

Figure 2.12: Overview of an audio-visual speech system (from Adjoudani et al. 1997)

synthesis module. Perceptive modules other than speech recognition may be necessary to capture additional modalities, e.g. head movements, and hand gestures. This will be described in Sections 2.8.5 and 2.8.6. A decision module is responsible for planning the actions of the synthetic agent, for deciding on the agent's behaviour and for selecting the nonverbal signals to be displayed (facial expressions, gaze, gestures) (Thórisson 1997).

2.5.2 Lip shape computation

The lip moves in complex ways. Many muscles are interwoven with each other around the mouth area. There is no simple mapping between a single speech sound and a lip shape. The same lip shape is involved in producing a variety of different speech sounds. Computing lip shape during speech is therefore extremely challenging. Many approaches have been developed:

- *Cartoon-type*: Early cartoon movies use a limited number of mouth positions. A relation is established between a speech unit and a mouth shape.
- *Set of parameters*: Research concerning lip movement during speech has shown that only a few parameters are necessary to describe lip shapes (Fromkin 1964; Benoît et al. 1990; Cosi et al. 1996).
- *Coarticulation*: Coarticulation arises from the temporal overlap of successive articulatory movements. Forward or backward influences on segments can happen. Forward coarticulation refers to the articulatory adjustment for one phonetic segment from an upcoming segment. Backward coarticulation refers to the articulatory adjustment of a phonetic segment over later segments. Good lip movements cannot be obtained by simply juxtaposing lip shapes, since lip shapes associated with a phoneme depend on the context given by the surrounding segments. A model of visual coarticulation has to be considered (Pelachaud et al. 1996; Cohen and Massaro 1993; Beskow 1995; LeGoff 1997).
- *Image-based*: The images of a talking face or of lip shapes are coded and stored in a library. Given a new audio signal, the problem is to select the most appropriate lip shape from the library.
- *EMG*: EMG electrodes are placed onto and around the lip area. Their measurements are used to get information about muscle contraction (Vatikiotis-Bateson et al. 1996). Artificial neural networks establish a mapping between muscle contractions and vocal tract articulators.

2.5.3 Talking heads: audio and video output synchronisation

To achieve a realistic talking head, the synchronisation of the audio and the video channel is crucial. Indeed the McGurk effect (McGurk and MacDonald 1976) describes the phenomenon that when we see a particular mouth shape (e.g. /ba/) but hear some other sound (/ga/), we perceive a blend of the two sounds (/da/). Humans are very sensitive to the synchronisation of audio and visual information. Moreover, different studies have shown that a delay of the audio channel over the visual one is noticeable if the audio channel is more than 130ms ahead of the visual one, and 260ms in the opposite case, that is, if the visual channel is being displayed before the audio channel (Dixon et al. cited in Guiard-Marigny 1996). Even tighter synchronisation requirements apply (75ms instead of 130ms and 188ms instead of 260ms) for sharp, transient noise such as a hammer knocking on a steel block (Dixon et al. cited in Guiard-Marigny 1996). The emission of phonemic items such as /p, b, m/ can be assimilated to a sharp noise. Asynchrony between sound and image starts to be perceived when the image anticipates the sound by 40ms or when the sound anticipates the image by 60ms (Summerfield 1992). This problem arises frequently in videophone technology where data transmission of the image in real time is an issue.
In some systems, the audio and the facial animation channels are computed separately and integrated when recorded onto a video tape (Nahas et al. 1988; Lewis and Parke 1987; Hill et al. 1988; Henton and Litwinowicz 1994).
Systems offering automatic interaction between the audio and visual channels (Waters and Levergood 1993; Adjoudani et al. 1997; Takeuchi and Nagao 1993) use the audio channel as the synchronous clock since the ear is more sensitive to delay. The audio module sends a signal to the image module. Then the lip shape corresponding to the sound is computed, and the audio signal and the image are presented. It is crucial not to miss important frames, for example, in the pronunciation of /p/, the system has to be sure to display the closure of the lips. In such a case the system selects to display a particular frame associated with the given sound (Takeuchi and Nagao 1993).
Since computation can be high in mainly multimodal speech systems, several computers can be involved and therefore need to communicate with one another. Parallel Virtual Machine (PVM) is a protocol of communication that can be used when different machines are involved in the system architecture (LeGoff 1997). To ensure synchronisation between both channels all computations for one frame are completed before going to the next. The system clock is used to synchronise the audio and the visual channels.

2.6 Speech input with modalities other than faces

This section describes multimodal interfaces that combine speech input with other human communication channels, as defined in the introduction of this chapter. First, Section 2.6.1 briefly reviews component technologies necessary to build such multimodal applications, and go beyond multimodal speech input and output (which was described in the previous section). Besides planning and dialogue managing modules, these are recognisers for different modalities, including speech, handwriting, gestures, and facial features. Then, Section 2.6.2

describes integrated multimodal systems: taxonomies of integrated systems (Section 2.6.2.1), and how different component recognisers are integrated. Such components are integrated for two main reasons: either to provide more flexible input to computer systems (Section 2.6.2.2), or to model user behaviour or intentions in multimodal user modelling (Section 2.6.2.4). For a summary of published integrated systems see Section 2.2, for evaluation issues of multimodal interfaces see Section 2.3. Details of the technology and reviews of state-of-the-art recognisers of on-line handwriting and gesture recognition can be found in Section 2.8.5 and 2.8.6.

2.6.1 Recognition of non-speech input modalities

Multimodal component technologies model human perceptual (or cognitive) skills in order to make the multiple information channels that people naturally employ available for human–computer interaction. The five human senses are: hearing, sight, taste, smell, and touch. To date, research has focused on imitating hearing and sight. Out of the multitude of studies on the various instances of these two modalities, the subsequent sections focus on the recognition technology for modalities that have been associated with speech recognition: on-line handwriting recognition and recognition of 2D and 3D gestures (Section 2.8.6), and recognition and tracking of faces, facial features, including gaze and lip regions (Section 2.4.1). Robust speech input using both audio and visual information (*speechreading*) has been reviewed in Section 2.4.4; the detection, recognition, and tracking of faces will be described in the technology section of this chapter (see Section 2.8.1).

2.6.1.1 On-line handwriting recognition

On-line handwriting recognition systems transform handwriting input, given as sequences of two dimensional coordinates, into text. Handwriting input can be at the level of characters, words, and sentences. To capture on-line handwriting input, tablets, touch screens, and light pens are used. However, currently available products still suffer from various usability problems. It is possible to achieve good handwriting recognition performance by retraining current HMM-based continuous speech recognisers, but only after modifying the preprocessing module appropriately. The feature vector needs to be adapted to the particular requirements of the HMM approach. Specialised on-line handwriting recognisers however achieve better performance. Most state-of-the-art systems implement the so-called *analytical approach* to handwriting recognition: the handwriting input is first evaluated on the level of constituent characters, and, in a second step, higher-level word or sentence hypotheses based on this information are identified. The best published writer independent recognition accuracies for analytical handwriting recognition systems are more than 95% for character recognition (Guyon et al. 1992), 93.4% for word recognition (with a 20,000 word vocabulary) (Manke 1998), and 86.6% for sentence recognition (with a 20,000 word vocabulary) (Manke 1998). However, these recognition rates appear to be overestimated, since humans reach only 50–70% recognition on isolated words written in free style. The high recognition accuracies reported above may be due to adaptation of heuristics and parameters in frequent iter-

ations of training and testing without changing data sets. For more details on the technology of on-line handwriting recognition, see Section 2.8.5.

2.6.1.2 Gesture recognition

Gesture input can mean many different things: pointing on the screen with an appropriate pointing device, 2D gestures drawn with a pen on a writable display or a tablet, and movements of fingers, hands, or the body in the three-dimensional space (called 3D gesture in this chapter). While pointing does not require recognition beyond identifying which displayed object the user wants to refer to, recognising 2D or 3D gestures is a typical pattern recognition problem. Standard pattern recognition techniques, such as template matching and feature-based recognition, are sufficient for gesture recognition. Input devices for gesture input include standard pointing devices (for pointing input), touch screens or tablets for 2D gestures, and data gloves or cameras for 3D gestures. Gesture recognition with standard methods is more than 90% accurate, provided the set of gestures is small (less than 20). For more details on gesture recognition, see Section 2.8.6.

2.6.2 Integration in multimodal applications

Multimodal input to computer systems offers several advantages, including higher accuracies for automatic interpretation (e.g. lipreading...) and more flexible input (e.g. using speech and gestures to refer to objects when interacting with maps). First, Section 2.6.2.1 presents different taxonomies of modality integration and clarifies important concepts. Section 2.6.2.2 discusses the technology of combining input events in different modalities, mainly on the semantic level. Section 2.6.2.4 briefly outlines the application of multimodal systems to the modelling of user behaviour and intentions. Relevant to integrated multimodal systems, but positioned among the common resources and guidelines section within this chapter, Section 2.9.4 describes architectures and toolkits for building multimodal applications.

2.6.2.1 Taxonomies of modality integration

This section summarises two taxonomies of multimodal systems that focus on central issues of integration of modalities in multimodal applications: What is multimodality? How are different modalities supported by a particular system? How do types of information from different modalities relate to each other, and how are they combined?

The first taxonomy is based on two dimensions: types and goals of cooperation between modalities (Martin 1997; Martin et al. 1995). *Modalities* are defined as "ways of exploiting a specific physical device enabling the exchange of information between user and computer system", and *multimodality* as "the cooperation between several modalities in order to improve the (human–computer) interaction". The 'goals of cooperation' describe the requirements of a human–computer interface in terms of improving interaction: by making the interaction more accurate, more intuitive, more efficient and adaptive to different users and environments. Six different types of cooperation are distinguished (for formal definitions see Martin et al. 1995):

- Complementarity: Different chunks of information belonging to the same command are transmitted over more than one modality (e.g. "put-that-there", while pointing at an object, and then at a location (Bolt 1980)).
- Redundancy: The same chunk of information is transmitted using more than one modality (e.g. a costumer saying "I want the second item on the right", simultaneously pointing in that direction).
- Equivalence: A chunk of information may be transmitted using more than one modality (e.g. option to choose from a menu by either mouse or voice selection).
- Specialisation: A specific chunk of information is always transmitted using the same modality (e.g. an information kiosk offers different services which are selected by touching the corresponding button). Specialisation may also manifest itself in user preferences, for example, if users consistently prefer speech over other input modalities for certain tasks.
- Concurrency: Independent chunks of information are transmitted using different modalities and overlap in time (e.g. talking over speaker phone while editing a document). Concurrency means parallel use of different modalities to initiate different actions.
- Transfer: A chunk of information in one modality triggers an event in another modality (e.g. in hypermedia interfaces: a mouse click causes an image to be displayed).

The second taxonomy suggests a design space for multimodal systems in terms of concurrency of processing and type of data fusion (Nigay and Coutaz 1993). Multimodal systems are characterised along three dimensions: levels of abstraction, use of modalities, and fusion.

- *Levels of abstraction* refers to the multiple levels at which data from a particular device can be processed. The level of abstraction ranges from the signal level to the semantic level. On the signal level, no interpretation has taken place yet. Lipreading is an example of signal level integration of modalities: the audio and image signal are merged on a low level, to obtain more accurate interpretation of the combined event. At the other end of the spectrum, the classic example of "put-that-there" requires the separate interpretation of speech and gesture input, and the merging of information on the semantic level. The distinction of modality integration according to the level of abstraction is equivalent to the distinction of tight coupling versus loose coupling of modalities (e.g. Sarukkai and Hunter 1997).
- *Use of modalities* refers to the temporal availability of multiple modalities. Modalities can be used either sequentially or in parallel.
- *Fusion* refers to the possible combination of different input events. Multimodal input events can either be interpreted independently, or they can be merged.

A multimodal system is described by a set of features (e.g. the commands it supports) which are located in the design space and are assigned a weight (e.g. frequency of use). The position of the whole multimodal system in the design space is the pivotal center of its features. According to the characterisation of an interaction along the two dimensions 'fusion' and 'use of modalities', four basic types of multimodal interaction can be distinguished: alternative, synergistic, exclusive, and concurrent multimodal interaction, as shown in Figure 2.13.
Obviously, synergistic systems subsume the other three classes of multimodal systems. Therefore, architectural models of multimodal integration (as pre-

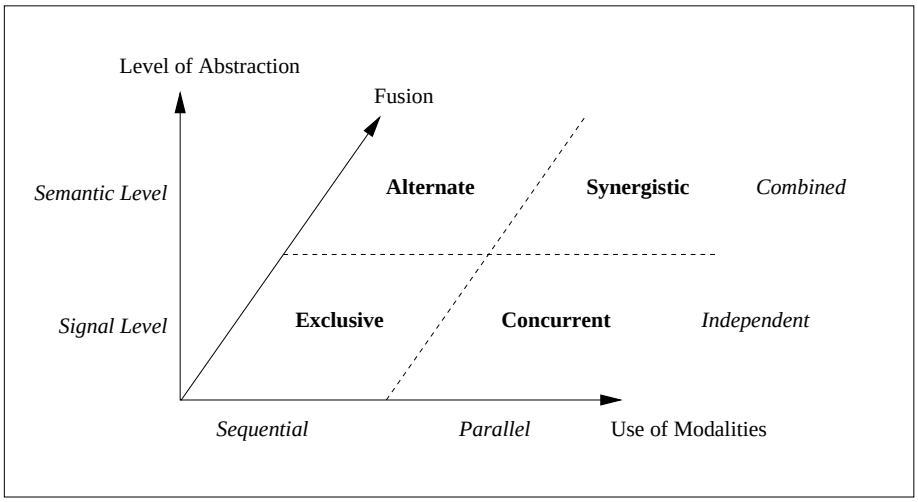

Figure 2.13: Multimodal Design Space (from Nigay and Coutaz 1993)

sented in the next subsection and in Section 2.9) are sufficient if they are able to model synergistic cooperation of modalities.

2.6.2.2 Fusion of multimodal input

Fusion of multimodal input events can occur on different levels, ranging from signal level to semantic level. *Signal-level fusion* (or *lexical fusion*) (Nigay and Coutaz 1993) performs the combination of multimodal input at the level of the input signal. Signal-level fusion has to date been tried for audio-visual speech recognition, combining speech as audio signals with lip movements as visual signals (see Section 2.4.4). Other types of signal-level fusion have been explored in the robotics field (e.g. combining image data with other sensor input, such as laser ranger finders, or infra red sensors), but that is beyond the scope of this chapter. *Semantic fusion* performs the combination of multimodal input at the meaning level. This raises the question of how to assign meaning to a multimodal input event. For multimodal applications that combine speech input with other modalities for more flexible input to a computer system, the emphasis is on the way the system responds to (multimodal) user input. For such applications, the meaning of a multimodal input event is commonly defined as the (parametrised) action that the application should perform in response to the input event (cf. Vo 1998). The remainder of this section on fusion of multimodal input will describe how semantic fusion can be realised, interpreting multimodal input events to trigger actions in the multimodal application.
Semantic fusion of multimodal input proceeds in two steps. First, input events in different modalities are combined in a low-level interpretation module by grouping input events in different modalities to *multimodal input events*. Next, the multimodal input event is passed on to the *high-level interpretation module* to derive the meaning of multimodal input events by extracting and combining

the information chunks. Thus the high-level interpretation module determines what type of action the user wants to trigger, and what its parameters are. This parametrised action is then passed to the application's *dialogue manager* which can initiate the execution of the intended action.
Two issues in fusion of multimodal input remain: How are multimodal input events represented? How is the meaning of a multimodal input event derived? Different approaches to the representation of multimodal events have been proposed, but no standard has yet emerged:

- Typed feature structures (Johnston et al. 1997; Cohen et al. 1997): Multimodal inputs can be transformed into typed feature structures that represent the semantic contributions of different modalities. These typed feature structures are then combined by unification operations.
- Syntactic representation (Faure and Julia 1993): Multimodal input events are represented as triplets {verb, object, location}. This representation is sufficient for speech input with deictic references expressed as gestures, but it is unclear how it can be generalised to other multimodal events.
- Melting pots (Nigay and Coutaz 1995): A *melting pot* encapsulates types of structural parts of a multimodal event. The content of a structural part is a piece of time-stamped information. Melting pots are built from elementary input events by different fusion mechanisms: microtemporal, macrotemporal, and contextual fusion. *Microtemporal fusion* combines information units that are produced simultaneously or very close in time. *Macrotemporal fusion* combines sequential information units in temporal proximity when the information units are complementary. *Contextual fusion* combines information units based on semantic constraints. Figure 2.14 shows an example of macrotemporal fusion of a speech event "Denver" and a pointing event "Boston" (the user pointed to Boston on a map) to a query "from Boston to Denver".
- Partial Action Frames (Vo 1998; Vo and Waibel 1997; Vo and Wood 1996): Input from each modality is interpreted separately and then parsed and transformed into semantic frames containing slots that specify command parameters (parameter slots). The information in these (partial) action frames may be incomplete or ambiguous if not all elements of the command were expressed in a single modality. A domain-independent frame-merging algorithm combines partial frames into complete frames. Each grouped sequence of input events is assigned a score based on their mutual information. A dynamic programming algorithm (similar to Viterbi search or Dynamic Time Warping used in speech recognisers) determines the best sequence of input event interpretations that fit the whole multimodal input event. This frame merging architecture is called multi-state mutual information network (MS-MIN).

2.6.2.3 Generalised input devices

On a more abstract level, multiple input modalities can be viewed as generalised input devices (Schomaker et al. 1995a). A multimodal computer system offers different ways of conveying the information necessary for the computer to perform the desired actions. The choice between different modalities is determined both by user preferences and by application constraints. For example, cancelling an object by voice or by mouse has the same effect, but the choice of action (voice vs. mouse) can either be forced by the application (e.g. speech in a hands-busy application), or determined by user preference.

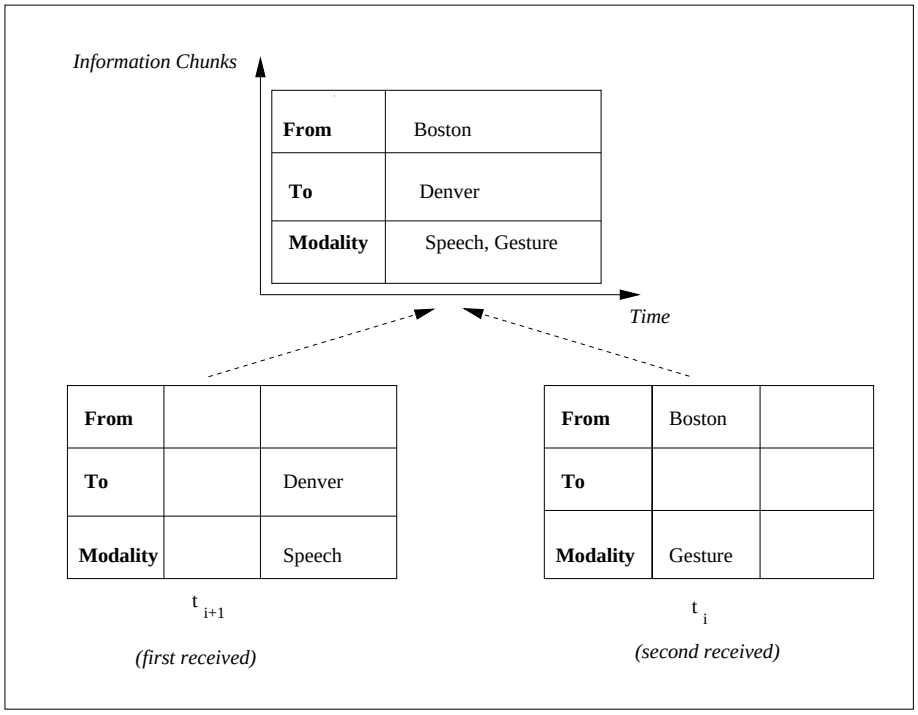

Figure 2.14: Fusion of two melting pots (from Nigay and Coutaz 1995)

In generalised input devices, the translation from input actions performed by the human user to input events that can be interpreted by the application is modelled as generalised device drivers. For modalities that are automatically interpreted, such as speech and pen input, these device drivers include recognition algorithms. Figure 2.15 illustrates the situation.

2.6.2.4 Modelling of user intentions

Modelling of user intentions attempts to track user behaviour and intentions in the context of human–computer interaction, using multimodal input channels. Multimodal user modelling is based on adequate multimodal recognition components, including gesture recognisers, prosodic and facial expression detectors. This research is motivated by the observation that situation awareness plays an important role in human–human communication, including back-channel utterances (which often provide feedback on whether the conversation partner follows the speaker's argument), and turn-taking mechanisms. By recognising situations and knowing what people usually do in those situations, the system can be aware of the user's intentions and be able to predict what would be likely to come next. Also the system can intervene with clarification requests whenever the user's behaviour does not match the system's predictions. Obviously, such user modelling would be very useful in making synthetic agents (animated with a talking face) more human-like. Research in this area is however still at

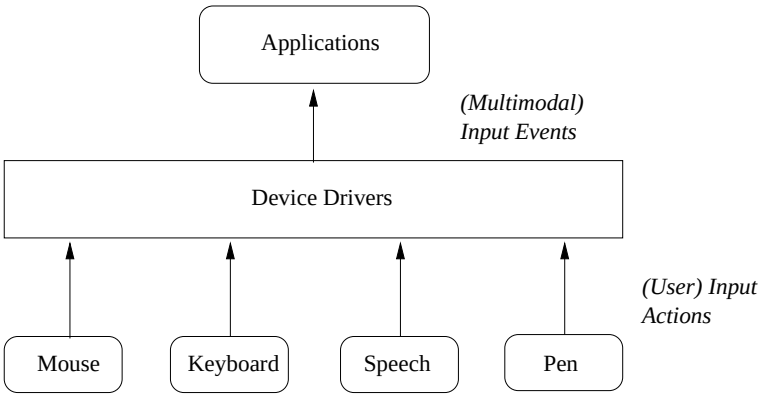

Figure 2.15: Generalised input devices (from Schomaker et al. 1995a)

an early stage.

2.6.2.5 Recommendations

- Input capture devices and recognisers should preserve timing information.
- Modality integration: Temporal proximity of multimodal events is the single most useful factor to group multimodal events for complementary interpretation. However, temporal proximity must be correlated with semantic coherence in order to avoid extracting meaningless multimodal events.
- System architecture for multimodal systems: Client–server architectures are almost indispensable in order to distribute the computational burden of processing multiple modalities and to increase the maintainability of a system consisting of several different recognisers. From a software engineering point of view, object-oriented programming is strongly recommended.
- Combination of isolated input events with multimodal input events should be encapsulated in a separate module.
- Recognition technologies: Recognition of speech and pen-input (gestures, handwriting) are currently reliable enough for small vocabulary tasks (e.g. controlling tasks in the user interfaces).
- Adequate error correction and dialogue management methods are necessary to compensate for unavoidable recognition errors.
- Information routing: Limit exchanges between different modules to messages containing semantic representations of the data, not the data itself, thereby avoiding high-bandwidth communication across low-bandwidth networks.
- Barge-in should be allowed in any multimodal interaction.

2.7 Speech output in multimedia systems

The goal of this section is to establish common problems that multimedia systems including speech output are facing when being developed, or when a manager has to decide which type of multimedia system to buy that would best fit the needs.
In a first section, we introduce a taxonomy of output modalities. Then we concentrate on the theoretical problems associated with multimedia systems with speech output. Multimedia concepts and standards are described. The specific

features of the speech output media/modality are discussed. The questions addressing the integration of speech output in multimedia systems are tackled and corresponding recommendations are provided. In a second section, some details on technical issues are discussed.

2.7.1 Taxonomy of output modalities

Bernsen (1997) proposed an output modaliy taxonomy that is based on the different representations of output modalities:

- *Linguistic representation*: Linguistic representation contains, for example, syntactic, semantic, and pragmatic information. This representation is based on a high level of abstraction and is not able to give relevant details for distinguishing specific entities as would an analogue representation. The string 'my book' distinguishes a particular book from other books within some utterance context, but it does not give any further specific information, for example, title, author, size, and collection.
- *Analogue representation*: Analogue representation is often complementary to the linguistic one. It is sometimes also called 'iconic' representation. It is based on the particular characteristics of the object it represents. Image, sound, graphics and haptic devices may be used to give such a representation. A picture of a book may give information on the title of the book, the author, the collection but it will not tell you the book is mine.
- *Arbitrary representation*: Arbitrary representation can be interpreted correctly only within a system of conventions. For example, in the case of linguistic representation, such a system of conventions is defined by rules of language usage; a representation such as a diagramme should be accompanied by the information necessary to interpret it (such as name axis or scale).
- *Static-dynamic representation*: Static-dynamic representation is considered static when it can be perceived in an identical form for a certain time. If the representation changes continuously over time, it is called a dynamic representation. A blinking icon is considered static while a movie or music will be characterised as dynamic. The classification is dependent on the size of the time window.

Based on these categories, Blattner and Glinert (1996) introduce three main groups of output:

- Linguistic, non-analogue and non-arbitrary representations include for example spoken language, handwritten text, written text, braille.
- Non-linguistic, analogue, and non-arbitrary representations are modalities such as movies, maps, images, non-sense sounds, graphs, diagrams.
- Non-linguistic and non-analogue representations group modalities such as windows or scroll bars.

2.7.2 Output devices

The visual, acoustic and haptic modalities are used by systems to communicate with users. We now introduce the different devices associated with each modality.

- Visual devices: Visual display by using a monitor is the most widely used means of communication via computer. Virtual reality, stereoscopic monitors, and immersive systems enhance spatial information by displaying data in 3D.

- Acoustic devices: Much research has been conducted with the goal of producing good quality synthetic speech, and there are several commercial products which achieve this goal. Non-speech sounds include beep sounds, auditory icons, and auditory display (visualisation of data through sound parameters (Kramer 1994)). Virtual reality systems or headphones can simulate spatial relations of sounds.
- Haptic devices: These devices are generally expensive (minimum $ 10,000 and more). Vibration generation seems to be an effective way of stimulating the tactile sense (Schomaker et al. 1995a). Work is underway to develop electrotactile stimulation. But most haptic devices do not act directly on the somatic sense of users (for example force feedback devices). DataGlove can also be equipped to send feedback to the user (Stone 1991). A simple 2D mouse can be transformed to produce force feedback and predict the user's next actions (Schomaker et al. 1995a).

2.7.3 Theoretical issues

2.7.3.1 Introduction to multimedia systems

Multimedia software combining text, graphics, pictures, video and audio is now widely available on the market on CD-ROMs and on the Web. Multimedia authoring tools enable developers to integrate and combine text with pictures, animations, video and audio clips. From a theoretical point of view, the development of multimedia interfaces addresses several issues (Maybury 1993; Dowell et al. 1995):

- content selection ("what to say"),,
- media allocation ("which medium to say it in"),
- media realisation ("how to say it in that medium") and
- media combination ("how to combine several media").

In this section, we focus on multimedia output and consider only limited user input such as hyperlink navigation or the classical graphical user interface using the mouse.

2.7.3.1.1 Solutions to the "media allocation" problem

Among these issues, the "media allocation" issue has received most attention, but without providing any universal solution. One approach is to map types of data onto types of medium (e.g. database entries onto tables). Another approach is to map properties of information (e.g. fixed versus varying) onto properties of media (e.g. persistent or not). Several suggestions on the salience of graphics versus text as a function of the types of information or the communicative act are provided in Maybury (1993).

2.7.3.1.2 Synchronisation and interaction issues

Complex and only partially solved problems regarding synchronisation and interactivity (even simple ones such as hypermedia or direct manipulation of buttons) occur in multimedia systems (Blakowski and Steinmetz 1996), which have consequences as to the functions a multimedia toolkit should be able to support (Bailey et al. 1998; Schomaker et al. 1995a):

- *Synchronised modalities*:
  The presentation freezes, and the user can interact. The different modalities have different spatial and temporal properties, but also share spatial and temporal properties to a large extent. Correct synchronisation enhances the interpretation and the comprehension of simultaneously processed modalities. Negative examples can be seen in dubbed ('synchronised') movies; somewhat different expectations are associated with cartoon movies, though advances in animation technology will no doubt change this in the near future. A typical example is the perception of speech, which is greatly enhanced when visual information is concurrently provided. Presentation of written material or spoken text may also be augmented by other modalities (e.g. video, cartoon character, deictic gestures, see Rist et al. 1997). An example of such a system is the WIP system developed by Rist and his colleagues (Rist et al. 1997). The WIP system presents material at the request of a user. Not only the presentation is displayed (text, illustration, video, graphs) but a 2D animated character, PPP Persona, helps the user to perceive the more relevant items in the material by pointing to specific objects/words. Two major issues in integrating audio and video output with animated characters must be considered:
  - Selection of material: Examples of display choices include text, speech, video, and graphics. These are not equally appropriate for all information types, and one should consider which modality is appropriate, based on the query of the user and on the interaction context.
  - Scheduling process: The temporal and spatial relationships for an animated multimodal presentation cannot be predicted in detail during the development phase of the system, and thus have to be computed on the fly: temporal constraints, such as when a deictic gesture occurs in the presentation, and spatial constraints, such as where a deictic gesture points to, are added to the system and are computed when required.
- *Asynchronised modalities*: The presentation continues at its own pace, and the user can interact when he wants to. In some situations, an asynchronous scheme is more appropriate than synchronisation. For example, multimodal interfaces should allow the user to interrupt a process, to cancel a request, to clarify a request at any time, including during processing (Bayer et al. 1995). Barge-in synchronisation would require the system to finish the execution of the current event first and then to accept the user's next request. On the other hand, an asynchronous scheme will interpret any new input event right away. IPP, a multimodal user interface on the web (Bayer et al. 1995), is an example of such an asynchronous system. Different applications of IPP can be executed any time as well. IPP accepts as input text, mouse-pointer deixis, and speech. The possible outputs are synthesised speech, text, graphs, and map displays. Which output modalities are chosen depends on the user's requests.
- Fine-grained temporal relationship: Tight / loosely coupled / synchronised at specific points in time.
- Conjunctions and disjunctions of the above items.

As an example, the toolkit described in Bailey et al. (1998) consists of two components, a declarative synchronisation definition language and a run-time presentation management system. The synchronisation definition language supports the specification of synchronous interaction, asynchronous interaction, fine-grained relationships, and combinations of each through the use of con-

junctive and disjunctive operators. The run-time presentation system uses a novel predictive logic to predict the future behaviour of a presentation. As the viewer makes decisions, the presentation is updated and new predictions are made.
One of the key problems of multimedia development is the specification of spatio-temporal composition and indexing for applications which use large numbers of objects, e.g. more than 10,000 for a 3D synthetic movie (Vazirgiannis et al. 1998).

2.7.3.1.3 Multimedia developer versus intelligent multimedia system

The problems mentioned above can be solved either before the execution (i.e. during the development), or during the execution. Thus, two families of multimedia systems can be distinguished.

1. In the first family, the multimedia design is made by a developer. The multimedia developer uses a multimedia authoring tool (or other generic development tools and a programming language) to specify the content of each media (which can be retrieved from existing databases). The above issues are addressed before execution.
2. In the second family, the multimedia design is made by the multimedia system itself. Some "intelligent" multimedia systems are able to generate on-line the content to be presented in each media thanks to decision making capabilities.

An example of a system combining both families is described in Denda et al. (1997), where speech synthesis is produced on-line along with pictures and maps. As described in the next section, current trends in multimedia standards also deal with this classification of multimedia systems.

2.7.3.2 Recommendations for the use of speech output in multimedia systems

This section elaborates a list of questions related to speech output that have to be answered by developers of multimedia systems, or managers faced with decisions about purchasing a multimedia system. These questions arise before buying a multimedia system (or multimedia system development tool), during system development, and during the execution of the system (for instance, whether the system is able to take on-line decisions during presentation). An obvious recommendation is to perform user studies, since no generic rules for the use of speech output in multimedia interfaces can be formulated at the present time. In this section, we provide questions and tentative answers with respect to experimental studies involving human–computer interaction. Readers interested in the corresponding guidelines should have a closer look at the experimental studies referred to in order to check whether they can be applied to their own needs.
We have used the following grouping for recommendations on the use of speech output in multimedia systems:

- Why should speech output be used in a multimedia system?
- When should speech output be used?
- When should the system use a combination of speech output and other media?
- When should the system prefer the use of speech output to other media?

- When should the system support equivalence of use between speech output and other media?

These recommendations will involve intrinsic features of the speech output modality: sequentiality, non-persistence, omni-directionality.

2.7.3.2.1 Why should speech output be used in a multimedia system?

Various claims about the use of several media, including audio and speech, can be found in psychology. For instance, Gibson (1966, 1979) argues that our senses are constructed to handle the very complex flow of information in natural environments, and that our senses are not constructed to handle simple stimuli. Gibson argues further that we are not passive receivers of information. Rather, our perceptual system is characterised by the pickup of information and by the integration of activities of the different senses. Although these arguments are directed at the experimental study of human perception they are, according to Marmolin (1991), also relevant to multimedia computing: information systems that need to support human information processing should make full use of human perceptual and cognitive capabilities, and should represent natural information flows to users and offer support for processing natural information flows. Experimental evaluations of voice interaction between two subjects have shown the effectiveness of the speech channel (both input and output) over other channels such as writing, typing, video, regarding the average time it took the subject to solve a problem (Ochsman and Chapanis 1974).
It has also been reported that the incorrect use of multimedia can easily result in negative cognitive side effects such as over-stimulation, cognitive overload, distraction, fatigue (Heller 1990). A number of studies have also demonstrated the attention grabbing, sometimes disrupting effect of audio, background speech and noise (Hapeshi and Jones 1992; Taylor 1989).
The constraint of using speech output is also related to the existing material chunks that designers want to use in the presentation: video clips, audio clips, images, etc. Although not yet available in existing authoring tools, techniques are explored to enable automatic processing of audio and video for information retrieval. Artificial intelligence techniques are used to analyse, index, extract and retrieve audio or video files containing speech signals (Maybury 1997). Content-based audio retrieval is also tackled for instance in a user-extensible sound classification and retrieval system that computes both acoustic and perceptual properties to enable content-based audio clip access (Blum et al. 1997). Systems using speech and language processing for video access may use techniques for video mail retrieval using spoken language indexing (Jones et al. 1997), or large vocabulary, continuous speaker independent broadcast news transcription (Hauptmann and Witbrock 1997).
In multimedia (as in any speech-only application), the quality of speech output is central. The quality of speech synthesis or concatenation of isolated words is generally not as good as recorded speech (Bunt 1989). Researchers in multimedia have found that synthetic speech may hinder verbal learning (Hapeshi and Jones 1992). On the other hand, high quality but repetitive recorded spoken messages can bore users (Wang et al. 1993).

The use of speech output can also be imposed by constraints linked to the application (eyes busy during the task, existing application of high graphical complexity, application over mobile telephone), the users (pre-school children, blind people) or the environment (no place for visual display or not enough light).

2.7.3.2.2 When should speech output be used?

Once it has been decided to include speech output in a system, designers have to decide when to use it during the multimedia presentation. The answer to this question depends on several features (Arens et al. 1993): the characteristics of the media used, the nature of the information to be conveyed, the communicative goal, the preferences/capabilities of the current user and the environmental conditions. The criteria used for answering this question is also of importance: the recommended media may not be the same when considering either the number of errors made by subjects, or the speed at which these users accomplish a task.
Managers who want to buy a multimedia system should ask multimedia companies the following questions:

- Is there any control over the speech output (especially regarding the content of the information)?
- What were the rules applied by the designers regarding which type of information is conveyed by which media?
- Are these rules extracted from experimental studies made with the multimedia system, by a review of the literature, or by informal procedures?

Several studies compare the use of graphics versus language regarding the content of information (no distinction is made as to whether speech output or text display is used). Concrete information such as visual properties (shape, colour) and hierarchical structures should be transmitted through graphics (André and Rist 1993; Bos et al. 1994). But graphics cannot describe objects which are not visible or are not displayed due to lack of space. Spatial information such as the position, the orientation, the composition of objects, and the actions and events using movements should be transmitted using graphics (André and Rist 1993). However, in the case of assembly tasks, information was perceived faster with graphics but subjects made fewer mistakes when they were provided with text (Bieger and Glock 1986). Sequential temporal information between states, events and actions can be represented by a sequence of frames. They are better represented by text when they describe events overlapping in time, or specifications such as “often”, “periodically”, “in the future” (André and Rist 1993), or future actions to be done by the user (Cohen 1992). Language seems more appropriate for abstract information (Bos et al. 1994) such as negations, quantifiers, or semantic relationships such as action/result, problem/solution. The selection of language content and graphics content also depends on the communicative goal (André and Rist 1993; Huls et al. 1994; Cohen 1992): attract attention, compare, elaborate, enable, elucidate, label, motivate, evidence, background, summarise. For instance, language can be used to direct the user’s attention to special aspects of the graphics.

These claims pertain to the comparison of text and graphics. But language can be conveyed either as text on a visual display or as speech output. Visual display of text can be read several times and is perceived faster than speech (Huls et al. 1994). Moreover, spoken messages in natural language should be less verbose than textual messages. Messages in telegraphic style text can be more easily understood than spoken telegraphic style messages (Bunt 1989).
Text-based information presentation systems for people who are not domain experts can build on rich existing knowledge of language, while a graphics-based presentation system must explain everything from scratch, except the meaning of a few hundred icons and a few general syntactic rules which should be easily understood by the "man in the street" (Reiter 1997) without significant training.
Several groups of claims regarding the use of speech output are described in Bernsen (1996):

- claims recommending the use of speech output (i.e. "speech output reduces visual clutter in graphical displays")
- claims positively comparing speech output with other modalities (i.e. "speech output may be preferable to static text for persuasive information")
- conditional claims on the use of speech (i.e. "speech output is attention catching and thus may require headphones in some work environments")
- recommendations against the use of speech output (i.e. "avoid speech output when spatial reference to the information source is important")
- claims negatively comparing speech output to other modalities (i.e. "speech output may lock people out of further interaction for its duration, whereas static visual displays can be sampled when convenient")

These claims are evaluated against properties derived from more general principles, such as modality theory. For example the claim "speech output may be preferable to static text for setting a mood" is justified by the following property: "discourse output modalities have strong rhetorical potential". A detailed list of claims can be found in Bernsen (1996).

2.7.3.2.3 When should the system use a combination of speech output and some other media?

People are indeed able to direct attention to groups of stimuli sharing sensory characteristics (Bearne et al. 1994). Users may be able to link several pieces of information by sensory characteristics such as prosody or the speaker's voice (male voice for one window, female voice for another window). But for efficient perception, the user should not have to listen to two passages of speech simultaneously, watch two videos simultaneously, or watch one video while listening to speech on another subject.
In Wang et al. (1993), it was observed that the redundancy between displayed text and vocal messages enabled faster learning of a graphical interface. But when users were familiar with the interface, they were annoyed by the redundancy between speech and text, especially since the content of the recorded spoken messages could not be modified on-line by the system.
The combined use of speech and text output is investigated in Huls and Bos (1995). Subjects were asked to move files into folders when the graphical rep-

resentation of the files came with textual and/or spoken messages. Some of the subjects in the text-only output condition did not use the descriptions provided by the system. The time taken to perform the task was shortest for text-only output, followed at a distance by the text-and-speech condition, the nonlinguistic output condition, and finally the speech-only output condition. The smallest number of errors was found in the speech-only condition followed by the text-only condition, the text-and-speech condition, and finally the nonlinguistic output condition.
The level of congruence is the degree to which different media are used redundantly to express the same concepts. With regard to the congruent use of synchronised video and audio, Hapeshi and Jones (1992) remark that the presence of moving images can serve to enhance comprehension and learning of spoken material. But they also describe studies showing that the presence of an incongruent video presentation significantly reduces recognition memory of audio material, however showing a congruent visual map results in better recall, particularly if the narrative structure is relatively simple. In the case of a monologue or a dialogue, visual display can facilitate processing of the auditory message if the speaker's face can be seen, because facial expressions, particularly lip movements enhance speech intelligibility. Visual display of text can also enhance speech intelligibility, e.g. to recognise the lyrics of a song.
A tool integrating guidelines with the development of a multimedia interface using an expert system is described in Faraday and Sutcliffe (1997). Their tool is based on experimental studies of attention and recall of expository multimedia presentation. The results of their studies were used to form a set of guidelines for attentional design which aim at predicting what would be attended to in a presentation and at flagging any potential design problem. What follows is a subset of their guidelines related to the use of speech output when combined with other media (for more details see Faraday and Sutcliffe 1996):

- Multiple concurrent strands of speech or sound will interfere with each other and distract focus. Speech and visual information can be focused upon concurrently, but no more than one language strand should be presented at once.
- Show (reveal) objects and labels when cued in the speech track. Cueing labels within the speech track will produce a shift of attention to the object and its label.
- Allow reading time after cueing a text. Avoid reveals or animation for the duration of speech segment which cues a label. If the label is complex, reading speed will be similar to that for speech track to pronounce it.

Existing development tools enable the specification of some temporal and spatial relations between media objects. Current research also explores the specification of more complex relations. For instance Nsync (Bailey et al. 1998), a multimedia synchronisation toolkit consists of two components, a declarative synchronisation definition language and a run-time presentation management system. The synchronisation definition language supports the specification of synchronous interaction, asynchronous interaction, fine-grained relationships, and combinations of each through the use of conjunctive and disjunctive operators. Fine-grained temporal relationships include: tight / loosely coupled relationships / synchronised at specific points in time within a media object.

The run-time presentation system uses a novel predictive logic to predict the future behaviour of a presentation. As the user makes decisions, the presentation is updated and new predictions are made. The toolkit enables the specification of different types of control the user may have on the multimedia output: ability to skip ahead, skip back, or adjust the playback rate during the presentation. When choosing between several multimedia development tools, a manager or a developer should study the types of interaction, temporal and spatial relations that can be specified. More precisely, questions related to the spoken output capacities of these development tools need to be answered:

- Does it allow the recording and playing of spoken messages?
- Can these spoken messages be dynamically combined?
- Does it enable the production of speech output from a textual representation?
- Can synchronisation cues be incorporated in order to enable fine-grained synchronisation between speech and other media?

2.7.3.2.4 When should the system prefer the use of speech output to other media?

This question does not focus on the combination of speech output with some other media. Rather, specialisation means that in similar parts of the multimedia presentation speech output will be involved.
These parts in the multimedia presentation can be selected in different ways:

- Parts similar with respect to temporal features: for instance when the multimedia presentation is made of several sequences and speech output is involved at the end of each of these sequences.
- Parts similar with respect to content: for instance, all help commands involve speech output.
- Parts similar with respect to the user: for instance, at the beginning of the presentation, the user is asked to give information regarding his preferences or computer familiarity and, depending on his answer, speech output is used or not used during the presentation.
- Parts similar with respect to environmental conditions: similarly, if the user selects a noisy environmental condition, speech output should be avoided during the presentation.

Three types of specialisation relations can link a case C_i (i.e. warning messages) to the use of speech output:

- absolute relation between the use of speech output and the case C_i,
- media-relative specialisation of speech output in the case C_i: that means that speech output is not used in other cases, but some other media can be involved in the case C_i (i.e. a warning is also displayed graphically),
- message-relative specialisation: the C_i case does not use other media than speech output but speech output is also used in other cases.

The main benefit of specialisation is that it is easier for the user to interpret the speech output message since it is always used for the same purpose.

2.7.3.2.5 When should the system enable equivalence of use between speech output and another media? What are the criteria?

In existing media, the cooperation between media is limited and if the user switches off the loudspeaker, there may be some information he will not be able to get out of the multimedia presentation. It is of importance to allow the switching of output from speech to another media when the user decides to do this, either due to his preferences or due to some unexpected environmental features such as noise. Output switching enables the user to have control over the multimedia presentation. As stated in Bearne et al. (1994), a multimedia presentation "should provide individual users with the flexibility to switch off certain forms of output where they are not essential to the task". Of course, presentations should always be designed such that all users will be able to attend to all essential information. Output switching may also enable the adaptation to environmental changes.

2.7.4 Summary of recommendations

This section summarises the recommendations regarding the use of speech output in multimedia systems. The reader is recommended to have a further look at the references for details and underlying assumptions of the claims.

2.7.4.1 Recommendations regarding applications

- Use speech output when the user's eyes may be occupied during the application. Use speech output commands in following procedures (using a video recorder) where limb and visual activity is required.(Bernsen 1996)
- Use speech output when mobility is needed. (Bernsen 1996)
- Use speech output if the message needs to be displayed to several people simultaneously. (Bernsen 1996)
- Use speech output where the graphical display is overloaded (i.e. aviation control). (Bernsen 1996)
- Avoid using speech output for spatial manipulation applications. (Bernsen 1996)
- Avoid using speech output if privacy protection is needed during the application. (Bernsen 1996)
- Use speech output alarm when immediate response is required. (Bernsen 1996)

2.7.4.2 Intrinsic properties of speech output

- People can listen to speech faster (up to 700 words/min) than they can read (up to 200 words/min) (Levy-Schoen 1969).
- Avoid repeating spoken messages (especially recorded messages) (Wang et al. 1993).
- Spoken messages should not be "too long". Natural language spoken messages should be shorter than visually displayed textual messages. (Huls et al. 1994)
- Synthetic speech should be of "good" quality (otherwise, recorded spoken messages should be used). Avoid telegraphic-style spoken messages (telegraphic style messages are more acceptable for visually displayed text). (Bunt 1989; Hapeshi and Jones 1992)
- Avoid multiple strands of speech or sound, which will interfere with each other and distract focus. (Faraday and Sutcliffe 1997)

2.7.4.3 Recommendations regarding the environment

- Use speech output if the environment has low-acoustics, no other people can be disturbed (otherwise, consider using headphones), or conditions are not good enough for visual display (darkness). (Bernsen 1996)

2.7.4.4 Recommendations regarding the user

- Use speech output for people who have difficulties with standard computer outputs (computer illiteracy, visual deficiency, pre-school children) (Bernsen 1996).

2.7.4.5 Recommendations regarding content

- Use speech output for low complexity information, short lists (versus text for long lists) since speech output implies severe cognitive processing limitations with respect to the amount of information that can be attended to in real time and remembered (Bernsen 1996).
- Use graphics in preference to speech output or text for presenting concrete information (shape, colour, texture) (André and Rist 1993).
- In assembly instructions spatial information may be perceived faster if pictures are used; on the other hand, subjects confronted with textual presentations make fewer mistakes when carrying out instructions (Bieger and Glock 1986).
- Textual presentations should be preferred for time specifications such as "mostly", "periodically", "in the future", "overlap in time" (André and Rist 1993).
- Covariant information (cause/effect, action/result, problem/solution, condition, concession) should be presented with text (or a combination of graphics and text) (André and Rist 1993).
- Quantifiers and negation are more easily described by text than by graphics (André and Rist 1993).

2.7.4.6 Recommendations regarding communicative goals

- Use speech output for warnings since speech output is attention grabbing (better than static text) (Bernsen 1996).
- Voice warning is more explicit and more intuitively appropriate than audio warning (Taylor 1989).
- Speech output may be preferable to static text for setting a mood, persuasive information (Bernsen 1996).

2.7.4.7 Recommendations regarding interaction

- Enable the user to have control over the speech output. Add facilities for reviewing speech output messages for enhancement of long-term retention (Bernsen 1996).
- Provide facilities for interrupting and resuming spoken output at the beginning of a semantic unit when resuming synthesis which has been interrupted (Bearne et al. 1994).
- Speech output may lock people out of an interaction for its duration while static visual displays can be sampled when convenient (Bernsen 1996).

2.7.4.8 Recommendations regarding the combination of speech output and other media

- Avoid simultaneous video and speech output on different subjects (Bearne et al. 1994).

- Speech output can be used to elaborate, summarise, elucidate graphic messages (André and Rist 1993).
- Use speech output features such as the speaker's voice to facilitate the link between the spoken message and other media (i.e. male voice for one window, female voice for another) (Bearne et al. 1994).
- Pay attention to complex synchronisation relationships between speech output and other media. Show (reveal) objects and labels when cued in the speech track. Cueing labels within the speech track will produce a shift of attention to the object and its label. Allow reading time after cueing a text. Avoid reveals or animation for the duration of the speech segment which cues a label. If the label is complex, reading speed will be similar to that for the speech track to pronounce it (Faraday and Sutcliffe 1997).
- Development tools should enable the specification of synchronous interaction, asynchronous interaction, fine-grained temporal and spatial relationships (tight and loosely coupled, synchronised at specific points in time) and combinations of each through the use of conjunctive and disjunctive operators. (Bailey et al. 1998; Vazirgiannis et al. 1998).
- Redundancy between recorded speech messages and textual output may enable faster learning of the interface (Wang et al. 1993).
- Redundancy between recorded speech messages and textual output may annoy expert users (Wang et al. 1993).
- Redundancy between recorded speech messages and text may direct the user's attention to the textual message (Huls and Bos 1995).
- Redundancy between recorded speech messages and text may enable the user to achieve a task faster than with speech alone but slower than with text alone (Huls and Bos 1995).
- Redundancy between recorded speech messages and text may drive the user to achieve a task with more errors than with text only or speech only (Huls and Bos 1995).
- Information provided by speech output should be planned to be presented in another modality if necessary (Bearne et al. 1994).

2.8 Technology of multimodal system components

This section reviews technical details of components necessary to implement multimodal systems as described in previous sections, including reviews of recognition algorithms, and the performance of state-of-the-art systems, where available. First, 2.8.1 describes techniques and algorithms related to processing the visual input channel, in particular face detection, face recognition, tracking of faces in a sequence of images, and locating and tracking of other facial features (e.g. lips). Then, 2.8.2 reviews technical concepts and techniques for the creation of 3D representations of faces and facial features, required for talking heads and synthetic conversational agents as described in Sections 2.8.3 and 2.8.4. Section 2.8.5 discusses the state-of-the-art recognition technology for on-line handwriting, and Section 2.8.6 for gestures.

2.8.1 Techniques related to face recognition systems

This section describes techniques related to face recognition systems, addressing the three fundamental problems of such systems: first, the detection of faces within a given scene; second, the identification (or recognition) of faces within a set of known faces; third, the tracking of faces within the moving scene.

2.8.1.1 Detection of faces

We review two main techniques for detecting faces in an input image:

- *Holistic detection of faces* (e.g. Yang and Waibel 1997; Stiefelhagen et al. 1997b). Colour has long been used for object recognition, and recently has been successfully applied to locating (and tracking) faces, based on the observation that the colour distribution of skin-colour is clustered in a small area of the chromatic colour space. Figure 2.16 shows how a face can be detected based on colour information only – provided the background does not contain skin colours. [5] Although colour information is an efficient tool for identifying facial areas, it depends on lighting and camera, and therefore is useful only in combination with other channels. If a sequence of images is available, motion analysis offers a quick way of locating moving objects, such as heads and hands. Moving objects are located by analysing the difference of consecutive frames. Spurious responses can be eliminated by combining motion with colour analysis: moving skin coloured objects are most likely either faces or hands.

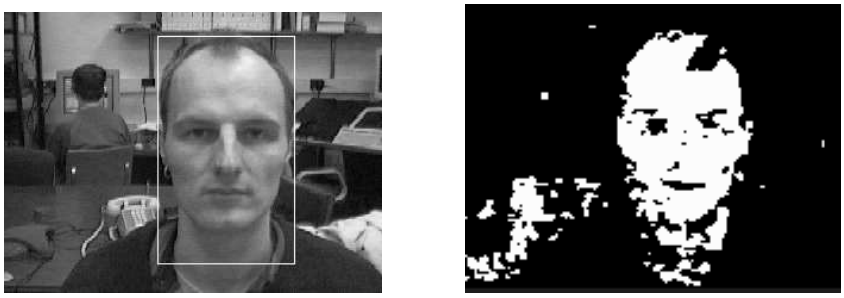

Figure 2.16: Application of the colour model to a sample input image. The face is marked in the input image.

- *Detection of faces based on the detection of facial features.* Shape and texture information of facial features can be extracted using standard image processing techniques (filtering and thresholding). Yuille et al. (1989) present a face detection algorithm based on deformable templates for eyes. Graf et al. (1997) describe an "n-gram" search technique that combines information from various channels (including colour and motion information, see above), such as grouping shapes bottom up, from individual shapes to whole faces. The search is kept efficient by using a hierarchical search and eliminating groups with low scores in each step (beam search).

In order to handle a wide range of conditions (e.g. lighting, camera, head orientation), hybrid approaches combine information of several analysis channels, including shape, texture, colour, and motion. Such hybrid face detection techniques have been reported to achieve almost perfect performance on large datasets with many speakers (Graf et al. 1997). Also, adaptation techniques have been shown to make colour-based face detection more robust against changes in the environment (Yang and Waibel 1997).

[5] The figure shows a black-and-white approximation. - Ed.

2.8.1.2 Face recognition

Two main approaches have been used by researchers to recognise faces: template-based recognition and feature-based recognition.

2.8.1.2.1 Template-based recognition.

Template-based recognition represents images as an array of pixel values. Subimages can be masks of the eyes, nose, or mouth. The pixel value can be intensity values or may have been pre-processed by gradient or Laplacian filters to achieve scale, translation, and rotation independency. The recognition is performed by computing a normalised cross-correlation for each template, and finding the highest cumulative score.

- Principal Component Analysis: The simplest version of template matching is principal component analysis (PCA). A test image is classified based on its (Euclidean) distance to templates generated from the faces in the training set (database). The Kurhunen-Loeve procedure (Kirby and Sirovich 1990) and eigenfaces (Turk and Pentland 1991) are based on this simple template matching method. Eigenfaces correspond to characteristic feature images and can be viewed as the principal components of a test image with respect to characteristic features obtained from the database of faces. This technique has been applied to the recognition of lip shapes (Bregler and Konig 1994). The Table 2.4 summarises the recognition rate results in word error.
- Geometric templates: Another approach builds geometric templates of specific facial features (such as eyes and lips) to describe and then recognise faces (Yuille et al. 1989). These templates are constructed based on a priori knowledge about the feature shapes. Templates are parameterised curves that can deform during model fitting. The curves follow the outline of the facial features, and their final shapes can be used to verify if the observed object is an eye, lip, or face. An appropriate distance metric has to be defined, for example a potential energy function. Minimising the potential energy is equivalent to forcing the templates toward salient features (valleys, edges, peaks and intensity). A problem with this technique is its relative dependency on position and lighting.
- Deformable templates: Deformable templates are used to model lip shapes and recognise faces (Chandramohan and Silsbee 1996; Yuille 1991; Vogt 1996; Silsbee 1994; Luettin et al. 1996). These templates are constructed based on a priori knowledge about the feature shapes as parameterised curves that can deform during model fitting. The curves follow the outline of the facial features and their final shapes can be used to recognise a particular lip shape or face. When multiple templates are used in the recognition process the results of correct recognition increases, for example, Chandramohan and Silsbee (1996) report 16% classification accuracy with one template, and 33% accuracy with six templates.
- Optical flow: Optical flow techniques allow to detect motion rather than facial feature displacements (Mase 1991; Essa et al. 1994; DeCarlo and Metaxas 1996; Ezzat and Poggio 1997). It works at the pixel level and computes the difference in image intensity between two consecutive frames. The computation is done pixel per pixel. This technique may be used to extract muscle contraction. Windows are placed around muscle locations. Velocity of each muscle contraction is computed.

- Neural networks: Another variant of template matching are neural-network based approaches to face recognition, for example applying Kohonen self-organising maps (Allison et al. 1992).

In order to alleviate dependency on the view, multiple views (also called virtual views) can be used (Beymer 1995; Beymer and Poggio 1995). Two approaches are current: either collect multiple views, or generate virtual views from one template.

2.8.1.2.2 Feature-based recognition.

Kanade (1973) first introduced the description of faces by using features. Since then, many others have further developed the feature-based approach to face recognition (e.g. Brunelli and Poggio 1993b; Craw 1992). First, facial features are located using on geometric measurements which are represented as feature vectors. Geometric features correspond to parameters such as angles, distances and curvatures of the eyes, nose, and mouth. Anthropometric features and profiles have also been used. Parameters can be extracted by first reducing the information from the video image: a binary image is generated using a threshold value (Brooke and Petajan 1986); the chroma-key technique is used to detach the lips from the image background (Lallouache 1991); reflective markers are placed onto and around the lip area (Cosi and Magno-Caldognetto 1996). Next, the face is identified by comparing its features with features of faces stored in a database. Before features can be compared, scale normalisation ensures that face images are of the same scale. Scale normalisation can be achieved by locating both eyes in the image and by applying rotation, translation and scaling to align them with reference faces. Extraction of facial features is a delicate and difficult task. To enhance feature extraction, model-based approaches exploit a priori the fact that faces have the same overall configuration: two eyes above a central nose and a mouth centered below the nose.

- Landmarks: Craw (1992) uses a generic set of landmarks that create a triangular mask and serve to standardise every test image. The landmarks identified in the current face are transformed to adapt as much as possible to the reference mask. Texture mapping of the face is conserved and non-linear warping can be used to adapt it to the deformed face. The current mask and the reference mask are compared using principal component analysis (PCA). An advantage of this approach is its invariance to lighting conditions, small changes in view angle, and to some facial expressions.
- Anthropometric data: Brunelli and Poggio (1993a) use anthropometric data and knowledge to extract facial features automatically. The eye-to-eye axis and the distance between the eyes are detected and used to achieve view independency. Integral projections (already used by Kanade (1973)) extract horizontal and vertical edge directions. Horizontal gradients allow the detection of the left and right boundaries of the face and the nose. Vertical gradients allow the detection of the top of the face, the eyes, the base of the nose, the mouth, and the chin. A Bayesian classifier can be used for face recognition based on these features.
- Snakes: Terzopoulos and Waters (Terzopoulos and Waters 1993; Parke and Waters 1996) propose a related approach. *Snakes* (or *active contours*) are first located on the face. Contours are tracked by applying an image force field that

is computed from the gradient of the intensity image. Muscle contraction is estimated from contour deformations. The effect of visual information on the recognition of audio signals is around 7% (Dalton et al. 1996).

2.8.1.3 Recommendations on face recognition systems

The following recommendations can be given for face recognition systems.

- Template-based systems give better recognition accuracy rates than feature-based systems (Brunelli and Poggio 1993b). They avoid the hard task of feature detection and are easier to implement.
- Eigenlip and eigenface methods suffer from a lack of precision in detecting particular points (Bregler et al. 1997).
- Extracting lip and facial motion with optical flow techniques offers an advantage as they easily detect word boundaries (Mase 1991). Lips mark a stop (zero motion) during fluent speech. Word boundaries are found by looking for zero velocity. Word recognition is done by matching computed results with stored words in a small vocabulary set.
- Feature-based face recognition can obtain real-time performance more easily, since a low dimensional feature vector is used for classification, but its main disadvantage is that feature extraction is difficult and some feature characteristics can easily be missed (Brunelli and Poggio 1993b). Recognition accuracies of 96% have been achieved (Harmon et al. 1981).
- Snakes work well with high contrast features. They are well suited to the recognition of outer lip contours but not for inner lip contours (Bregler et al. 1997).

When acquiring a video image one has to be careful that it is composed of two fields. A video image is displayed in two stages: the odd lines of an image are displayed separately from the even lines. For SECAM or PAL systems the frame rate is 25 interleaved frames per second, i.e. one frame every 40ms. It takes 20ms to display the odd lines and 20ms for the even lines. When capturing the screen image the camera should take this fact into consideration. The camera should record at 50Hz (or 60Hz), depending on the frame repetition rate of the video.

2.8.1.4 Face tracking

When a face needs to be tracked in a sequence of images, it should be recognised in the first image of the sequence and then be tracked in the successive ones. Yang and Waibel (1997) present a real-time face tracker based on colour, geometric and motion information. To eliminate dependency of colour-based face detection to change in viewing environment (lighting, camera type, clothing, background), the skin-colour model is dynamically adapted using a temporal filter. To increase tracking speed, the search window is adapted by using motion predictions. The motion of a face can be predicted in real-time using the current position and velocity. Motion modelling techniques such as Kalman filters tend to be computationally expensive and are currently not feasible in real-time. Finally, in order to track people moving freely within a large area, a camera model predicts camera motion (panning, tilting, and zooming), which is

necessary in order to keep up with varying distances to the tracked face. Tracking speeds of 15, 20, and 30 frames per second are reported at 0.5, 1.0, and 2.0 meters distance from the camera, respectively (using a HP 9000 workstation and a Canon VC-C1 camera).

2.8.1.5 Approaches to locating and tracking other facial features

Model-based approaches can achieve higher accuracies. Stiefelhagen et al. (1997b) present a non-intrusive model-based tracking system for gaze (and other facial features, including eyes and nostrils). The face and facial features are located by top-down search based on a statistical colour model and knowledge of the geometric configuration of a face. First, the face is located as the largest connected region of skin-coloured pixels. Then, eyes, nostrils, and lip corners are located by searching in appropriate subregions of the located face, exploiting geometric relations between the different facial features. The 3D pose of a user's head is estimated and tracked based on six facial feature points (both eyes, lip corners, and nostrils), using a full perspective model. A frame rate of 15+ frames per second using a HP 9000 workstation and a Canon VC-C1 camera is reported. Figure 2.17 shows tracking points for eyes, nostrils, and lip corners, and the respective subimages that were searched to identify these points within the face.

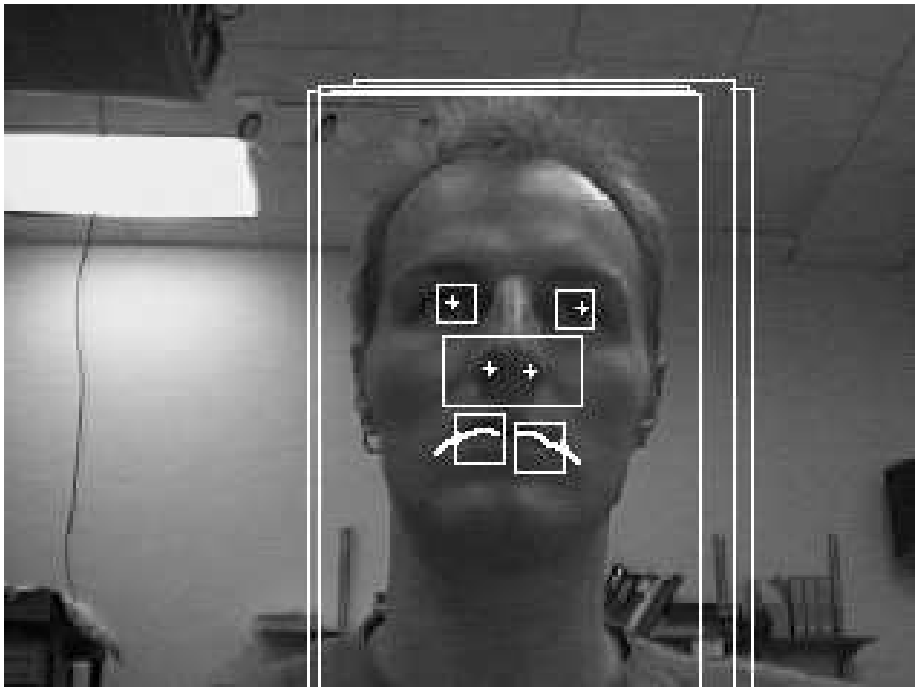

Figure 2.17: Tracking of eyes, nostrils, and lips corners

2.8.1.6 Recommendations on recognising facial expressions

- Representations: Maintain several different representations of each area of interest at different levels of detail (enumerated with increasing complexity: position, bounding box, colour information, and binary/greyscale pixel map), and analyse them using simple representations first, backing off to more complex representations only in cases of ambiguity.
- Necessary spatial and gray level resolutions: The lower bound on spatial and gray scale resolution for human face detection and identification systems is 32x32x4 bpp, and can be considered as the lower bound for automatic face recognition systems as well.
- Robustness of face recognition under varying lighting and camera conditions can be achieved with algorithms that combine multiple cues (colour, motion, and shapes/textures), and with adaptation techniques.

2.8.2 Synthesis module

This section presents the different steps involved in the creation of a 3D model. Animation and control techniques are presented in Sections 2.8.3.1 and 2.8.3.3. All these sections will concentrate particularly on facial models and animation. Section 2.8.3.5 will discuss the problem of lip shape computation during speech. Finally, Section 2.8.4 introduces behaviour patterns and functions necessary to create a conversational agent. More detailed information can be found in Parke and Waters (1996); Foley et al. (1990); Prevost and Pelachaud (to appear).
The steps involved in the creation of a 3D model are discussed below: model creation, geometric representation, adding of features, rendering.

Model creation

Different steps are required to create a facial model: acquisition of the model, geometric representation, colour determination. Various techniques are available to generate a model, and more specifically a facial model. One can use a modeler, one can acquire data from different technologies (digitiser, sensor, laser scans), and one can build a canonical face or use anthropometric data.

Geometric representation

Most facial models use a polygonal representation. This is is a model which consists of a mesh of connected polygons defined by a set of points and a topology linking these points. To obtain smooth surfaces splines are often used. They are defined by a set of control points and weights. A spline surface is given by an array of control points. Each row and column of the array has the same number of points. To overcome the problem of detailed representations, hierarchical B-splines are used (Forsey and Bartels 1990). Local refinement is possible by overlaying a detailed surface. Points are only added to the specific area and not to each row and column of the array structure, with special attention to the boundaries of the overlaid regions. This process can be repeated interactively to obtain a multi-level representation of the model (Wang 1993).

Adding of features

Adding the eyes, ears, back of the head, neck, teeth, tongue, wrinkles and hair greatly enhances the aesthetics, naturalness and realism of the facial model.

But there is a trade-off in modelling between the quantity of data which can be handled the quality of the representation.

Rendering

The rendering process requires different steps: choosing colours, lights, and shading effects. Shading an object requires selecting the number of lights and the type of lights illuminating the scene. Shadow is important in a scene because it helps to create the 3D illusion and can produce dramatic effects. Lighting computation can be extremely complex. The simplest illumination models are presented in Foley et al. (1990); Parke and Waters (1996). The picture of a face can be texture mapped onto the facial model. It gives good results for the eyes and is a simple way of modelling the teeth and the inside of the mouth. The matching of the facial features of the texture and the model requires special attention. Other techniques producing high quality images can be used: simulating the texture of the skin (Ishii et al. 1993), raytracing and radiosity. Even though very realistic results can be obtained with these last three techniques, they cannot be considered for our purpose since their computation time is still very high and they are therefore not suitable for interactive programs with real-time requirements, where at least 15 frames per second are necessary.

2.8.3 Facial models

In this section we first introduce the different types of facial model and then the methods of controlling them.

2.8.3.1 Face modelling

In early systems, modelling was done by digitising a face (or part of the face) with different expressions. Each model was stored in a library. The animation was obtained by interpolating between two expressions. It is a very simple but extremely time consuming method: one has to create a new plaster model for each new expression and then digitise it. Even though the technique is rudimentary, very expressive animations can be obtained (Emmett 1985; Kleiser-Walczak 1988, 1989). The following presents different animation control techniques.

- *Parametric model*: A facial model is created and animated through a set of parameters. Generally, parameters can be divided into two groups: conformation and expression parameters. The former refer to parameters acting on the facial topology (including position and size of the nose and eyes, global size of the head). The latter specify facial expressions such as brow action, mouth movement, and blink. The separation between these two groups implies the independence of the facial model and of the facial expression. The animation is obtained by changing the parameter values and by interpolating between the key frames (Parke 1972; Hill et al. 1988; Pearce et al. 1986; Cohen and Massaro 1993; Nahas et al. 1988; Lewis and Parke 1987; Guiard-Marigny et al. 1994).
- *Physically-based*: Skin properties and muscle actions are simulated using an elastic spring mesh and forces (Platt 1985; Reeves 1990; Viaud and Yahia 1992; Pieper 1991; Waters 1987; Lee et al. 1995; Takeuchi and Franks 1992).

- *Structural model*: The face is structured as a hierarchy of regions (forehead, brow, cheek, node, lip) and subregions (upper lip, lower lip, left lip corner, right lip corner).(Platt and Badler 1981; Platt 1985) Each region corresponds to one muscle or a group of related muscles. These regions can, under the action of a muscle, either contract or be affected by the propagation of movement from adjacent regions. A region is defined by a special point (the point of insertion of the muscle), and its connection information (to which regions it is connected). Connection information is necessary for computing the movement propagation. The muscle is defined by three or five segments that follow the bone structure of the face. This model is well adapted to the Facial Action Coding System (FACS) to encode its motions (Ekman and Friesen 1978) (see Section 2.9.5.1). Action Units (AUs) can be defined to act on large regions (lip pucker (AU18)) or on subregions (inner brow raiser (AU1)). An example of an AU coding is given below. AU4 corresponds to the action of frowning, and involves both the right and the left brows as well as the regions between them. During frowning the brows become closer and are lowered. Therefore there are two types of action: central (the left brow moves towards the central position, that is, it moves towards its right, and, symmetrically the right brow moves towards its left) and down. Each brow is decomposed into three subregions: lateral, medial and central.

```
MACRO AU4
BETWEEN-ABOVE-BROW DOWN 0.1
BETWEEN-BROW CENTRAL 0.35
RIGHT-BROW-CENTRAL DOWN 0.35
RIGHT-BROW-CENTRAL CENTRAL 0.5
RIGHT-BROW-MEDIAL CENTRAL 0.5
RIGHT-BROW-LATERAL CENTRAL 0.5
LEFT-BROW-CENTRAL DOWN 0.35
LEFT-BROW-CENTRAL CENTRAL 0.5
LEFT-BROW-MEDIAL CENTRAL 0.5
LEFT-BROW-LATERAL CENTRAL 0.5
```

- *Muscle-based model*: This method integrates anatomical features (e.g. skull, skin, muscle) and properties of the face (elasticity of the skin and muscle contraction). The spring-mass model simulates skin and muscle behaviour (Waters 1987; Pelachaud et al. 1993; Waite 1989; Patel and Willis 1991). Each muscle is characterised by a vector that represents a direction, a magnitude, and a zone of influence. Three types of muscle can be modelled: linear, sheet, and sphincter. Two end points define linear muscles: the point of insertion in the skin (mobile point) and the attachment point on the skull (fixed point). Linear muscles pull in one direction. Most facial muscles are linear (e.g. zygomatic, risorius). The direction of muscle movement is towards the point of muscle attachment. The magnitude of the force is zero at this point and increases to a maximum at the other point of insertion into the skin. The contraction of a muscle acts in the zone of influence associated with the muscle. The sheet muscle is a flat broad muscle (e.g. frontalis). It is represented by a set of parallel vectors within a region. The sphincter muscle (e.g. orbicularis) squeezes towards a center. Its zone of influence is assimilated to an ellipse. Dynamic formulation of the spring system is used (Terzopoulos and Waters 1990, 1993). Different forces have to

be considered:

- Friction: The movement of points is dampened by the kinetic energy dissipation due to friction.
- Volume preservation: The skin is an incompressible material. This property can be incorporated into the model by adding a force on each point of the skin layer. The amplitude and direction of the force depends on the deformation force applied to the node. Such a force allows the skin to bulge or crease.
- Penetration constraint: The skin does not penetrate the skull but slides over it. This penetration constraint is simulated by a force whose direction is the same as the normal skull (Lee et al. 1995). When the force applied on a node causes the node to penetrate the skull, the non-penetration force counteracts it.

- *Facial tissue*: The skin consists of different layers whose density and thickness vary according to their function. It is divided from the outermost to the innermost layers into the epidermis, the dermis, and the subcutaneous layers. A three layer deformable lattice structure can simulate the various skin layers (Terzopoulos and Waters 1990, 1993; Lee et al. 1995). The lattice is made of points connected by springs. The lowest layer corresponds to the bone structure. A muscle layer connects the bone structure to the fascia surface, which is also connected by springs (representing the dermis layer) to the epidermis surface (see Figure 2.18). The stiffness value of the springs on each layer depends on its biomechanical properties. For each layer, the springs have different stress–strain relationships.
- *Finite element method*: Biomechanical properties of skin have been widely studied (Gou 1970; Veronda and Westmann 1970; Scherer et al. 1984; Manschot and Brakee 1986; Fung 1993). The skin, which is mainly anisotropic, shows viscoelastic behaviour under stress (force) and strain (deformation). Three properties characterise such a behaviour: hysteresis, stress relaxation, and creep:
  - Hysteresis implies that the stress–strain curves corresponding to strain loading and strain unloading are different.
  - Stress relaxation corresponds to the decrease of stress under a constant strain.
  - Creep describes the increasing strain under a constant stress. Under compression, the skin exhibits a phenomenon called the *Poisson effect*, whereby the skin bulges perpendicularly to the compression direction.

  Facial muscles are skeletal muscles. They are sometimes called mimicry muscles because one end of the muscles is inserted into the superficial fascia of the skin. Facial muscles are very difficult to individualise. They are difficult to separate from the skin due to their superficial location and yet they are all interwoven with each other. Muscles do not contract linearly but show a viscoelastic behaviour.
  Finite Element Methods (FEM) (Larrabee 1986; Pieper and Zeltzer 1989; Pieper 1991; Deng 1988) have been applied to simulate the viscoelasticity properties of the skin. These models have mainly been applied to facial surgery simulation. They model the skin and muscle actions with good approximation, but the complexity and duration of the computation forbids its use in interactive applications for the time being. The computation time even on very powerful machines does not allow

for real-time animation. FEM has also been used to model lip shapes during speech (Basu and Pentland 1997).

- *Procedural model*: This method is not based on biological studies. Rather, the idea is to simulate the action of a muscle by a few parameters. Muscles are simulated by specialised procedures (Magnenat-Thalmann and Thalmann 1987). These procedures are called Abstract Muscle Action (AMA) and can have up to 24 parameters. They work on specific facial regions that correspond to one muscle and compute the displacement occurring under muscle contraction. This method was developed in close relation with the definition of FACS. We now give an example of the procedure 'upper lip raiser'. The upper lip under this action looks rather like a wave with a maximum at the mid upper lip and a minimum at both lip corners. The procedure involves all the points of the upper lip. Different parameters are defined for the mid-points and for the lip corners where different translation displacements apply.
- *Free form deformation*: The technique of free form deformation (Coquillard 1990; Coquillard and Jancene 1991) and rational free form deformation can be applied to model muscle action (Kalra et al. 1992). A deformation box is set to act on a set of points. The box can stretch, squash or bend. The points inside the box are moved according to the next shape of the box. In the case of rational free form deformation a weight is assigned to each control point of the box, giving more control over the deformations. The face is decomposed into regions based on anatomical considerations. Each region corresponds approximately to one muscle or to a group of related muscles. A parallelepiped box is defined to include all the points of each region. Each box affects only the points inside one region and therefore simulates the action of one muscle. Functions of deformation can be developed to simulate the behaviour of different muscles. These functions are: stretch and squash (pull and push action in a linear direction), shear (pull or push action on two parallel sides of the box in opposite direction), shift (pull on one side and push on the other side), circular (radial motion from the center point), rotational (rotation of the points around the center point) and hybrid (combination of actions). Any type of muscle action can be created with these functions. For example, orbital muscles can be simulated with the circular function while the shear function applies to the sheet muscle type.

2.8.3.2 Recommendations on facial modelling

In this section, we give recommendations on each type of facial modelling technique. If the system has to deal with limited bandwidth, parametric models are the most suitable (see the use of MPEG-4 in Section 2.9.5.2). But if realism and naturalness of facial movement is desired, physically-based models are more appropriate.

- *Parametric model*: As Parke (1991) pointed out, one major difficulty with parametric models is to develop a complete set of parameters; that is, a set that can describe any facial expression and any facial configuration (see Figure 2.19). Moreover, parameterised models do not model movement propagation and do not simulate muscle movements. On the other hand, the precise control of parameterised models is valuable in reproducing exact lip shapes during speech. The parametric model has the advantage of simplicity and low data storage requirements.

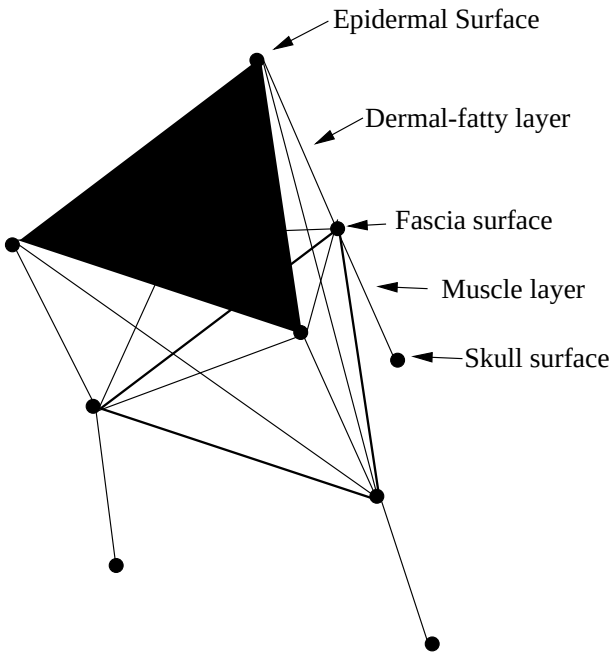

Figure 2.18: Fascial tissue layers (from Parke and Waters 1996)

- *Physically-based model*: This model is the most appropriate for simulating the skin. Simulating the visco-elasticity of the skin using a three-layer lattice produces subtle facial expressions but computation time increases greatly. One major disadvantage compared to parametric models is the difficulty of model control. Knowing which muscle needs to be activated and with what intensity to perform complex expressions such as lip movements during speech is a hard task. This is one of the reasons why parametric models have been chosen when dealing with lip movement during speech. Using EMG measurements of muscle contraction to drive a facial model may overcome this difficulty. This method has been successfully applied to lip shape modelling (Vatikiotis-Bateson et al. 1996).

    - *Structural model*: This model simulates movement propagation. The structure of the model in a hierarchy of regions and subregions is well-adapted to the definitions of AUs in FACS. The distinction of physical and functional information has the advantage of making the definition of an AU independent of the regions it is applied to. But this model cannot simulate wrinkles and neither models the different types of muscle action nor the visco-elasticity property of the skin. Other behaviour patterns such as non-penetration of the skull and volume preservation are not handled by the model.
    - *Muscle-based model and facial tissue model*: The muscle-based model is based on anatomical and biological studies. It can easily be applied to different models since the human head has the same set of muscles that are anchored on certain facial parts. This method has an advantage over the previous methods in that it is able to model different muscle

activities by changing the function of contraction. By using dynamic formulations, complex behaviour patterns such as the penetration constraint and volume preservation can be modelled, which cannot be done using other methods. One disadvantage of these methods is the regular deformation obtained during muscle contraction since visco-elasticity properties of muscles and skin are simplified to an elastic model.

- *Procedural model*: This method is based on empirical data and not on biomechanical studies. No movement propagation is considered. The definition of an AMA procedure is not independent of the other definitions, and the order of the procedure calls matters. On the other hand, the procedures are independent of the facial model. Another advantage of this method is that it allows the hierarchical definition of actions. AMA procedures define shapes on the lowest control level. But one can combine these procedures to produce facial expressions and/or lip shapes for speech.
- *Free form deformation*: The interactivity of this method offers a great advantage to the user. It is also more intuitive than defining a muscle by a vector, as in the muscle-based model.

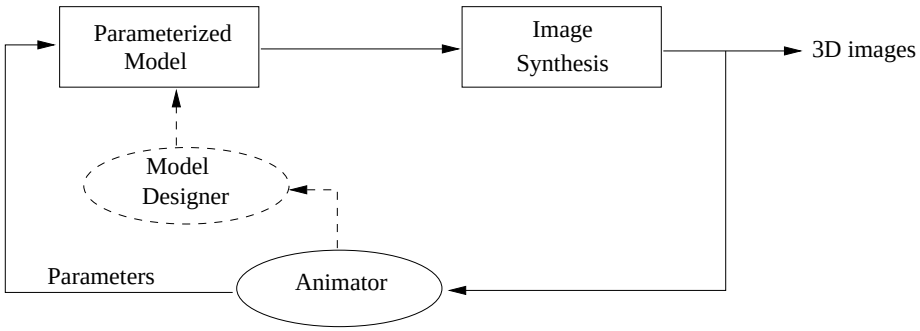

Figure 2.19: Parameterised facial model system (from Parke and Waters 1996)

2.8.3.3 Animation control

Animating a face by hand is a very tedious task which requires a skilled animator. A mechanism for animating facial models automatically is therefore needed. Three facial animation techniques are available: rule-based, analysis-based and performance-based animation.

- *Rule-based animation*: A set of rules drives the animation system (see Figure 2.20). Hierarchical structures and script languages are often used. Hierarchical structures allow the user to work at the phoneme, word, sentence, or emotion level rather than specifying each muscle action manually (Kalra et al. 1991; Patel and Willis 1991; Reeves 1990; Pelachaud et al. 1996; Beskow 1997a; McGlashan 1996). Faces have their own language. Facial expressions are not only related to emotions, but also to the intonation and the semantic content of speech. Some are tied to the intonation of the voice, some are used to highlight a word or to underline a pause. These relations can be encoded by a set of

rules working at different levels (e.g. phoneme, word, sentence) (Pelachaud et al. 1996). A script language offers a scheduling mechanism to coordinate and synchronise several parallel or sequential actions (Kalra et al. 1991). For example, actions can happen simultaneously, sequentially, triggered by certain actions, or after a certain lapse of time. Each action has an inherent default duration that can be modified if necessary, and can be specified in discrete time units or relative to previous or successive actions.

- *Analysis-based animation*: The analysis-based technique extracts information from live animations. The computed movement data are interpreted as muscle contractions and given as input to the animation system (Essa and Pentland 1994; Terzopoulos and Waters 1991). Examples include deformable contours (snakes) (Terzopoulos and Waters 1991) and optical flow (Essa and Pentland 1994; Mase 1991; DeCarlo and Metaxas 1996).
- *Performance-based animation*: This method has been introduced in Section 2.5.1. Various points are visually marked on a live actor and are tracked. Their movements can be used to drive a 3D model (deGraf 1990; Patterson et al. 1991; Litwinowicz 1994; Guenter et al. 1998).

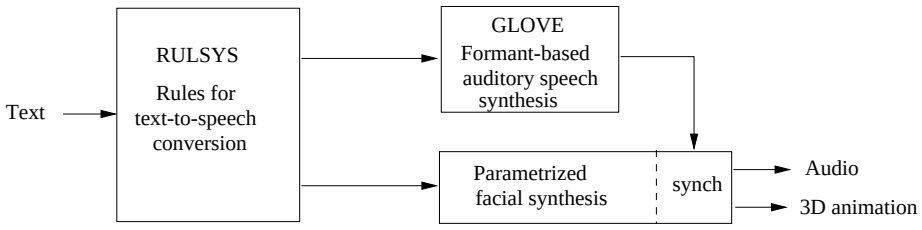

Figure 2.20: Overview of a audio-visual rule-based system (from Beskow 1995)

2.8.3.4 Recommendations on animation control

Depending on the application, either interactivity is acceptable or full automatism has to be achieved.

- *Rule-based method*: This method is quite straightforward, but has the handicap of not being interactive and the animation produced with such a method appears repetitive. If a raised eyebrow is assigned to appear with an accent, every accent will be marked by a brow action. People do mark accents with brow raising, but certainly not every accent. Fine tuning of the rules can become very complex as the number of the rules increases and as the rules get interwoven. Such a method can be used successfully to drive cartoon-like faces (Reeves 1990). A rule-based approach has the advantage of automatically driving the facial animation model from an input text (Pelachaud et al. 1996; Beskow 1997a).
- *Analysis-based method*: This method has great potential for enhancing graphics systems. It does not require any special markers or other intrusive devices. System robustness can be enhanced by achieving invariance of lighting conditions, head movements, and background (Petajan and Graf 1996; Meier et al. 1997; Revéret et al. 1997; Beymer 1995). It offers the possibility of extracting subtle facial actions with timing information. Temporal information on muscle

contraction could be obtained for facial expressions (Essa et al. 1996) and for speech (Cosi and Magno-Caldognetto 1996).

- *Performance-based method*: It has the advantage of reproducing animation with the right timing and movements since this information is directly extracted from a live video. However, animations other than those recorded cannot be generated. It is difficult to edit the recorded information, to manipulate it, and to change it. Moreover, it is intrusive since it requires the use of markers, of head mounted displays, and of special lighting. It is also recommended not to use fluorescent lights near video cameras due to flickering effects. Cameras with plastic lenses are also not recommended since they require constant re-calibration and de-fogging (DeCarlo 1998).

2.8.3.5 Lip shape computation techniques

- *Acoustic*: Early works (Boston 1973; Erber and deFilippo 1978) converted speech signals into visual signals on an oscilloscope. This visual signal had the form of an ellipse representing the lip shapes.
- *Cartoon-type*: Most cartoon-type movies use few mouth shapes to model speech. Speech units are clustered in 8–9 groups. They are (in SAMPA notation): /b,p,m/, /a,E/, /o,aU/, /o:/, /e:/, /f,v/, /d,n,g,k,l,r/, /s,t/ (cited in Emmett 1985). In cartoon-like movies the liveliness of the character is not given by perfect lip synching, but by the co-occurrence of movements (Reeves 1990; Emmett 1985; Kleiser 1989). Anatomically correct movement is therefore not the prime problem; rather, using exaggeration/simplification as well as some anticipation of the movement is often more valuable in producing more expressive animation (Lassiter 1987). A large number of mouth models (19 models) are used by British Telecom (Walker and Sheppard 1997). When phonemes are combined, a smooth transition between them is computed by interpolating the vertices of the mouth model between two consecutive frames.
- *Set of parameters*: Following research by phoneticians (Fromkin 1964; Benoît et al. 1990; Cosi et al. 1996), only few parameters are considered to define lip shapes. These parameters are (Benoît et al. 1990): the horizontal width of the lip, the vertical height of the internal lip contour, and the distance between a vertical profile and the lip contact protrusion. Parametric models, such as that of Parke (1972), can be extended to include these parameters (Hill et al. 1988; Lewis and Parke 1987; Nahas et al. 1988; Pearce et al. 1986; Cohen and Massaro 1990; Guiard-Marigny et al. 1994; Beskow 1995).
- *Coarticulation*: Several models of coarticulation have been proposed, differing primarily in their way of analysing the timing of event dependencies. Four main models of coarticulation can be distinguished: the time-locked model, the look-ahead model, a hybrid model (Kent and Minifie 1977), and the expansion model (Abry and Lallouache 1995).
  - The *time-locked model* is based on the principle that an event starts from an inherent time (a locked time). The protrusion influence due to a vowel appears at a given time before the vowel.
  - In the *look-ahead model*, the influence of a vowel on segments does not start from a given time but rather from the last preceding vowel (in the case of forward coarticulation) or the following vowel (in the case of backward coarticulation). The look-ahead model has been used by Pelachaud et al. (1996). Three parameters influence sequences of consonants: targets, features, and goals. There are three corresponding variants of look-ahead models.

    - In the *target-based model* (Kent and Minifie 1977), positions are invariant in the sense that the articulator (lip shape) is forced to assume a given target without regard to the pattern of muscle contraction or how such a position might be achieved. Only the final target is considered. Depending on the context (i.e. the surrounding segments), a given target may be executed differently and different muscular contractions may be involved. Beskow (1997a) derived a set of rules to compute lip shapes based on such a coarticulation model.
    - Coarticulation in *feature-based models* starts as soon as features involved at the articulatory level in segments are compatible with the features used to realise the current segment (Benguerel and Cowan 1974).
    - The *goal-based model* considers the sequence of goals to be achieved in computing articulator behaviour patterns.
  - The *hybrid model* (Kent and Minifie 1977) combines aspects of the look-ahead and the time-locked model. Coarticulation effects occur in two phases of influence. The first phase starts as predicted by the look-ahead model, while the second phase begins at a locked time. In the first phase the movement due to the influence of a certain vowel makes a slow appearance, while in the next phase the appearance of movement is faster.
  - The *expansion model* (Abry and Lallouache 1995) is based on the fact that the protrusion effect of a vowel can be expanded. The zone of influence depends on the number of consonants to the next (or from the previous) vowel but it cannot arise in less than a constant time.

  No single method can explain the coarticulation effects of all languages. Turkish and Swedish are apparently better modelled by the look-ahead model, while American rather appears to correspond to the time-locked model (Boyce cited in Cohen and Massaro 1993). To include the diversities of each language, Lofqvist introduced the notion of the *dominance function* for each articulator (Lofqvist 1990). A dominance function specifies the time-invariant influence (that is, the dominance) that an articulator has over the articulators involved in the production of preceding or succeeding segments. A two-phase model is proposed in which the influence of a given segment first increases, then decreases, having maximal influence at its own point of articulation. The dominance functions are characterised by several different parameters, including magnitude, duration and offset. Duration, which affects the time when an influence really starts, can be varied so as to simulate either the time-locked model or the look-ahead model. The variation of offset simulates differences in voicing (Cohen and Massaro 1993; LeGoff 1997). The magnitude represents the degree to which the current segment influences its environment.
- *Image-based*: This technique makes use of a library of lip shapes. It works at a more abstract level rather than at a feature level. Different coding systems exist:

  - *Hidden Markov Model (HMM)*: HMMs have been used by different systems (Yamamoto et al. 1997; Brooke and Scott 1994; Morishima 1996; Lavagetto and Lavagetto 1996; Adjoudani and Benoît 1996; Potamianos et al. 1997; Goldschen et al. 1996). An HMM is a finite state machine that represents the variation in time of visual and audio features. For each triphone made of a phoneme and its surrounding phonemes,

an HMM is built (Brooke and Scott 1994). The output of an HMM is a Principal Component Analysis (PCA) vector. PCA represents the image as a vector. Each image is stored as a vector made of principal component values (15 values are sufficient to encode 80% of the information). The corresponding image of the PCA vector is then output on the graphics screen. HMMs have been combined with vector quantisation (VQ) and ANN methods to generate lip shapes from speech. (Yamamoto et al. 1997).

- *Codebook*: A codebook of lip images can be stored (Bothe 1996; Woodward et al. 1992). The codebook is based on diphone clustering (e.g. /bb/, /ba/, /br/). The input text is decomposed into a sequence of diphones. The closest image chosen from the codebook is displayed. Image interpolation techniques smooth the transition between successive images.
- *Finite Element Method* (FEM): A 3D mesh of the lip shape is built from a linear elastic behaviour modelled by the FEM (Basu and Pentland 1997). Strain and stress values of muscle contractions are used to establish the equilibrium equation. The 3D lip mesh is trained on accurate 3D data extracted from video footage of human lip movements.
- *Morphing*: Real video footage of a person can be used to generate videos of the same person saying arbitrary text/utterances (Bregler et al. 1997; Ezzat and Poggio 1997). A set of phonemes is labeled automatically (Bregler et al. 1997) or manually (Ezzat and Poggio 1997) from training data, as well as from the new audio track one desires to animate. The system selects the closest mouth video image and stitches it into the background image using a morphing technique. Head direction and orientation have to be adapted accordingly.

- *EMG*: Once EMG measurements of the lip area are obtained, a mapping between muscle contractions and vocal tract articulators should be established. The mapping has to be dynamic since its relation is between muscle force and articulator acceleration. The final position of the articulator can be obtained by double integration from the acceleration data using the classical dynamic formulation: $m\ddot{x} + b\dot{x} + kx = 0$ where m, b and k stand for: mass, viscosity and stiffness and $\ddot{x}, \dot{x}$ and x for acceleration, velocity and position. Lip muscles have practically no mass (Gray 1973) and are heavily dampened and stiffened due to their attachment to the viscoelastic skin structure. This model is well suited to driving muscle-based models (described in the previous section).

2.8.4 Building conversational agents

It has long been a dream to simulate spoken conversation by computers. It is still a big challenge. Some systems simulate face-to-face conversation between a synthetic agent and a user (Thórisson 1997; Takeuchi and Naito 1995; Beskow 1997a; Ball and Ling 1994; Bates 1994; Beskow et al. 1997; Nitta et al. 1997) (see Figure 2.21 for a typical system architecture) or between two synthetic agents (Cassell et al. 1994). Such systems embody rules from cognitive science studies (Ekman 1979; Beattie 1981; Kendon 1990; Argyle and Cook 1976; Goodwin 1986; Scherer 1980; Duncan 1974; Condon 1988). Conversation is organised as an exchange of speaking turns (Schegloff and Sacks 1973; Sacks et al. 1974). Speech is the main stream of information but not the only one. Nonverbal signals are important means of conveying meaning and information at the linguistic, semantic and emotional level. An accent may be marked by

a pitch rise, but also by a raised brow, a head nod or even a blink. Raised eyebrows can mark an accent, but they can also be a signal of surprise, or they can mark questions, especially syntactically unmarked questions. A large number of studies have been conducted to understand the role of non-verbal cues in human–human communication (Ekman 1992; Chovil 1989; Harper et al. 1978; Fridlund 1994). They also point out the property of synchrony, linking the verbal or nonverbal signals (Condon 1988; Kendon 1990).

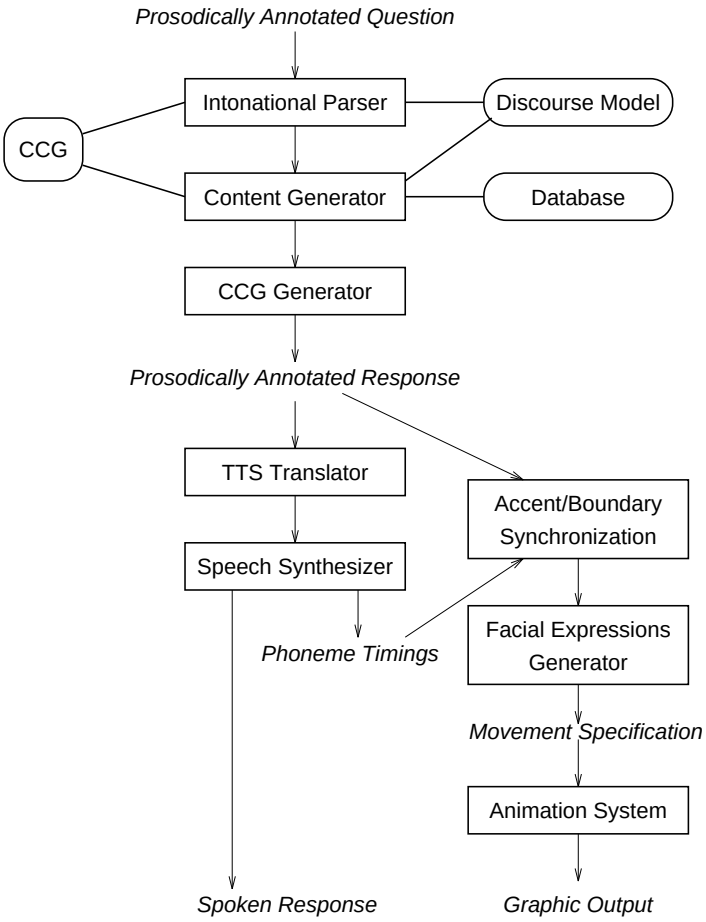

Figure 2.21: Architecture of a system automatically generating an answer with the appropriate intonation and facial expression starting from a query (from Pelachaud and Prevost 1995)

2.8.4.1 Conversational nonverbal behaviour patterns

Different human behaviour patterns are exhibited during the communication process:

- *Turn-taking systems* refer to the protocol followed during an exchange of speaking turns (Duncan 1974; Sacks et al. 1974; Goodwin 1986). Different modalities are involved during the exchange of speaking turns: gaze direction, hand

gesture, body posture, paralanguage, and facial expression. Following such a protocol ensures a dialogue with no overlap or interruption during the exchange of speaking turns between conversants.

- *Backchannel signals* indicate the listener's participation in the conversation (Duncan 1974; Kendon 1974) and give valuable feedback to the speaker.
- *Emotions* play an important role in human conversation. Ekman (1989) claims to have found six emotions associated with universal facial expressions, namely anger, disgust, fear, happiness, sadness, and surprise. Later on, Keltner (1995) added the emotion of embarrassment to this list. Most of the existing facial animation systems use this set of emotions (Patel and Willis 1991; Kalra 1993; Pelachaud et al. 1996; Waters 1987). For each emotion a facial prototype can be expressed with FACS (see Table 2.5).
- *Hand and arm gestures* contribute significantly to speech. They can be a reproduction of what is being said (opening widely the arms in a round shape in front of oneself while saying "She is in her last month of pregnancy"), an addition (opening the fingers of one hand vertically in a C shape as to mean a small quantity while saying "I want that much"), a substitution (gestures are more easily performed during word search; or in a noisy environment such as a bar, holding up the hand with three fingers extended to mean "I would like three beers"), contradiction (hand showing the number one while saying "two") (Poggi and Caldognetto 1996). Hand gestures can be classified into four symbolic classes (Cassell et al. 1994):
  1. *deictic* indicates a point in space;
  2. *iconic* depicts an object;
  3. *metaphoric* represents an abstract idea;
  4. *beat* marks the utterance rhythm.
- *Facial expressions* constantly accompany human conversation (e.g. Figure 2.22). They may accompany the flow of speech, punctuate an accent, a pause (Ekman 1979; Scherer 1980; Chovil 1991) (one can raise the eyebrows on the accented word "amplifier" in "The British AMPLIFIER produces clean treble"). They can also replace a word (frowning to mean "I don't understand"), or refer to an emotion (Ekman 1979) (showing a happy face while mentioning a past event "Yesterday it was a really nice day").
- *Gaze* is a powerful means of communication. Eye and head movements may be used to control the communicative process (Argyle and Cook 1976; Beattie 1981; Kendon 1990). Their main function in a conversation is to regulate and synchronise the flow of speech (Argyle and Cook 1976) (breaking the gaze when taking the speaking turn), to look for feedback (the speaker looking to check whether the listener is following), to express emotion (staring at the object of fear), to influence another person's behaviour (looking directly into his eyes to exert power), to show one's attitude toward the other (friends look at each other more often). Head movements can also replace words (shaking the head while refusing something).

2.8.4.2 Interpretation of communicative signals

Simultaneous signals in different modalities are different tools for achieving the same goal. They do not convey the same information, rather they frequently complement each other. Each emitted signal cannot be interpreted separately, but the overall meaning of the discourse is the result of the combination of all signals: gaze, facial expressions, words, intonation, hand gestures, and body

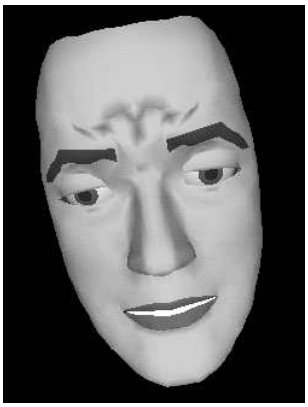

Figure 2.22: Facial expression of imploration (from Pelachaud and Poggi 1998)

Figure 2.23: Facial expression accompanying the accented word 'amplifier' (from Pelachaud and Prevost 1995)

postures are all active elements of the conversation. The different modes of communication overlap. The redundancy of the information coming from different modalities have been shown to reduce the error rate in a human–machine conversation (Bolt 1987). Therefore both verbal and nonverbal behaviour patterns are considered for next generation multimodal human–computer interfaces, including facial animation systems. More information on capturing non-verbal cues can be found in Section 2.4.2.

2.8.4.3 Synchronisation on communicative signals

Synchronisation between all modalities is challenging in such applications. As mentioned above, speech has to be synchronised with lip movement, but this includes also facial expressions, gaze, and hand gestures. A delay in the synchronisation process is easily perceived by the user and is very disturbing. Synchronisation should occur at computation time as well as during the output

phase. Nonverbal signals are highly synchronised with speech (Condon 1988; Kendon 1974). Synchronisation occurs at different levels of the utterance: an eye-blink starts closing on a phoneme, remains closed on the next phoneme(s) and starts opening on the following one; a raised eyebrow is synchronised at the word level, an emotion at the phrase level; hand gestures are also synchronised at the phrase level. The hand stops gesturing at the end of the spoken turn (McNeill 1992). Moreover, there is intersynchrony between speaker and listener (Kendon 1974) which serves as a metric for the conversation. Change of body position, gaze movement patterns of the speaker and of the listener follow this metric.

2.8.4.4 Pros and cons of talking faces

For various reasons, establishing the pros and cons of the use of talking faces in a multimodal user interface is still very difficult. For one thing, the relation between 'real' facial expressions and the meanings they convey is not fully understood, and these facial expressions may be highly ambiguous; in practice, much work has been based on simplified stereotypes of facial expressions. A further reason is that the state of the art of facial modelling and animation has not yet achieved a significant degree of realism and naturalness (except if very sophisticated performance-based techniques (Guenter et al. 1998) are used, such as the techniques used in the film industry). Another reason concerns the types of application in which a 3D face should be used. It is still unknown to what degree a 3D face makes a user interface "better". No criteria establishing the usefulness of a 3D face have been elaborated. Moreover, nobody is entirely sure if using a 3D face exhibiting conversational skills is more appropriate than using a 2D face or even using a caricatural face. It may even be that using a human face is less desirable than using cute cartoon type animals such as dogs or parrots.
Different studies (Walker et al. 1994; Takeuchi and Naito 1995; Koda and Maes 1996) have tried to answer some of these questions and problems, and have suggested that productivity and user performance are enhanced by such a multimodal speech system. They have shown that a talking agent makes the system more attractive to the user for the following reasons:

- The user spends more time interacting with the system. He has positive feelings towards it, which is a positive effect of using a talking agent. But interpreting the facial expressions of the agent requires effort and attention on the part of the user. His performance might deteriorate if he requires time to interpret what the agent is saying and which actions it is performing (Brennan and Ohaeri 1994), leading to loss of concentration on the task he is performing. This drawback is mainly due to the technical aspects of facial modelling and should be resolved as new models will be developed.
- The user might not be able to interpret all the subtle communicative facial expressions the agent is exhibiting (Takeuchi and Naito 1995). Nevertheless he will respond to them in a conscious or unconscious manner. He can react to the agent's behaviour without being fully conscious of it. This result is quite encouraging for using a talking agent.
- Although current synthesised faces might not make the agent appear more friendly, showing some emotional expressions increases user attention (Walker

et al. 1994). For example, the user spends more time interacting with an agent with a stern face than with a neutral expression. In some applications, user attention is more important than perceived friendliness of the synthetic agent, for instance in education applications.

2.8.4.5 Recommendations on building conversational agents

Building conversational agents requires the inclusion of nonverbal signals in the agent's behaviour patterns. Indeed, if the synthetic agent does not show natural movements, the user will have the feeling of talking to an inexpressive robot. But the choice of the signals to exhibit, and the time of appearance of these signals in the conversation, are very important:

- Image size: It is recommended that the image size be a minimum of 100×100 pixels (Schomaker et al. 1995a). Smaller image sizes would result in a lack of details, especially in the lip area.
- Appropriate signals: As humans we are very sensitive to any errors perceived in the emitted signals. Wrong movements, wrong timing of the appearance or disappearance of the signals, as well as wrong duration of the signals are immediately detected. This is especially true as synthetic agents are becoming more and more realistic: in a 3D model, the fine simulation of muscle actions and of skin elasticity and good lip movements create the illusion of a realistic model, and therefore humans are becoming more demanding as to the quality of the animation. The use of cartoon faces, caricatures, or other non-human animated objects (the animals or even lifeless objects which Walt Disney animation has accustomed us to) bypass such difficulties.
- Timing of signals: Synchronisation among the nonverbal signals and speech is an important property of human–human conversation. Nonverbal behaviour patterns do not occur randomly but at specific times during speech. If nonverbal signals are badly placed (such as a raised eyebrow appearing on the wrong non-accented word) the user will be confused: which signals (verbal or non-verbal) should prevail to interpret which is the accented item? That is, which feature designates the accented word, the raised eyebrow or the pitch accent?
- Lip shapes: Speaking rate has to be considered during the computation of lip shapes, as it has a strong modifying effect on lip shapes. As the speaking rate increases, lip shapes tend to be less articulated (hypo-speech), whereas during emphasised speech exaggerated articulation is produced (Schomaker et al. 1995a).
- Eye blinks: Eye blinks mark accented items, but also perform the biological necessity of keeping the eyes wet. On the average, they appear every 4.8 seconds (Argyle and Cook 1976), lasting about 1/4 sec., with 1/8 sec. of closure time, 1/24 sec. of closed eyes, and 1/12 sec. of opening time (Grant 1969).

2.8.5 On-line character and handwriting recognition

Written language recognition transforms language represented in the spatial form of graphic marks into an equivalent symbolic representation as ASCII text. In this section on handwriting recognition we first discuss the challenges of handwriting recognition and taxonomies of handwriting recognisers. Secondly, issues in and devices for sampling handwriting input are described, along with

file formats for storing handwritten input. We then review the main approaches and the state-of-the-art systems in character and handwriting recognition. The material presented is based on several recent surveys (Govindaraju et al. 1997; Hildebrandt and Liu 1993; Manke 1998; Nouboud and Plamondon 1990).
For evaluation, the field of handwriting recognition has adopted the same evaluation methodology as the field of speech recognition: measuring the item accuracy (character, digit, or word accuracy) by aligning the output of a recogniser on a benchmark testset against the "truth". Since this methodology is well-known from speech recognition, the evaluation of handwriting recognition systems is not further discussed here.

2.8.5.1 Taxonomies of handwriting input

Handwriting recognition has many challenges in common with speech recognition, including writer independent recognition, recognition at the level of characters or digits, words, or sentences; writing styles (printed versus cursive, North American versus European), vocabulary size, and hardware dependencies.
According to the mode of data acquisition used, automatic handwriting recognition systems can be classified into on-line and off-line systems. In *off-line systems* (which can also be classified as Optical Character Recognition (OCR) systems), the handwriting is given as an image, without time sequence information. In *on-line systems*, the handwriting is given as a sequence of coordinates that represents the pen trajectory. For integration in multimodal interfaces, on-line systems are required. The following discussion will therefore focus on on-line systems.
Handwriting recognition can be either at the level of isolated characters (or digits), at the level of words, or at the level of sentences.

- *Character recognition* is a typical pattern recognition problem: shape and time features are extracted from the trajectory (given as time sequences or spatial representations) and are used to assign it to the appropriate class. Artificial Neural Networks (ANNs), Hidden Markov Models (HMMs), and hybrid approaches that combine neural network modelling techniques with HMMs have been successfully employed as classifiers for character recognition.
- *Word recognition*: There are two basic approaches to word recognition, which correspond to different theories of human cognition: The *analytical approach* first identifies the individual characters (using character recognition methods), and then builds word-level hypotheses from character-level hypotheses. In contrast, the *holistic approach* identifies the word directly from its global shape. In both cases, constraining recognition to a vocabulary increases accuracy substantially. Algorithms of analytical and holistic handwriting recognition are discussed further below.
- *Sentence recognition*: Recognition of sentences builds on word-level recognition methods. In addition, language models are used to incorporate statistical information about word sequences, similar to the use of language models in automatic speech recognition systems. For instance, a trigram language model increased the performance of an on-line handwriting recognition system with a 21,000 word vocabulary from 80% to 95% (Srihari and Baltus 1993).

2.8.5.2 Input devices for handwriting input

Hardware necessary for the processing of handwriting input (digits, characters, or words) in multimodal applications includes the device used to sample handwriting and methods of storing handwriting input. This section summarises available digitisers for handwriting, and the technical requirements determining the choice between different digitisers. The UNIPEN file format to store handwriting input is presented later in Section 2.9.
Different digitisers can be used to sample on-line character and handwriting input: digitising tablets (e.g. WACOM tablets), touch-sensitive displays (resistive or capacitive technology), and light pens. Recently, LCD tablets have become available (e.g. WACOM PL series). The key usability issue is in how far the device achieves the feel of paper and pencil. Known usability problems of current devices include: no immediate visual feedback (all graphic tablets), significant delay in sampling of input device movement (some light pens, LCD tablets, and resistive touch-sensitive displays), and no possibility of resting one's wrist while writing (capacitive touch-sensitive displays).

2.8.5.3 Recommendations for handwriting input devices

The following technical requirements have to be considered when deciding on an input device for handwriting input (cf. Hartung et al. 1996):

- *Usability of device*: Check for the following properties: visual feedback on handwriting movement, sampling of writing movement, possibility of resting one's wrist while writing. The minimum sampling rate is 5 points per stroke, or about 50 samples per second.
- *Sampling rate*: From a theoretical point of view, the Nyquist theorem suggests that sampling rates of 15–20 samples/second are sufficient for reconstructing the signal. However, for many applications it is easier to use higher sampling rates of 50–100 samples/second than to sample at much lower rates and reconstruct the trajectory using interpolation techniques.
- *Resolution and accuracy*: Resolutions of 0.02–0.1 mm and accuracies of at least 0.1 mm are recommended.
- *Sampling bursts*: Depending on the type of handwriting input, different typical durations of one input item can be identified. For handprint characters, the duration is less than 1 second, for cursive words less than 10 seconds, and for phrases typically more than 10 seconds.
- *Sampling modes*: *Continuous sampling* is equidistant in time. The sampling can occur either only when the input device touches the writing surface (pen-down only), or both during pen-down and pen-up phases. *Tracking* generates samples whenever a threshold distance has been travelled with respect to the last sample, i.e. equidistant in space. *Pointing* generates samples when the user taps on the writing surface with the input device.
- *Signals*: Currently available digitisers provide some of the following signals: x/y position, pressure, pen-up/pen-down information.

Several simple file formats for storing handwriting input have evolved. A minimal data format captures the x/y location of the pen and pen-up/pen-down information. A time-stamp for each sample, and pen pressure may be added. For scientific use, the UNIPEN format has established itself as a pseudo standard (see Section 2.9).

2.8.5.4 Holistic and analytical approach to handwriting recognition

Since handwriting recognition shares many challenges with speech recognition, it is not surprising that similar algorithms and techniques are successful. With adaptations of the preprocessing components and the topology of the basic modelling units, a continuous speech recognition system can be trained on handwriting data and achieve very reasonable performance. For instance, using the BYBLOS continuous speech recogniser without any changes to its algorithms, a word accuracy of more than 95% was achieved on a 3,000 word vocabulary (Starner et al. 1994). However, specialised handwriting recognisers can achieve better performance. In the following, we will therefore review handwriting recognition algorithms. The next paragraph briefly describes the different features that are extracted from the input image and that serve as input to the ensuing classification step. The subsequent two paragraphs review the main algorithms developed for the classification of handwriting: holistic and analytical approaches to word and sentence recognition.

Features extracted from the input image in order to identify characters within the handwriting input can be classified into local and global features. *Local features* represent the main topological characteristics of a small subsection of the trajectory. *Global features* represent the relationship of different line segments within a trajectory. While local features are applicable to any character set, global features attempt to capture specific characteristics of certain character sets (e.g. strokes in Chinese characters). For a detailed discussion of different local and global features the reader is referred to Hildebrandt and Liu (1993); Manke (1998).

Holistic approaches to handwriting word recognition identify words directly from global shapes. After extracting features, standard pattern classification methods are applied to assign the shape to one of the words within the vocabulary. Unlike some analytical methods that can recognise arbitrary words, holistic methods therefore have to constrain the search to a given vocabulary. The following features have shown to be useful for holistic handwriting recognition: word contour (e.g. ascenders, descenders, holes, i-dots), length of word (e.g. estimated by the number of crossings of the center line), and identification of "significant" visual structures, called *graphemes* (Hildebrandt and Liu 1993). Additional methods are necessary to make holistic recognition feasible for large vocabularies. Lexicon reduction determines the set of words from a large lexicon (vocabulary) that is likely to match some handwritten input (Madhvanath 1996). Performance of holistic methods is sufficiently high for small vocabularies (e.g. 98% on a 10 word vocabulary, Farag 1979). Lexicon reduction can make holistic methods applicable to large vocabulary tasks: a 3,000 word lexicon can be reduced to 50–100 words with a 95% accuracy (Madhvanath 1996).

Analytical approaches to handwriting word recognition first identify the constituent characters. Then, based on character-level information obtained in the first step, a second step identifies word-level hypotheses. Analytical approaches can be further classified into approaches with *explicit segmentation* (also called OCR postprocessing), and approaches with *implicit segmentation*. OCR postprocessing has two distinct stages: the first stage identifies sequences of characters, and the second stage matches character sequences on ASCII rep-

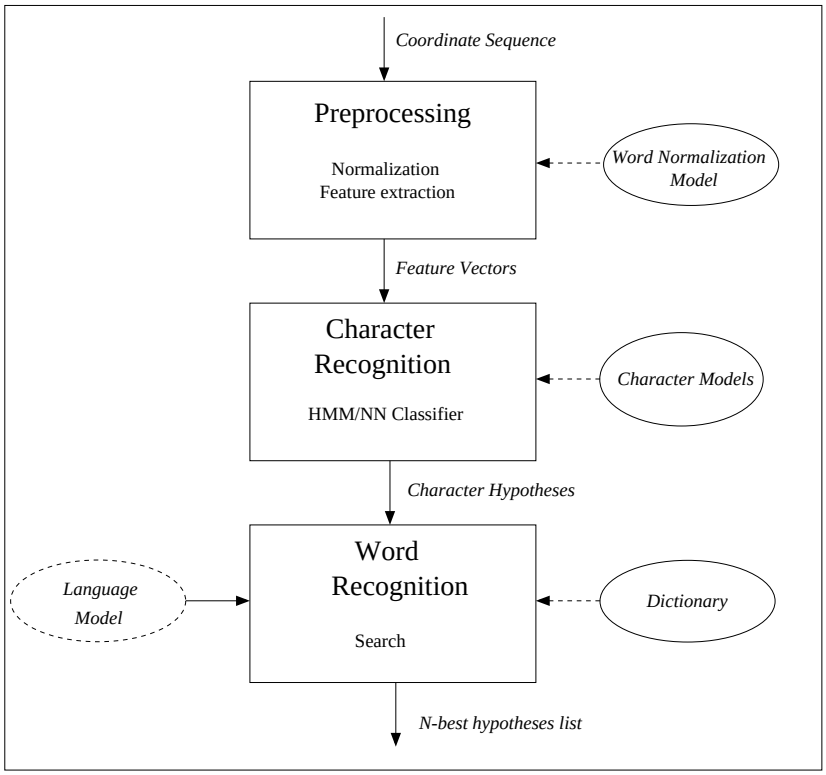

Figure 2.24: Handwriting recognition with explicit segmentation

resentations of words.
Figure 2.24 shows the basic architecture of a handwriting recogniser following the explicit segmentation approach.
Approaches with implicit segmentation use a lexicon to drive the segmentation and the recognition process. In a single step, both character and word-level information is used to match the input with words within a given vocabulary. For sentence recognition, the search is typically supported by a statistical language model. Figure 2.25 shows the basic architecture of a handwriting recogniser following the implicit segmentation approach.
While approaches with implicit segmentation have superior accuracy, they inevitably fail when the word input is not present in the given vocabulary (new word). OCR postprocessing can be extended to recover from the presence of new words.
The best published writer independent recognition accuracies for analytical handwriting recognition systems are more than 95% for character recognition (Guyon et al. 1992), 93.4% for word recognition (with a 20,000 word vocabulary) (Manke 1998), and 86.6% for sentence recognition (with a 20,000 word vocabulary) (Manke 1998).

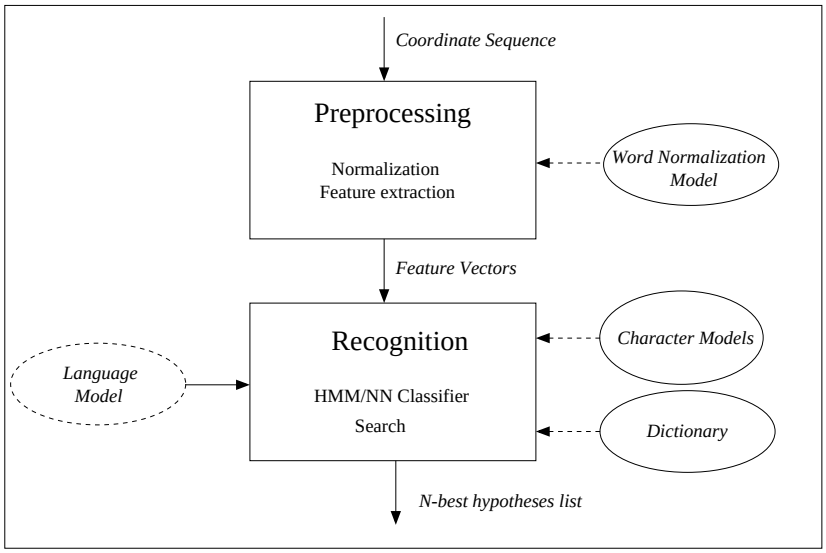

Figure 2.25: Handwriting recognition with implicit segmentation

2.8.5.5 Recommendations for handwriting recognition algorithms

- The holistic approach is applicable to off-line handwriting recognition and small-vocabulary on-line handwriting recognition; the analytical approach is easier to apply and achieves higher accuracy for most on-line handwriting recognition tasks. If a continuous speech recogniser is available, good performance can be achieved by retraining the recogniser on handwriting, of course with appropriately modified signal preprocessing.
- *Achieving sufficient recognition accuracy*: The following factors are correlated with high recognition accuracy on handwriting input: small vocabulary size, long input items, printed rather than cursive input, and finally a good match between the digitiser used to collect the training database and the digitiser used in the actual application.

2.8.6 Gesture recognition

With the development of gesture recognition algorithms, so-called *gesture-based interfaces* have become feasible. Gesture-based interaction with a computer offers an alternative to traditional interfaces driven by keyboard, menu and direct manipulation input. Gesture-based interaction may appeal to both novice and expert users for a number of reasons (Wolf and Morrel-Samuels 1987): objects, operations and optional parameters can be specified efficiently in the same movement, and learning and recall is facilitated since gestures tap into well-practiced human–human communication behaviour patterns, and use of pencil and paper. Gestures can be viewed as intuitive extensions of direct manipulation interfaces which have significantly improved the usability of human–computer interfaces.

Gesture recognition is a typical pattern recognition problem: gesture input (as 2D or 3D gestures) has to be assigned to certain gesture categories. This section on gesture recognition first summarises taxonomies of gesture input that may inspire innovative yet "natural" uses of gestures in multimodal applications. Then input devices to capture gesture input are briefly described. Finally, state-of-the-art recognisers for 2D and 3D gestures are reviewed.

2.8.6.1 Taxonomies of gesture input

An understanding of how people naturally use gestures is obviously necessary when considering gesture input in multimodal human–computer interaction. Several taxonomies have been proposed for categorising gestures that occur with speech (Blattner and Dannenberg 1990; Koons et al. 1993; Nespoulous and Lecours 1986). We adopted a taxonomy based on the dimensionality of gesture input, distinguishing pointing from 2D and 3D gesture input.
People sometimes employ gestures as the only means of communication, e.g. indicating affirmation and disagreement by a head gesture by a nod or shake of the head, or by means of hand gesture of pointing the thumb up or down. However, in most cases gestures occur simultaneously with other modalities, particularly accompanying speech. Usually, the following four gesture types are distinguished:

- *Symbolic gestures* can be directly translated to some meaning, e.g. a thumbs-up gesture to indicate agreement.
- *Deictic gestures* refer to objects or events in the environment, e.g. the famous "put-that-there" expression accompanied by pointing with the mouse or fingers.
- *Iconic gestures* refer to objects, spatial relations, or actions by describing them visually using a representation which is familiar to everyone, similarly to icons which represent applications in graphical user interfaces.
- *Metaphoric gestures* involve the manipulation of some abstract object or tool.

While symbolic gestures can be interpreted without context references, deictic, iconic, and metaphoric gestures can be interpreted meaningfully only with additional information from other modalities that occur simultaneously. However, this taxonomy has also been critically reviewed (Butterworth and Hadar 1989). In what ways can such gestures be used in multimodal human–computer interaction? According to Blattner and Dannenberg (1990), gestures in multimodal human–computer interaction can provide:

- *Semantic categories.* The gesture can identify the kind of action the user wants the system to perform, including: object manipulation, creation, and destruction; establishing relationships; retrieval or storage; confirmation; modification.
- *Attributes.* The gesture can refer to attributes of an action, including object location, direction, intensity, accuracy, size, orientation, and velocity.
- *Relationships.* Relationships like order, selection, aggregation, and implication can be indicated by speech and clarified using gestures.

Such gesture taxonomies can be used to discover new, yet obvious, gestures in order to convey information transparently. But finding and defining sets of

useful gestures will probably remain an application specific development effort until gesture-based interaction has been understood in depth.

2.8.6.2 Input devices for gesture input

Different input devices have emerged, classified by the type of gesture input (pointing, 2D gesture, 3D gesture). Pointing is typically sampled either using standard pointing devices (e.g. mouse), or using standard pointing devices emulated on a touch-sensitive display (e.g. finger moving on the display). 2D gestures are sampled using the same devices as handwriting input. These devices, and issues associated with them, have been described in Section 2.8.5. Finally, position trackers and sensing gloves (or data gloves) have evolved to track 3D gestures of hands and other body parts. Position trackers and sensing gloves are worn on the user's body or hand. They measure body and hand positions. More details can be found in Burdea (1996).
Burdea (1996) discusses technical details of position trackers and sensing gloves, including:

- *Sampling rate*: Although for most purposes, sampling rates of more than 30 samples per second are sufficient, some new devices offer sampling rates of up to 200 samples per second (e.g. "3-D probe" by Immersion & Co.).
- *Sampling bursts*: 3D gestures typically last a second, whereas position tracking is obviously a continuous process.
- *Sampling modes*: Absolute positioning determines the hand or head position with respect to a fixed system of coordinates, whereas relative positioning detects only incremental motion, relative to the current position.
- *Signals*: Both position tracker and sensing glove provide three-dimensional translations (x,y,z coordinates) and orientations (angles).

2.8.6.3 Recognition of 2D gestures

Given one or more pen strokes, each consisting of a sequence of coordinates, a gesture recogniser attempts to classify the stroke combination as one of many possible shapes. There are three main approaches to the recognition of 2D gestures: hand-coded algorithms, template-based approaches, and feature-based approaches.
While creating hand-coded gesture recognisers is feasible (e.g. Coleman 1969), it makes the resulting system difficult to build, maintain and modify. Since hand-coded gesture recognisers are useful only within the application they were created for, they are not described in this review.
Template-based gesture recognisers compare the input pattern with prototypical templates and choose the best matched template. Each gesture is characterised as a class of shapes and is represented by one or more templates (i.e. prototypical gestures of that shape). The input is compared to each template by first transforming the gesture to match the templates as closely as possible, then computing the residual difference using a *mean squared error (MSE) measure*. Allowable transformations include translation, rotation, and linear scaling along each coordinate axis. The template that yields the lowest residual difference below a set threshold is considered the best match. The input gesture can be labelled as unknown if all residual difference scores exceed the threshold.

Feature-based approaches first extract features from the stream of input coordinates, and then apply some pattern classification algorithm to assign the gesture to one of a set of pre-determined gesture categories. First approaches to creating feature-based gesture recognisers include a dictionary lookup method presented in Newman and Sproull (1979). Zoning features are derived by dividing the pen trajectory into zones, and by representing input strokes by the zones they traverse. A second early approach, linguistic matching, applies formal language theory to pattern recognition. An input gesture is represented by pattern primitives and composition operators that express relationships between primitives. Such representations can be parsed using a grammar that specifies how each gesture category can be generated from the pattern primitives. Shaw (1970) first introduced this approach, proposing a picture description language. Fu (1981) proposes a hybrid approach that uses statistical techniques to classify path segments and linguistic techniques to classify the input into a gesture category based on the relationships between the path segments.

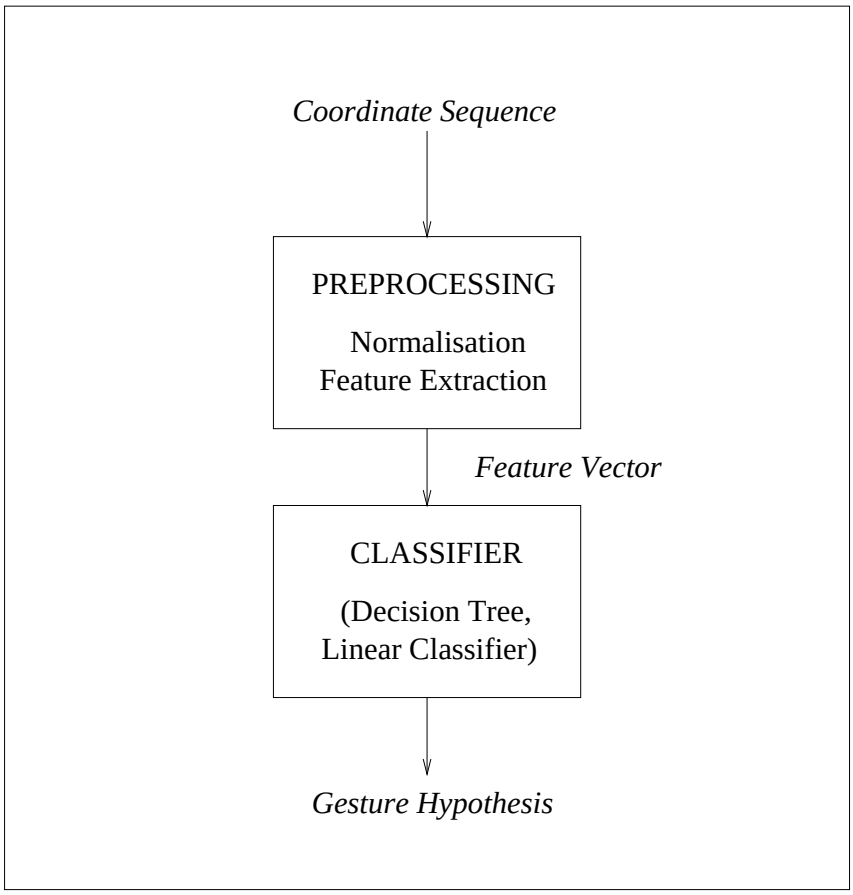

Figure 2.26: Architecture for a feature-based gesture recognition system

The generic feature-based approach to gesture recognition is illustrated in Fig-

ure 2.26. As in handwriting recognition (described in Section 2.8.5), features are extracted from the gesture input. Then standard pattern classification algorithms can be used to classify the gesture. Smoothing and filtering during feature extraction improve the recognition accuracy. The features characterise the shape, size, direction and orientation of the gesture. For a detailed description of a feature set that provides sufficient accuracy on small sets of gestures, see Rubine (1991b). Most gesture recognisers in the literature follow this general scheme. They differ with respect to the features and the pattern classification algorithm. The following briefly outlines different recognition algorithms:

- *Template matching* algorithms compare a given input template to one or more prototypical templates of each expected gesture. Variations of template matching include statistical approaches that derive classifiers from average feature vectors per class (Rubine 1991a) or via per-class variances and correlations of individual features (Hand 1982), and template matching based on multiple templates at different resolutions (Lipscomb 1991).
- *Decision tree* classification algorithms classify inputs represented as feature vectors by testing features one by one using conditions at each node until a leaf node is reached. The tree can be hand-crafted (Coleman 1969), but learning decision trees based on a sufficient number of gesture examples (typically less than 30 per category) is obviously recommended (Berthod and Maroy 1979).
- *Neural networks* have also been used successfully as classifiers in feature based gesture recognisers (Hollan et al. 1988; Shankar and Krishnaswamy 1993).

For integration in multimodal graphical user interfaces, toolkits that support gesture recognition would obviously be very helpful. Such toolkits free the designer from having to deal with the internals of a gesture recogniser. Some graphical user interface toolkits have been enhanced with integrated gesture recognition, thus facilitating the development of gesture-based interfaces. Rubine's GRANDMA system (Rubine 1991a) allows the developer to specify gestures with small sets of examples. Typically, 15–20 examples per gesture class are sufficient. Although his template based gesture recognition algorithm was designed for single-stroke gestures only, it can be applied without modification to multi-stroke gestures by processing a multi-stroke gesture just like a single-stroke gesture. This simple trick works as long as the set of gestures does not contain (multi-stroke) gestures that are ambiguous when interpreted as single-stroke gesture. Rubine's recogniser achieves a writer dependent accuracy of 97% on gesture recognition problems with no more than 15 gesture classes (trained on around 40 examples for each gesture class), and writer independent accuracies of around 85% (Suhm 1997). Further examples of interface toolkits that have gesture recognition integrated include CMU's GUI toolkits *Garnet* (Landay and Myers 1993) and *Amulet* (Myers et al. 1997), a toolkit for 3D virtual interaction (Bohm et al. 1992), and *HITS* from MCC (Hollan et al. 1988).

2.8.6.4 Recognition of 3D gestures

There are two main approaches to recognising movements of hands or other body parts in three dimensions (here called 3D gestures). The first approach directly captures gesture movements using dedicated input devices (e.g. sensing gloves or position trackers), and then applies pattern classification techniques.

The second approach uses computer vision techniques, observing the user with one or more cameras, and applying computer vision algorithms to segment and classify the image data. The big advantage of this method is that no intrusive devices are necessary, but the recognition is less robust, compared to the first approach.
As representative for the first approach (based on dedicated input devices), we review the 3D gesture recognition algorithm presented in Koons et al. (1993). Raw data from a sensing glove runs through two layers of abstraction before it is passed on to a gesture parser that integrates gesture information with information from other modalities. This general structure of a 3D gesture recogniser is shown in Figure 2.27. The first layer of abstraction transforms the raw data from the sensing glove into a feature representation. Koons et al. (1993) suggests three features: posture (for each finger: straight, relaxed, closed), orientation (direction of the hand's two normal vectors, the first out of the palm, and the second indicating where the hand is pointing), and hand motion. Thus every

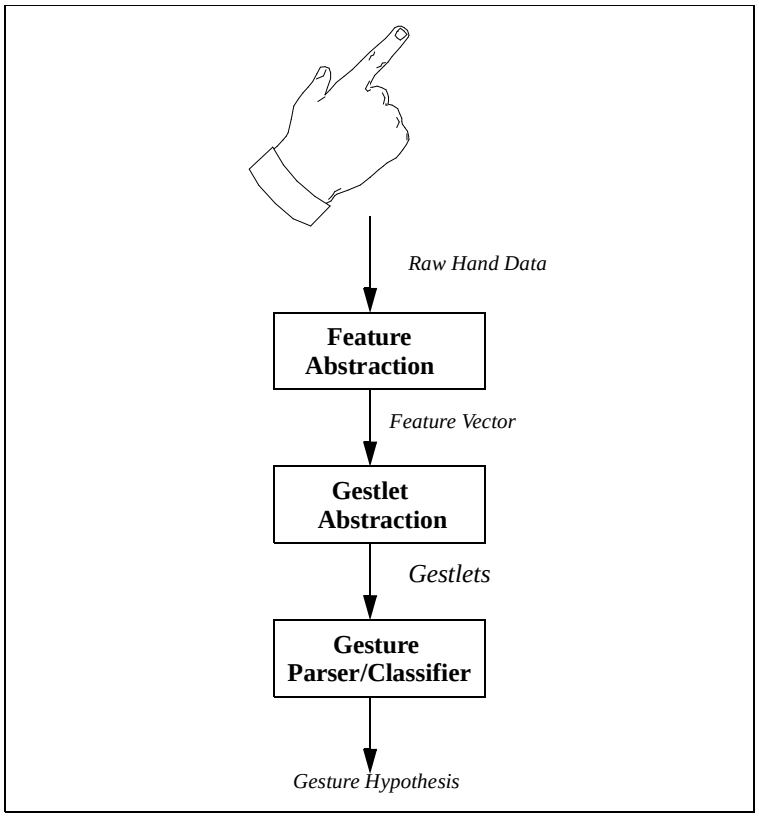

Figure 2.27: Architecture for a 3D gesture recognition system

The second layer of abstraction collapses the stream of triple feature tags into structures similar to speech phrases, called gestlets. *Gestlets* are pieces of ges-

ture that are formed from the stream of feature tags using certain rules. One rule could be to group all contiguous data sets between stops of the hand movement. The resulting stream of gestlets can be buffered, either for search if input from other modalities suggests important information may be contained in gestures that accompanied this input, or for classification using standard pattern classification algorithms. More information on glove based input can be found in a recent survey (Sturman and Zeltzer 1994).
Computer-vision based approaches to gesture recognition have been reviewed in Sharma et al. (1995), including a general framework for recognising 3D gestures within multimodal systems. The following is based on material from this review. Analysis begins with time-varying video images. The first task is to locate the active user who is performing gestures. The active user can be located either using motion information, visual cues (e.g. skin colour modelling), or non-visual cues (e.g. sound source localisation based on acoustic beamforming, Bub et al. (1995)). Second, the user's arms and hands have to be located. Segmentation based on colour histograms of human skin have so far shown the best results (Sharma et al. 1995). Segmentation can be aided by hand and arm models. Finally, the hand gesture has to be classified. Different approaches have been explored, including model-based approaches (Downton and Drouet 1991; Kuch and Huang 1995; Lee and Kunii 1995; Rehg and Kanade 1993): recognition using feature markers that are placed on finger tips (Cipolla et al. 1993; Davis and Shah 1993; Maggioni 1985) and other computer vision based approaches (Downton and Drouet 1991; Kuch and Huang 1995; Lee and Kunii 1995; Rehg and Kanade 1993). Sharma et al. (1995) present an algorithm that transforms stereo camera input into image geometric moment features, and uses a standard feature-based recognition algorithm (the HTK HMM toolkit, originally developed for speech recognition applications) to classify gestures based on geometric moment features.

2.8.6.5 Recommendations on gesture input and recognition

- Input device: In deciding between dedicated input devices (trackers, sensing gloves) and computer vision algorithms, intrusiveness and accuracy of sampling and gesture interpretation are important trade-offs. Intrusive tracking devices are attractive if the application requires the use of other intrusive equipment anyway, for example in virtual reality and wearable computing applications.
- Haptic (force, tactile) feedback is very important in order to ensure realism and avoid usability problems when tracking devices or sensing gloves are used.
- Sampling rates of 20–50 samples per second are sufficient for most applications.
- Gesture recognition: Use one of the publicly available GUI toolkits with built-in gesture recognition capabilities (e.g. CMU's Amulet (Myers et al. 1997)); otherwise, a simple template matcher can be implemented within a few days, which provides sufficient accuracy for applications with limited gesture sets (< 10 gesture classes).
- Combining handwriting and 2D gesture input is technically challenging, since no state-of-the-art handwriting recogniser can handle both types of input. Therefore, algorithms that automatically classify pen input in handwriting versus 2D gestures are necessary to process both types of pen input.

2.8.7 Technical issues

Multimedia development may involve a scripting language such as Macromedia's Lingo but also other WWW capable techniques such as the Java programming language developed by Sun Microsystems, or Microsoft's Active-X controls. Application Programming Interfaces (API) or "libraries" for speech output (SAPI) are provided by several companies such as Sun Microsystems (speech API), IBM (ViaVoice developer tools) and Microsoft.
When choosing a development tool or an API, questions related to its spoken output need to be answered:

- Does it enable recording and playing of spoken messages?
- Can these spoken messages be dynamically combined?
- Does it enable the production of speech output from a textual representation?
- Can synchronisation cues be incorporated in order to enable fine-grained synchronisation between speech and other media?

- Examples of call to API's and combination with other media?
- A list of authoring tools.
  "`http://lorien.ncl.ac.uk/ming/resources/cal/mmedia.htm`"
- Java Speech API
  "`http://www.javasoft.com/products/java-media/speech/index.html`"
- IBM ViaVoice SDK
  "`http://www.software.ibm.com/is/voicetype/dev_vvsdk.html`"

2.9 Standards and resources for multimodal/multimedia systems

2.9.1 Standards and resources for monomodal processing

2.9.1.1 Writing / 2D gestures

The *UNIPEN* format input is a standard for representing handwriting input. It can be used for 2D gesture input as well. It is defined as follows:

- Comments: Lines starting with *.COMMENT* are ignored.
- Header information, includes information on
  - the data: source of data after *.DATA_SOURCE*, contact after *.DATA_CONTACT*;
  - the data collection setup: some general information after *.SETUP*, input device after *.PAD*, information on the kind of data after *.DATA_INFO*, the alphabet after *.ALPHABET*;
  - the writer: id of writer after *.WRITER_ID*, writing style after *.STYLE*,
  - specific information on the writer, including handedness, age, sex after *.HAND*, *.AGE*, *.SEX*, and *.WRITER_INFO*.
- Actual data section: represents the pen trajectory as coordinate sequences of triplets {X, Y, T}. The information whether the pen went down or up is stored in a separate line *.PEN_DOWN*, and *.PEN_UP*, respectively.

The UNIPEN format can be extended to 3D gesture input data in a straightforward way by using 4-tuples {X, Y, Z, T}.
The Open Agent Architecture (OAA). The OAA architecture (a trademark by SRI) is useful for implementing distributed multimodal applications. It was recently made available publicly at `www.ai.sri.com/~oaa`.

2.9.1.2 Speech recognition and synthesis

With its speech recognition and synthesis systems, IBM provides ViaVoice development tools which can be used by developers to integrate speech technology in their human–computer interfaces (IBM n.d.).
Sun Microsystems are defining a more generic Application Programming Interface along with several other companies such as IBM, Apple Computer, Inc., AT&T, Dragon Systems, Inc., IBM Corporation, Novell, Inc., Philips Speech Processing, and Texas Instruments Incorporated.
The so-called JavaTM Speech API (Javasoft n.d.) should allow Java applications and applets to incorporate speech technology into their user interfaces. The API defines a cross-platform API to support command and control recognisers, dictation systems and speech synthesisers. The Java Speech Grammar Format (JSGF) will provide cross-platform control of speech recognisers. The Java Speech Markup Language (JSML) will provide cross-platform control of speech synthesisers.

2.9.2 Towards standards for multimedia systems

Standards are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics, to ensure that materials, products, processes and services are fit for their purpose. The International Organisation for Standardisation (ISO) is a non-governmental organisation established in 1947. ISO is currently working on several standards related to multimedia: a coded representation of moving pictures and associated audio (MPEG ISO 1998), the coding of multimedia and hypermedia information (MHEG ISO 1998), a middleware framework encompassing the management of distributed media resources (PREMO 1998). Standards are also being developed by W3C (SMIL 1998) or independent researchers (Bordegoni et al. 1997).

2.9.2.1 MHEG

MHEG will provide the concepts and mechanisms for the creation of multimedia applications. From MHEG's perspective, a multimedia application typically consists of a set of scenes. A scene, in turn, consists of a collection of media objects of various formats representing graphic, textual and audio-visual entities. Navigation between the scenes can be triggered by user interaction or events emanated from other scenes or media objects. The presentation of the media objects within a scene takes into account their spatial and temporal attributes and can be triggered by a variety of events generated as a result of user interaction or events emanating from other media objects. The playback of time-dependent media objects, such as streams of multiplexed audio-visual data, is also supported via VCR control functions (play, stop, pause, fast forward, etc). There is also the capability for capturing and processing the events generated during the playback. This generic standard will provide the coded representation of multimedia and hypermedia information objects to be interchanged within or across open applications and services by any means of interchange.
This standard should be applicable in any field where multimedia/hypermedia applications need to exchange information according to the following require-

ments:

- Need for an interchange form in which spatio-temporal and conditional relations between entities can be expressed, as well as interactivity (specific structures to support the dialogue with the end-user),
- need for an interchange form suited to real-time (requirement for an efficient mechanism to optimise interchange of composite structures in the correct sequence),
- need for a final form representation of information (without additional processing needed to restructure the information before its presentation).

2.9.2.2 PREMO: A middleware framework for the management of distributed media sources

As described in Duke and Herman (1998), SC24, the subcommittee of ISO/IEC JTC 1, completed work on PREMO (PResentation Environments for Multimedia Objects), a new standard that defines a middleware framework encompassing the management of distributed media resources, such as video, audio (both captured and synthetic), which is in principle extensible to new modalities such as haptic output and speech or gestural input. It also provides an object-oriented programming environment to support the development of such applications. PREMO also serves as a reference model. The PREMO environment allows existing media devices to interoperate and be interfaced to an application. While the ISO MPEG specification describes the details of a video format, PREMO concentrates on how an MPEG coder/decoder can be used together with other media processing entities like a graphics renderer.

2.9.2.3 SMIL

SMIL is a W3C recommended mark-up language for publishing synchronised multimedia presentations via the Internet. It uses sequential and parallel grouping tags and supports link-style navigation (asynchronous interaction). It has the ability to specify temporal subparts of media objects.

2.9.2.4 A standard reference model for intelligent multimedia presentations

Multimedia presentation design is not just a question of merging output fragments but requires fine grained co-ordination of communication media and modalities. A multimedia presentation system should be able to generate various presentations for one and the same information content flexibly, in order to meet individual requirements of users and situations, and resource limitations of the computer. Instead of being manually preset, the multimedia presentation is automated, with intelligence based on appropriate design decisions pertaining to presentation types and contextual knowledge. In Bordegoni et al. (1997) there is an attempt by several independent researchers in the field of "intelligent multimedia presentation systems" (IMMPS) to propose agreements on terminology, the functional definition of an IMMPS, and on a generic architecture which reflects an implementation independent view of the processes required for the generation of multimedia presentations.
This functional architecture consists of four expert knowledge modules (application, context, user, design) and five processing layers (control, content,

design, realisation, presentation) which share these expert modules. Data are exchanged between these layers from the goal formulation to be reached by the multimedia presentation down to the presentation sent to the user. The control layer receives goal formulations and related commands (start / interrupt, refine goal ...). The content layer contains several co-operating components: a content selection component, a media allocation component, and ordering components. The design layer generates multimedia design specifications thanks to a media design component and a layout design component. These specifications are processed by a realisation layer which produces the final presentation.

2.9.2.5 Ergonomic recommendation on multimedia

In commercial tools supporting the development of multimedia presentations, no advice is provided on how to solve the issues of content selection, media allocation and media combination with respect to cognitive constraints. Recommendations on the impact of multimedia on cognitive properties is discussed in Bearne et al. (1994). For instance, people are indeed able to attend to more than one stimulus at a time (drive a car while holding a conversation, listen to music while reading a book), but it is easier if these activities are dissimilar (different modes), highly practical and simple. In Hare et al. (1995), subjects were asked several questions, the answers to which they could find during the exploration of a multimedia system. It was observed that they found it more difficult if they had to combine several media for a single answer. Furthermore, subjects tend to spend more time on text than on other media such as video.

2.9.3 Towards standards for hypermedia systems

Several researchers aim at building standards for hypermedia systems (Grœnbæk and Wiil 1997).
The Dexter Model (Halasz and Schwartz 1994) attempts to provide a standard hypermedia terminology coupled with a formal model of the common abstractions found within contemporary hypermedia systems. This model is based on a layered conceptual data model including: application layer, communication layer, runtime layer, storage layer. The Flag Taxonomy (Østerbye and Wiil 1996) attempts to capture the functionality and interaction of hypermedia systems in such a manner as to aid classification. The Open Hypermedia Protocol (Goose et al. 1997) aims at enabling third party applications to access open hypermedia link service functionalities in a consistent and standard manner.

2.9.4 Architectures and toolkits for multimodal integration

General architectures for modality integration and publicly available toolkits that support the development of multimodal applications would obviously be useful. Unfortunately, most multimodal systems discussed in the literature are still ad hoc implementations; only few architectures have been proposed, and even fewer multimodal toolkits have been developed. This section first presents architectural principles for multimodal (and multimedia) interfaces (from Hill et al. 1992), and then briefly reviews multimodal architectures and specification languages/toolkits that the authors could identify from their survey.

2.9.4.1 Architectural qualities and principles

- *Blended modalities:* The user should be able to blend (simultaneously use) modalities at any time.
- *Inclusion of ambiguity:* The system needs to be able to handle ambiguous use of input modalities.
- *Protocol of cooperation:* The user should be able to interrupt input and output at any time (called *barge-in* for speech input).
- *Full access to the interaction history:* The history of interpreted interactions must be accessible on-line as well as after finishing an interaction session.
- *Evolution:* The interfaces should be open to improvement, either during or after interactions.

Architectural principles address the problem of input interpretation and representation, and the system architecture. There are different levels of interpretation: signal, syntax, semantics, pragmatics, and application. Processing should be shared wherever possible. In representing input, a compromise has to be made between performance, achievable by specialised device drivers, and recognition algorithms on the one hand, and homogeneity that allows for the blending of modalities on the other hand. With respect to system architecture, a separation of interface aspects from both application aspects and input interpretation aspects is recommended.

2.9.4.2 Architectures for multimodal integration

This section presents PAC-Amodeus, SRI's Open Agent Architecture, MIAMI's PVM, and the general CORBA architecture as architectures that either have evolved from implementations of multimodal applications, or that are highly suitable for the implementation of multimodal applications.

- PAC-Amodeus (Nigay and Coutaz 1993, 1995) supports the generic "melting-pot" fusion mechanism (described in Section 2.6.2.2), providing a reusable global platform that is applicable to the development of multimodal applications with synergetic (and thus any) cooperation of modalities. The core component of PAC-Amodeus is the Dialogue Controller – a set of cooperating agents that capture parallelism and information processing at multiple levels of abstraction.
- SRI's Open Agent Architecture (OAA) (Moran et al. 1997) provides access to agent-based applications through intelligent, cooperative, distributed agent-based user interfaces. It currently supports a mix of spoken language, handwriting, 2D gestures, in addition to standard input modalities (keyboard and mouse input). Since only the primary user interface agents need to run on the local computer, it lends itself to emerging mobile multimodal applications such as personal digital assistants (PDAs). The OAA has been used to develop a range of multimodal applications, including office assistants, map-based tourist information, summarisation of conversation, air travel information, multi-robot control, and emergence response systems, thus demonstrating its high grade of re-usability.
- MIAMI PVM (Schomaker et al. 1995b) supports multi-tasking within tcl/tk, based on the public domain software package PVM (Parallel Virtual Machine). Multi-tasking is crucial in the implementation of any multimodal system.
- Common Object Request Broker Architecture (CORBA) (Vinoski 1997) has emerged from an effort to standardise object-oriented design in distributed het-

erogeneous environments and communication in such computing environments. It is therefore targeted to a much more general audience than just developers of multimodal applications. However, since multimodal applications typically are distributed systems, and since object-oriented programming appears to be very adequate for highly modularised multimodal systems, future multimodal applications may increasingly be implemented CORBA compliant.

2.9.4.3 Specification languages and toolkits

- The *Multimodal User Interface Design Tool* (Kamio et al. 1994) supports rapid prototyping of multimodal interfaces. User interface objects can be placed on a panel, and links between objects describe plan-goal scenarios (what to do when a certain input event occurs). The design tool then generates a script that drives the multimodal interface.
- *Specification Language for Multimodal Application*: Martin et al. (1995) present a simple specification language to describe the cooperation of modalities in a multimodal application. For each task, it is specified what modalities can be used to express certain parameters, and how these parameters are passed to the execution module. For each modality, a list of elementary "events" describes what information chunks can be expressed in that modality, and how. The latter information is used by the recognition modules to define vocabularies.
  In addition, User Action Notation (Hartson and Gray 1992) has been used to specify multimodal interfaces.
- *Multimodal Grammar Tools*: Vo and Waibel (1997) present a toolkit for multimodal application development. The toolkit consists of a set of grammar tools that support the specification of multimodal applications using context free grammars, and the automatic transformation of such multimodal grammars into the configuration files necessary to implement the application with given multimodal component and integration modules. Graphic tools allow the interface developed to specify grammars using drag-and-drop interactions in a graphical user interface.

2.9.5 Notational systems

Several notational systems to encode body movements and facial expressions exist. Birdwhistell (1952) and Kendon (1990) developed a language to encode body movements. Their goal was to develop a methodology that analyses the communicative behaviour of the body and describe it as a linguistic model would (Birdwhistell 1952). Grant established a detailed repertoire of non-verbal behaviour patterns (Grant 1968) and especially about facial expressions (Grant 1969).
The following sections introduce the two most common notational systems used in facial animation: FACS (Section 2.9.5.1) and MPEG-4 (Section 2.9.5.2).

2.9.5.1 FACS

FACS has been developed by Ekman and Friesen (1978). It is designed to describe visible facial actions but it does not look at which muscles are activated to produce the facial actions. It is based on anatomical studies. FACS is composed of basic units called Action Units or AUs. An AU corresponds to the action of a muscle or a group of related muscles. Each AU describes the direct effect of muscle contraction as well as any secondary effects due to movement

propagation and the presence of wrinkles or bulges. A facial expression is the combination of AUs. Most of the AUs combine additively. But they may also be subject to rules of dominance (i.e. an AU disappears for the benefit of another AU), substitution (i.e. an AU is eliminated when others produce the same effect), alteration (i.e. AUs cannot combine). Table 2.6 lists all AUs.

2.9.5.2 MPEG–4

MPEG standards have so far concentrated on the issues of defining a coding scheme for audio and video data (Chen et al. 1994; Doenges et al. 1997). The most recent extension, MPEG-4, and especially the MPEG-4 Synthetic and Natural Hybrid Coding (SNHC) group, proposes an architecture for the efficient representation and coding of synthetically and naturally generated audiovisual information. This group is developing a set of parameters of human face and body description and animation. The group is also working on defining a representation for synthetic audio, static and dynamic mesh coding with texture mapping, on the description of an interface for TTS systems, and on a synchronisation scheme for audio and visual data. Another model based on a coding scheme similar to MPEG-4 has been proposed by Provine and Bruton (1996).
MPEG-4 also derived a standard for facial animation coding (Petajan 1997). A bitstream of sets of parameters define the geometry, texture and expression of the face. These sets can be either Facial Definition Parameter sets (FDP) or Facial Animation Parameter sets (FAP). FDPs control the shape of the face as well as its texture. FAPs control the animation of the face. They are defined for every frame of the animation of the facial model. All FAPs need to be expressed as a function of Facial Animation Parameter Units (FAPU). FAPU represent particular distances among the facial features (e.g, eye separation, mouth width). This process corresponds to the calibration phase and allows the FAPs to be applied to any facial model. The definition of FAPs is based on anatomical studies and corresponds to minimal (and basic) facial actions. An expression can be expressed as a combination of FAPs. The FAPs are applied to the neutral expression of the facial model. There are 66 FAPs clustered in different groups (such as outer lip, cheeks, eyebrow). Examples of FAPs are: vertical jaw displacement, horizontal displacement of right inner lip corner, vertical orientation of left eyeball, rolling of the tongue into a U shape. Apart from these 66 parameters there exist two other parameters defined at a higher level: one for visemes and one for expressions. Fourteen visemes have been defined (but no standard exists to convert phonemes into visemes) as well as six expressions of emotion (anger, joy, fear, sadness, disgust and surprise).

2.9.6 Face and audio databases

The FACS manual offers a large number of faces with different expressions (Ekman and Friesen 1978). At least one photo representing each AU with different intensity illustrates the manual. Many other photos of people showing a large variety of emotions can also be found. They are used to test one's ability to decode expressions with FACS.

The Digital Audio-Visual Integrated Database (DAVID) was developed by the British Telecom Laboratories and the Department of Electrical and Electronic Engineering of the University of Wales in Swansea, UK (Mason et al. 1996). The purpose of DAVID is to offer a database for research in speech or person recognition, synthesis of talking heads, facial image segmentation, visual speech feature assessment, and voice control of video-conferencing resources. The database contains material including isolated digits, the English-alphabet E-set, some "VCVCV" nonsense utterances, and some full sentences. Some of the speakers have been recorded over six months. Others had only one recording session. Most recordings were performed with plain background, but some were done in complex scenes. Some of the database elements show both front and profile images of the speaker, others are a frontal and profile close-up view of the speaker's lips only. This last set is useful for assessing automatic lip segmentation systems. The database contains data of about 100 persons.

Multimodal Verification for Teleservices and Security Applications (M2VTS) (M2VTS 1996) is a European project part of the ACTS program. The project is concerned with the issue of secured access to local and centralised services in a multimedia environment. The project has developed a database of 37 subjects. The database contains audio and visual material. The extended M2VTS database is composed of 250 subjects. It has taken 4x2 shots of each subject. The database contains speech and video elements. All the material is digital.

The US Army FERET database (Rauss et al. 1996) offers a very large collection of face images. The images have been collected under different lighting conditions, backgrounds, locations and times. The distance between the camera and the subject varies. For each individual, the database contains frontal and a variety of profile views taken at different times and under varying background and lighting conditions.

L. Bernstein and her colleagues have recorded a large database, distributed on 8 discs (Bernstein and Eberhardt 1986; Bernstein 1991; Bernstein et al. 1995, 1996a,b,c), four laser video discs and four optical discs using Panasonic optical discs. Recently a new video setting via the SGI Indigo 2 and a special purpose device (ACOM video recorder) has been used for data collection. In each recording session a teleprompter was used. It forces the speaker to look at the camera and it reduces speaker eye and head movements. Each video contains the speakers' electroglottograph signals on one of the audio tracks. The database contains six speakers. For each of them there is a large number of materials: monosyllabic nonsense syllables, disyllabic nonsense words, isolated words, sentences, nonsense sentences, repeated syllable strings, and special stimulus sets with particular properties for experimental tests. The materials of each recording session were established for particular scientific questions in mind.

The AT&T audio-visual database has been developed for bimodal ASR (Potamianos et al. 1997). The database was obtained using an SGI Indigo2

workstation. The image of the speaker is captured using a high quality 3CCD camera. The image resolution is high (560×480 pixels). Each recording is 30 sec. long with 30 interleaved frames per second. Four desktop quality microphones are used simultaneously to record four speech qualities (9, 14, 18 and 28 dB SNR). The database is divided into four parts. Part1 consists of a small vocabulary of highly confusable, mostly monosyllabic, isolated, "CVC" words. 50 subjects are used in Part1 to record 1250 isolated words. Part2 consists of sequences of connected letters. The same 50 subjects recorded 1250 connected letters. Part3 is under development and consists of phonetically balanced sentences from North American business news. Finally, Part4 will consist of spontaneous spoken utterances. The subjects are 10 women and 40 men. Among the men, 12 have moustaches and 9 have beards. Almost half of the subjects (21) wear glasses. 13 subjects are American English speakers.

Table 2.1: Results from Survey of Multimodal Interfaces – Part I: Domain, Input/Output modalities, and Cooperation

Reference	Application, Domain	Input	Output	Coop.	Fusion
		Multimodal Maps			
Neal & Shapiro '91	Road Maps	ASR, P	GUI, SS	E, S	sem
Koons et al. '93	Geogr. Maps & Blocks World	ASR, 2D GR 3D GR, GT	GUI	C, E	sem
Nigay & Coutaz '93	Air Travel Info Web Navigation Notebook	ASR, P, K ASR, P, K	GUI GUI, S GUI	C, E, CC, S E C, E	sem n/a sem
Vo & Wood '96	Calendar	ASR, 2D GR	GUI, SS	C, E, R	sem
Hollan et al. '88	Image Analysis	ASR, K, 2D GR	GUI, SS	C, E, S	sem
Cheyer '97 Cheyer & Julia '95	Tourist Inform. Office Assistant Robot Control Image Analysis Emerg. Dispatch	ASR, P, 2D GR and HR	GUI	C, E	sem
Oviatt et al. '97	Service Transaction	ASR, HR, 2D GR	GUI	C, E	sem
Martin '97	Tourist Map	ASR, P, K	GUI	all forms	sem
Sarukkai & Hunter '97	Train Scheduling	ASR, Gaze	GUI, SS	R	int
		Data Input			
Leopold & Ambler '97	Visual Progr.	ASR, HW, P	GUI	C, E, S	sem
Suhm '97	Error Correction	ASR, HW, 2D G, K	GUI	E	sem
		Virtual Reality			
Pentland & Darrel '94	Virtual World	ASR, 3D GR	SS, FS	S	sem
Wang '95	Virtual Reality	ASR, ET, 3D C		E, R, S	sem
Chu et al. '97	VR based CAD	ASR, ET, 3D GR	GUI, A, H	C, E, S	sem
Flanagan '97	Collaboration	ASR, SV	SS, VC	S	n/a
		Security / Access Control			
M2VTS '96	Access Control	SV, FR		R	sem

Table 2.1 (cont.): Results from Survey of Multimodal Interfaces – Part I: Domain, Input/Output modalities, and Cooperation

Reference	Application, Domain	Input	Output	Coop.	Fusion
Integrated Talking Heads					
Beskow et al. '96	Interactive TV	ASR, 3D GR	SS, FS	S	n/a
Nitta et al. '97	Discussion System	FR , ASR WWWBrowser	FS, SS	n/a	n/a
Takebayashi '95	Food Ordering	ASR	SS, FS	S	n/a
Takeuchi & Naito '95	Games		SS, FS	S	n/a
Thórisson '97	Talking agent	ASR, GR, prosody	SS, FS	n/a	signal int sem
Applications for Special User Populations					
Anglade et al. '94	Telephone Switchboard	ASR, Braille-K	SS	E, S	sem
Beskow et al. '97	Communic. tool	ASR, GR	FS, SS	E, R	signal
Brooke & Scott '94		ASR, FR	Image, Audio	n/a	sem
Chen et al. '96	Rehab. Robot	ASR, OR, 3D GR		C	sem
Dufresne et al. '95	GUI for Blind	AF, HF	E	n/a	
Lavagetto & Lavagetto '96		GR	FS	C, T	int
Brooke & Scott '94		ASR, FR	Image, Audio	n/a	sem
Miscellaneous					
Beskow et al. '97	Consumer Information	ASR, K	FS, SS	n/a	n/a
Decarlo & Metaxas '96		FR	FS	S	sem
Doenges et al. '97		ASR, OCR, K	SS, FS	C, T	signal
Kamio et al. '94	Directory Assistance	ASR, touch	GUI, SS	S, E	sem
McGlashan '96	Product Info	ASR, K, GR	SS, FS	E, R, C, S	sem
Provine & Bruton '96	Video conferencing	FR	SS, FS	C, CC, T	sem
Stock '93	Information Retrieval	ASR, K, 2D GR		C	sem
Brondsted et al. '98	Campus Info	ASR, GR	SS	E, C, S	sem
Gauvain et al. '95	Info Retrieval	ASR, P	SS	C, S	sem

Table 2.2: Results from Survey of Multimodal Interfaces – Part II: Evaluation

Reference	SW Archit.	Criterion	Measure	Methodology
Multimodal Maps				
Nigay & Coutaz '93	PAC-Amod. PAC-Amod.	Time, User Satisfaction Modal. Usage	TCT Usage Frequ.	User Study Heuristic Evaluation
Vo & Wood '96	MMI	Cost Error Rate	TCT Sem. Accur.	Simul. Study benchmark
Oviatt & VanGent '96 Oviatt et al. '97	n/a	Modal. Usage complexity of interaction	Freq. Multi-/ unimodal Perplexity utterance length vocabulary size	Simul. Study
Martin '97	in-house	n/a	n/a	Iterative Design
Sarukkai & Hunter '97	in-house	Error Rate	City Accur.	benchmark
Data Input				
Leopold & Ambler '97	commercial	User Satisfaction		User Self-reports
Suhm '97	in-house	Cost Error Rate	Correc. Speed Correc. Accur.	User Study
Integrated Talking Heads				
Beskow et al. '96	in-house	quality of output	Intelligibility	User Study
Takebayashi '95	in-house	Cost, Error Rate	TCT	User Study
Takeuchi & Naito '95	n/a	Usefulness, Value		User Self-reports
Thórisson '97	in-house	error rate, cost quality of output	user subject	informal test user subject with prototype
Applications for Special User Populations				
Beskow et al. '97	TCL/TK, OpenGL	Wizard-of-Oz	Percep. correct response	post-exp.
Brooke & Scott '94	in-house	error rate, cost qual. of output		identification experiments
Dufresne et al. '95	n/a	Performance	% Compl. Tasks TCT	Simul. Study
Lavagetto & Lavagetto '96	in-house	quality of output	quantitative	user study with prototype
Yamamoto et al. '97	in-house	qual. of image error rate	quantitative intelligibility t.	benchmark

Table 2.2 (cont.): Results from Survey of Multimodal Interfaces – Part II: Evaluation

Reference	SW Archit.	Criterion	Measure	Methodol.
Miscellaneous				
Beskow '97	TCL/TK, OpenGL, Prolog	Wizard-of-Oz	Percep. correct response	post-exp.
Decarlo & Metaxas '96	OS Level Syst. calls	quality of output		
Kamio et al. '94	in-house	Cost	TCT	User Study
McGlashan '96	in-house			informal test
Provine & Bruton '96	JDK	cost		informal test

Table 2.3: Performance results of TDNN systems for speaker dependent (from Meier et al. 1997)

Test-set	Visual only	Acoustic only	Combined
clean	55%	98.4%	99.5%
16 dB SNR	55%	59.6%	73.4%
8 dB SNR	55%	36.2%	66.5%

Table 2.4: Results in word error (from Bregler and Konig 1994)

Task	Acoustic	Eigenlips
clean	11.0%	10.1%
20dB SNR	33.5%	28.9%
10dB SNR	56.1%	51.7%
15dB SNR crosstalk	67.3%	51.7%

Table 2.5: Prototype universal facial expressions of emotions and their corresponding FACS action units

Emotion	Action Units
anger	AU2 + AU4 + AU5 + AU10 + AU20 + AU24
disgust	AU4 + AU9 + AU10 + AU17
embarrassment	AU12 + AU24 + AU51 + AU54 + AU64
fear	AU1 + AU2 + AU4 + AU5 + AU7 + AU15 + AU20 + AU25
happiness	AU6 + AU11 + AU12 + AU25
sadness	AU1 + AU4 + AU7 + AU15
surprise	AU1 + AU2 + AU5 + AU26 + rotate-jaw

Table 2.6: List of AUs

AU	Name	AU	Name
AU1	Inner Brow Raiser	AU31	Jaw Clencher
AU2	Outer Brow Raiser	AU32	Lip Bite
AU4	Brow Lowerer	AU33	Cheek Blow
AU5	Upper Lid Raiser	AU34	Cheek Puff
AU6	Cheek Raiser & Lid Compressor	AU35	Cheek Suck
AU7	Lid Tightener	AU36	Tongue Bulge
AU8	Lips Toward Each Other	AU37	Lip Wipe
AU9	Nose Wrinkler	AU38	Nostril Dilator
AU10	Upper Lip Raiser	AU39	Nostril Compressor
AU11	Nasolabial Furrow Deepener	AU41	Lip Droop
AU12	Lip Corner Puller	AU42	Slit
AU13	Sharp Lip Puller	AU43	Eyes Closed
AU14	Dimpler	AU44	Squint
AU15	Lip Corner Depressor	AU45	Blink
AU16	Lower Lip Depressor	AU46	Wink
AU17	Chin Raiser	AU51	Head Turn Left
AU18	Lip Pucker	AU52	Head Turn Right
AU19	Tongue Show	AU53	Head Up
AU20	Lip Stretcher	AU54	Head Down
AU21	Neck Tightener	AU55	Head Tilt Left
AU22	Lip Funneler	AU56	Head Tilt Right
AU23	Lip Tightener	AU57	Head Forward
AU24	Lip Presser	AU58	Head Back
AU25	Lips Part	AU61	Eyes Turn Left
AU26	Jaw Drop	AU62	Eyes Turn Right
AU27	Mouth Stretch	AU63	Eyes Up
AU28	Lip Suck	AU64	Eyes Down
AU29	Jaw Thrust	AU65	Wall-eye
AU30	Jaw Sideways	AU66	Cross-eye

3 Consumer off-the-shelf (COTS) product and service evaluation

3.1 Introduction

3.1.1 Purpose and scope of this chapter

This chapter deals with the assessment of speech related services and systems. It is meant for the reader who wants to conduct a comparative evaluation of systems that can be bought 'off-the-shelf', or services that can be used directly, such as voice dialing systems or travel information systems. The chapter is not written for technical specialists who want, for example, to choose a speech recognition system that will be embedded in a telephone switching system. These readers are referred to the detailed evaluation chapters of the *EAGLES Handbook of Standards and Resources for Spoken Language Systems* (Gibbon et al. 1997).
The products and services covered in this chapter are classified into three main categories:

COMMAND AND CONTROL SYSTEMS These systems contain an automatic speech recognition (ASR) system as an interface for controlling the environment of the user. The systems can be as simple as the graphical shell of the user's computer or as complicated as control software for all operational functions of a fast fighter aircraft.

DOCUMENT GENERATION These systems employ an ASR system in order to support the fast and flexible generation of documents, forms and reports. A simple application might be a dedicated system for filling out simple forms, or for data entry. More complicated systems allow full dictation of free text into a word processor of the user's choice, using continuous speech and permitting control of all the features of the word processor.

SERVICES AND TELEPHONE APPLICATIONS These systems generally require more speech technologies than just ASR alone. Usually speech synthesis is necessary for feedback, and sometimes speaker verification is required. Such systems also often contain some kind of dialogue control component. Services include information kiosks which can interpret spoken commands, automated call centers, and voice dialing systems in telephone exchanges and travel information systems.

A number of important product or system types are not covered in this chapter, for example language learning tutorial systems and audio indexing software.
Table 3.1 contains a classification of the main kinds of current speech technology products and services.

3.1.2 Introduction to speech technologies and classification

Another way of looking at product classification is to examine the speech technologies which are used in the products. Combinations of various technologies allow many different applications to be designed. The technologies themselves are more stable over time than specific applications: typically a new technology emerges every five years, while new applications or new application versions are

Table 3.1: Categorisation of some products and services into the categories described in Section 3.1.1.

System	Command& Control	Document Generation	Services & Telephone	Other
PC Dictation system		×		
Video recorder control	×			
Language learning				×
Spoken document retrieval				×
Voice Dialing (telephone)	×			
Voice Dialing (telephone exchange)			×	
Wheel chair control	×			
Radiology report dictation		×		
Transcription service		×	×	
Information kiosk			×	
Travel information			×	

released every few months. We will confine ourselves to the description of some of the major technologies in the field.

3.1.3 Automatic speech recognition

Automatic speech recognition (ASR) is the main technology discussed in this chapter because it is included in almost every product or service which uses speech technologies and underlies the front ends of systems that use spoken language input. The capabilities of a speech recognition subsystem can vary along many dimensions; cf. (see Gibbon et al. 1997), Chapter 10, on ASR system assessment. Here we restrict ourselves to an overview of the main characteristics of ASR systems.

SPEAKER DEPENDENCE An ASR system can be speaker independent, speaker adaptive, or speaker dependent. A speaker dependent ASR system needs to be trained for the user the system has been designed for. A speaker independent system is trained in the factory, and can therefore be used directly after unpacking. Recognition performance of a speaker independent system is generally lower than that of a comparable speaker dependent system. A speaker adaptive system starts out as a speaker independent system but gradually changes its speech models such that the system adapts to a specific user. Performance (after adaptations) is typically that of speaker dependent systems. There are also word recognition systems which use mixed speaker dependent and speaker independent models.

SPEECH CONTINUITY An ASR system can deal with isolated words, connected words, or continuous speech. An *isolated word* recognition system can only recognise speech units (words or fixed expressions) that are separated by (possibly tiny) pauses. A *connected word* recognition system still uses isolated words as speech models, but is capable of recognising these words when they are connected in running speech. A *continuous speech* recognition system can recognise running speech, and is also trained (possibly in the factory) with continuous speech. Some systems are hybrid; they are basically isolated word recognition systems

but can cope, for instance, with continuous digit strings.

> *Recommendation 1*
> Be aware that there are ASR system manufacturers who claim that their product can deal with 'continuous speech', while in fact the systems are isolated word recognisers, which still need tiny pauses, or which can only deal with limited vocabularies (e.g. digits) in continuous speech.

VOCABULARY SIZE Vocabulary size (coverage) is defined as the number of words a recogniser can handle. There are further specifications in terms of the *active* vocabulary size, i.e. the maximum number of words the system can recognise at any given moment, the *passive* vocabulary size, i.e. the number of words the system has in store to be loaded into the active vocabulary, and the *exception/user/extension* vocabulary size, i.e. the number of words a user can add himself.

Finally, we need to draw the reader's attention to the existence of various speaking styles, which greatly affect the performance of ASR systems. For the purpose of this chapter, we define the following general speaking styles:

READ SPEECH This speaking style is that of a radio or television news reader, somebody giving a prepared lecture (or perhaps delivering a somewhat unimaginative paper at a conference). Although this speaking style hardly ever occurs in everyday life, many dictation systems are trained on this type of material. The style is characterised by well formulated sentences, very few hesitations and intonation which is more or less predictable on the basis of the text alone.

SPONTANEOUS SPEECH This is the most representative talking style. In everyday life people generally communicate by talking spontaneously to each other. For an ASR system, this talking style is particularly difficult to handle. The style is characterised by large variations in volume (level), the use of unpredictable intonation, hesitations, errors in vocabulary, corrections, re-starts and incomplete or otherwise grammatically incorrect sentences.

DICTATION SPEECH For the purpose of clarity of presentation in this chapter we have coined the term 'dictation style speech' (Hunt n.d.) to refer to the way a skilled user of an automatic dictation system speaks to the computer. The intonation is similar to that of read speech, but the grammatical constructions and error corrections are more like those of spontaneous speech.

Finally, there is the issue of accent and dialect. Recognition systems are usually trained only for a very limited number of speaker accents. In practially all language communities, dialects differ in so many ways from the officially recognised standard language that a separate recogniser would be necessary to cope with them.

3.1.4 Text-to-speech and speech synthesis

In some respects, text-to-speech synthesis is the opposite of automatic speech recognition: given a machine-readable text, the system will read the text aloud, rather than producing a machine-readable text from speech as in ASR. However, formally, speech synthesis in the strict sense is only the last link in the complex chain of procedures required for converting text into speech, and involves the generation of the actual sounds that make up speech. Text-to-speech covers

also a number of other procedures, including the expansion of text (numbers, abbreviations, etc.), grapheme to phoneme conversion, the modelling of unit durations and rhythm, the definition of stress and focus, and the generation of pitch contours (intonation). These procedures involve analysis of the written text before the actual expansion, generation and synthesis procedures apply.
There are several techniques used in the field of speech synthesis:

PLAYBACK The simplest technique is synthesis by playback of pre-recorded words or phrases ('canned speech'). This generally provides good voice quality but low flexibility. This technique provides no way of adapting the intonation or the voice properties; this must be implemented by pre-recording all possible voices, intonations and vocabulary items (words and phrases). The vocabulary is limited by the recordings made. Sometimes longer units are constructed, as when a string of digits is merged into a standard carrier sentence; this provides some flexibility.

CONCATENATION By playing back sub-word units of pre-recorded speech contiguously, whole words and phrases can be synthesised. In general the units chosen are diphones, i.e. the interval composed of the last half of the previous phone and the first half of the next phone. Usually the voice quality of these systems is high. By using algorithms such as PSOLA the pitch and temporal properties of the pre-recorded waveforms can be changed, enabling the superimposition of controlled intonation and accentuation patterns. However, a genuine change of voice characteristics is not possible. The vocabulary is limited by the applicability of pronunciation rules.

PRODUCTION MODEL By using a physical model of the vocal folds and the vocal tract it is possible to produce sounds that resemble speech more or less closely. These models are often LPC (linear prediction coefficient) based, i.e. they define sounds by the position and width of formants in the signal spectrum. In practice, the voice quality of production model synthesis, also known as formant synthesis or parametric synthesis, is not as good as the playback and diphone concatenation techniques, but this technique constitutes what might be called 'pure' synthesis, in which in principle every parameter is controlled. Using this technique, a wide range of voice characteristics, in particular prosodic features such as pitch and intonation,, or accent (in the sense of or stress) can be exploited flexibly by the system. Again, the vocabulary is limited by the applicability of pronunciation rules.

When pronunciation rules are used, usually exceptions are made for proper names. Often, they are taken from name pronunciation databases.
Evaluation of speech synthesis systems of these types is treated in depth in Gibbon et al. (1997), Chapter 12. In this chapter we touch on this subject briefly with a number of specific examples.
A step further than text-to-speech (TTS) architecture is *concept-to-speech* (CTS) architecture. This is generally what is needed in information retrieval systems. In this case, the 'concept' is the piece of information requested by the user, and it is mapped into speech in order to be conveyed to the user. At the present state of technology, this is carried out through a concept-to-text generation component, followed by a text-to-speech system. In future systems, the text stage will be bypassed, since it introduces artefacts due to irregularities in punctuation and in grapheme-to-phoneme conversion.

3.1.5 Speaker recognition and verification

A speaker recognition system uses characteristics of the voice in a speech signal in order to identify the speaker. This is a special type of 'meta-information' that is conveyed by speech. Other related forms of meta-information are the specific characteristics of individual languages, dialects, speaker mood, and the health of the speaker. There are two main kinds of task associated with speaker recognition:

SPEAKER IDENTIFICATION Here the task is to identify an unknown speaker as one of a closed set of known possible speakers. The typical implementation is carried out by comparing the test utterance with recordings of all known speakers, and choosing the speaker that fits best.

SPEAKER VERIFICATION Here the task is to decide whether a test speaker is the speaker he claims to be. The claimed identity is known, and a typical implementation accepts the speaker if his speech matches recordings of the claimed identity closely enough.

There are two types of information available to speaker recognition:

TEXT DEPENDENT Here, the content of the utterance is known from other sources, which makes it possible to carry out a detailed comparison of the words or sub-word units in the speech signal to aid speaker recognition.

TEXT INDEPENDENT In this case it is not known beforehand what words have been said. The speaker recognition systems must either use the information globally or first perform speaker independent speech recognition.

This classification is very much like the distinction between speaker dependent and speaker independent ASR systems. Theoretically the distinction is independent of the identification/verification task, but in practice combinations of text dependent speaker verification and of text independent speaker identification are frequently encountered. More about the classification and the evaluation of speaker verification systems can be found in Gibbon et al. (1997), Chapter 11.

3.1.6 Speech understanding

A *speech understanding system* goes one step further than a speech recognition system. Not only is the speech recognised, but the words are also interpreted in terms of their meaning. One could perhaps call these systems 'speech-to-concept' as opposed to 'concept-to-speech,' by analogy with the terminology of speech synthesis. In dialogue systems that ask open questions (of the "What do you want?" or "How can I help you?" type), *speech understanding* plays an important role. If the questions are closed and binary ("Do you want information about trains?") or specific ("Where do you want to go by train?"), the system relies on *speech recognition* to a greater extent or even on *word spotting*, the identification of individual words in an utterance. Generally, however, any speech interactive system that reacts to reasonably complex spoken input 'sensibly' could be called a speech understanding system.
A large family of techniques involved in speech understanding is covered by the discipline of Natural Language Processing (NLP), which is mainly concerned written language processing; of particular importance is the parsing of written

language into semantic representations or 'concepts'. The evaluation of this specific aspect is beyond the scope of this chapter, and specialised literature on the subject should be referred to (see for example Galliers and Sparck Jones (1996)).
We mention the technique here because it can be part of a dialogue system that is implemented in a service or telephone application.

3.1.7 Dialogue control

Dialogue control is necessary in order to provide a fully automated information service. The dialogue control component is responsible for the interaction between the user and the service. It must not only handle events triggered by the user, but must also trigger the user to provide the system with information, as well as sending requests to the information retrieval engines and providing input to the text-to-speech engines.
Further general information about the evaluation of dialogue systems can be found in Gibbon et al. (1997), Chapter 13.

3.2 General remarks

In this section we will discuss matters concerning all types of products and services. In Gibbon et al. (1997), Chapter 9, more detailed information on evaluation test design can be found.

3.2.1 Assessment methodology

There are different types of spoken language system assessment. The main ones are *diagnostic* versus *comparative* assessment. Diagnostic assessment involves setting up a framework for testing the product with the aim of giving feedback to the developer in order to improve the system. Comparative or benchmarking assessment is used to select the best available system, or to publish an article in a consumer magazine, or just to determine the state of the art of the technology. Different reasons for assessing spoken language systems will in general lead to the selection of different methodologies. For assessing speech products and services we distinguish two main methodologies: one using the judgments and reactions of test subjects, and one using pre-recorded speech in a semi-automatic procedure. These are known as 'subjective' and 'objective' test methodologies, the terms should obviously not be over-interpreted.

3.2.1.1 Subjective test methods

In a subjective assessment the test is designed in such a way that human test subjects interact with the system. The evaluation measures can be objective, such as the percentage of successful task completions, the time taken to complete the task, or the number of interactions necessary per task. Subjective measures also exist, such as level of intelligibility, general impression, annoyance, user friendliness, intuitiveness, level of difficulty, the subjective impression of system response time, etc.
Subjective tests are often used for the purpose of comparative assessment or for benchmarking. Important design issues for subjective tests are the number of subjects, their gender, and the order in which they perform various tests.

3.2.1.1.1 Number of subjects

The number of subjects to be used is a very important parameter in evaluation design. For many psychological tests, there is enormous variation in subject performance. Speech is inherently variable by nature along many dimensions, and there exist marked differences in speech between different speakers and listeners. This means that for an individual subject one dictation system may perform better than another while for a second subject the results may be reversed.
The decision on how many subjects should be used should ideally come from a *power analysis* (Cohen 1988). First, a decision must be taken on how large an effect should be in order to be interesting. For instance, it might be possible to prove that for one recogniser which scores 78 % accuracy is significantly lower than another that scores 80 % by using 200 test subjects. However, the small difference in performance may not be at all interesting in practice because there are other factors such as ease of use, error recovery limitations, reaction time, etc. which are more important for the quality of the products in their intended applications.

> *Recommendation 2*
> Decide on the minimum difference in performance which can be considered an interesting difference.
>
> Choose a level of significance, e.g. $p < 0.05$.

Given a minimum interesting difference in performance, one should estimate the *mean* and *variance* in performance. These estimates can be based on earlier tests of similar products or be the outcome of a small pilot test. With these estimates, one can find the minimum number of subjects which is needed in order to show a significant difference in performance between systems. This can be found in any good introductory statistics book; cf. Gibbon et al. (1997) for further discussion and references.

> *Recommendation 3*
> If possible, make an estimation of the mean and variance of the performance measure, and base the number of subjects needed on these estimates and the minimum interesting performance difference.

This procedure may well seem too much effort for many consumer evaluation situations. It is difficult, however, to give rules of thumb for the number of subjects needed for a test. We would give the following advice:

- use a *minimum* number of four subjects,
- balance them evenly in gender, and
- balance them evenly in test ordering.

With four subjects, a minimum indication of variance between subjects can be estimated.

3.2.1.1.2 Gender of the subjects

If the systems involve speech input (recognition), a crucial factor is the gender of the subject. This is because male and female voices differ in spectral shape and content, i.e. there is usually a considerable difference in both fundamental frequency and formant positions between male and femle voices.
Unless the system to be evaluated is to be used solely by speakers of a single gender (no doubt it would be 'politically incorrect' to give examples of such uses, though many come to mind), it is very important to use both male and female subjects for evaluation. It is best to use the same number of males and females, even if in actual practice the gender ratio may be different from 1. In this case, post-scaling of the results for male and female speech based on the gender ratio should be performed.

Recommendation 4
Try to use at least 4 test subjects, 2 male and 2 female.

3.2.1.1.3 Order of conditions among subjects

A very important psychological effect is the *learning effect*. This is the effect that a first test for a naive subject is always harder than later tests, because the subject learns during the first test how to deal with the system, what kind of events to expect, and so on. Often therefore, either the first few results are not used at all in scoring, or an explicit training/adaptation period is taken into account. As a result of the learning effect, the order in which two systems are assessed with the same subject is important. For the first system, the subject may find the whole concept of the service difficult and may therefore react slowly or inconsistently, while by the time the second service is tested the subject may have become accustomed to this kind of service.
For this reason, it is important to balance the order among subjects. Balancing means that there are just as many subjects involved in assessing systems in one order as there are subjects that assess them in the opposite order (or other orders).

Recommendation 5
Try to balance the testing order of the systems among the subjects.

As an example, we combine gender and system order, so that we can divide 4 subjects in the following way in order to compare two different systems: In this

Order	male	female
System A, system B	Subject 1	Subject 2
System B, system A	Subject 3	Subject 4

way both male and female subjects are involved in both orders, averaging out any gender-dependent learning effects.
If more than two systems are compared, the orders should be such that systems occur in all possible positions the same number of times. It is customary to use a multiple of the number of systems as the number of subjects. In this way a

'Latin square' design of system assessment order can be used, an example for four systems is given in Table 3.2.

Table 3.2: The order of systems for different subjects. The numbers indicate the system number, time runs left to right.

Subject	order of system			
Subject 1	1	2	3	4
Subject 2	2	1	4	3
Subject 3	3	4	1	2
Subject 4	4	3	2	1

Other conditions should also be balanced among the subjects. For instance, a recognition service might be assessed in different noise conditions, such as a quiet room as opposed to a noisy office. Again, the order of the conditions 'quiet room' and 'noisy office' should be mixed with the conditions 'system A' and 'system B.' In such an orthogonal design one of the conditions forms the 'outer loop' and the other the 'inner loop.' Which parameter is selected for the outer loop depends on the effort needed to change the condition. In this case it might take considerably more time to change the system than the environmental condition, so system order would be placed in the outer loop of the design, as in Table 3.3.

Table 3.3: Test order for a combination of two systems under two test conditions. The numbers indicate the order, e.g. the number 2 indicates that that (system,condition) is tested second for the test subject.

Subject	System A		System B	
	quiet	noise	quiet	noise
Subject 1	1	2	3	4
Subject 2	2	1	4	3
Subject 3	3	4	1	2
Subject 4	4	3	2	1

We see from these balancing arguments that often the number of test subjects chosen is a multiple of the product of the dimensions of all variables. So if there are 2 genders, 3 systems, 3 noise conditions and 5 telephone sets, a full design would use $2 \times 3 \times 3 \times 5 = 90$ orders and thus 90 subjects in order to make sure that all order effects balance out. This example has a rather extreme number of variables, so in order to reduce the number of subjects needed one of the variables could be sacrificed, for instance the order of the telephones used (for which it may be unimportant whether there is a learning effect or not). Another approach to dealing with learning effects is to *randomise* the order. This approach can be used if there is no reason to assume that two variables influence each other, for instance telephone and noise condition.

3.2.1.2 Objective test methods

In an objective test the role of test subjects is reduced, and their behaviour is 'controlled'. One might still need human stimuli such as speech signals in order to feed information to the system, but they will be recorded on digital media, and can be used repeatedly in order to test different systems under identical conditions. Human interaction might also be necessary, for instance in the reaction to the dialogue conducted between a service and the user. In this case, the interaction is controlled by a skilled and experienced tester.
The advantages of objective methods over subjective methods are that the stimuli are controlled, and a test can be repeated reliably under different conditions and at different times. The disadvantages are that the test is not really representative and that human-machine interaction is not tested.
For an objective test it is tempting to automate the evaluation setup into a testbed. However, the efforts needed to do this are generally quite high, especially in making the testbed error free and robust against small changes of the system that is being tested. For small tests or tests that occur only seldom, our advice is to perform the test manually.

> *Recommendation 6*
> Before automation of an evaluation test setup into a testbed, assure yourself that the investment effort will be returned.

3.2.2 Subjective assessment measures

It was mentioned above that in addition to objective measures of the performance of systems there are also subjective measures. These can be very important for the global evaluation of a service or product, because in the end a human being has to use the system and if it is annoying or impractical it is likely that the system will be neither bought nor used.
Normally, subjective measures are obtained by asking test subjects to give their judgments on several system properties. This can be done either at regular intervals during the experiment, or at the end. The information is often given by asking the test subject to fill out a questionnaire. In this case, there should be a time slice in the test protocol during which this can take place. It is also possible that the tester may ask the subjects the questions, and fill out the forms or enters the data directly into a computer database. If there are too many questions, or if the questions occur too often, the subject can get bored or irritated, resulting in inaccurate answers.

> *Recommendation 7*
> If a questionnaire is to be filled out, reserve a spot in the test protocol where there is no intrusion in the test itself. Also, the test subject should not be put under time stress.

> *Recommendation 8*
> Do not ask the subjects to answer more questions than necessary. Do not ask the same set of questions more often than necessary.

In many cases the subjective measure is expressed in terms of a numbered scale. For example the subject can be asked if he/she was annoyed by the system at some point. The usual measure for this is a 5-point scale, as indicated in Table 3.4. In experimental psychology it is very common to use a 5-point scale, although sometimes a 7-point scale is used; experience has shown that five different levels give a good level of consistency among subjects. An odd number of levels allows the subject to give a 'neutral' answer by choosing the central option.

Recommendation 9
For subjective measures, use a five-point scale for the answer.

Table 3.4: Examples of the use of a five-point scale

Scale	Annoyance	Quality	Sound Level
1	not annoyed	bad	too soft
2	slightly annoyed	poor	soft
3	fairly annoyed	fair	good
4	annoyed	good	loud
5	very annoyed	excellent	too loud

3.2.3 Acoustic environment

3.2.3.1 Noise

For all speech related products and services it is very important to make an inventory of the acoustic environment in which the system is going to be used. This environment influences both the speech input and the speech output side of the system under assessment. When there is speech input, recognition and all further steps (understanding, dialogue, information retrieval) will typically suffer from environmental noise or distortions introduced by the transmission channel. Similarly, for speech output, intelligibility (and hence usability) will decrease if the signal-to-noise ratio (SNR) is low or if the acoustic environment is very reverberant.

Recommendation 10
Make an inventory of the properties of the acoustic environment in which the product or service is typically used.

During the test, an acoustic environment should be realised which is comparable to the real life situation. The noise level and the acoustic spectrum of the noise should be the same, or span a similar range. For instance, a car radio with speech recognition may be used in several cars driving at several speeds. If this system is going to be evaluated in the laboratory, a representative set of 'car noises' should be generated in an environment which is acoustically similar to a car, e.g. diffuse sound field, no reverbarations, etc.)

For very specific use — such as the car radio — the noise spectrum should be similar within 5 dB in third octave bands. For less specific circumstances, e.g. dictation in an office environment, the noise spectrum is less strict. For noise levels, it is usually sufficient to test under the highest noise level that can normally occur. Sometimes, however, fluctuations of noise can influence the results.
In cases where microphone positioning is an issue, it is advisable to use either test subjects in representative positions or a head and torso simulator ('artificial mouth'). Examples of such circumstances are an information kiosk in a train station (noise from people and trains, reverberation in the kiosk), or a car radio. When test subjects are used, humans react by 'speaking up' (increasing their vocal effort) when there is noise. This is called the *Lombard effect*. Not only the level, but also a change in speaking style can be observed, which means that it is definitely not sufficient simply to add noise to pre-recorded speech. Consequently, even when pre-recorded utterances are used for testing the speech input system, these should be recorded by putting the speaker in a similar noisy environment, for instance by applying noise at a calibrated level over a headset. The sound level, in such case, may never be higher than 80 dB(A).

Recommendation 11
Be aware that for noise conditions higher than 60 dB(A), the Lombard effect may change the level and intonation of a speaker.

Recommendation 12
Do not expose test subjects to ambient noise levels higher than 80 dB(A), because otherwise a permanent hearing loss could result from the experiment. For experiments in higher levels, be sure to consult a specialist on human hearing first, and have a proposal of the experiment checked by your local ethical committee.

3.2.3.2 Microphone electrical input

Many dictation and command and control systems are based on personal computer programs and a standard sound card. Because PCs are in many cases built with low-budget hardware, the sampling quality of the microphone signal is very low and the internal impedance high. The reason for this is that the voltages generated by a microphone are quite low and inside a PC crosstalk, i.e. inter-circuit interference, from high frequency digital circuits is therefore quite likely to occur.
Although representative conditions would require testing with a selection of available sound cards, the assessment is usually set up not in order to test computer hardware but rather because of the recognition software. Therefore, we would advise using a high quality microphone amplifier with the *line input* of the sound card, not the *microphone input*. For most sound cards, the line level input can be sampled with the desired accuracy.

Recommendation 13
For PC-based systems, check that the microphone is recording without distortion. If necessary, use a separate microphone amplifier with the line input.

Recommendation 14
Make sure that the electrical input signals do not overload the system.

3.2.4 Comparing several systems

If systems are assessed in order to be compared, we recommend making a list of capabilities for each product. Only the capabilities that all systems have in common can be compared quantitatively at a later stage. Nevertheless, the overall assessment of a product should be based on all capabilities, not only on those that a product has in common with other products.

Recommendation 15
Make an inventory of all the capabilities for all the systems that are evaluated. In comparative testing, only the common subset of capabilities can be compared quantitatively.

The detailed test design should therefore only test items that all systems have in common. This may reduce the number of measures drastically. If a certain capability is implemented on all but a minority of the systems under evaluation it may be interesting to include this item in the test anyway.
Capabilities that are unique to certain products may make such products stand out from the average. Therefore these capabilities must be explored and described in the test report. In the overall assessment extra capabilities could make all the difference in choosing between two systems that score more or less the same on their common capabilities.

Recommendation 16
Unique capabilities may be very important to the general quality of a product or service.

3.3 Command and control systems

3.3.1 Typical systems

A command and control system is a system that controls operation in a certain work environment. Often in situations where a person has to use both hands for carrying out his job (a 'hands busy task') or in adverse environments in which manual control is not possible, it is desirable to be able to give commands to the operating environment by means of speech commands. The spoken input interface adds an extra input mode to switches, buttons, levers, and knobs.
In many cases the environment is adapted in such a way that all control switches are electronic switches rather than switches that physically change state. This makes it possible to control the switches electronically, for instance by computer, in which case case control by speech input is also possible.
Since the introduction of graphical user interface (GUI) environments to computer systems, more and more work environments have become centred on the computer monitor, and voice input has become a third input modality next to keyboard, mouse and graphics pad. Since many ASR products run on a

personal computer or workstation, the integration of speech into control procedures for graphical environments is an obvious step. A particularly important motivation for using speech interfaces concerns input/output facilities for the physically handicapped, such as people with speech production problems (for speech output), the blind (for speech output), or the manually impaired (for speech input).

3.3.1.1 Managing computer systems by spoken commands

When PC and workstation applications are managed via speech interfaces, this usually involves menu control, filling in forms, checking checkboxes and similar events; in this context, speech control can be seen as an extension of keyboard and mouse input. There are two approaches to making an application 'speech aware.'

LINKED INTO THE APPLICATION Here, the developer of the application had speech input in mind from the start, and the ASR functions are explicitly linked into the application. For the developer the advantage is that the vocabulary is usually known, and the actions taken after a spoken command has been recognised are determined by the programmer. For the user, this has the consequence that it may not always be clear when the speech recogniser is in operation, and what it can recognise.

CONTROLLED THOUGH AN EXTERNAL APPLICATION In this case, the speech recognition system is a separate application ('voice manager') that is able to generate mouse and keyboard events for the target application. The target application is not aware that there is a speech recogniser controlling its inputs, so the level of integration is generally lower than in the case of the linked-in recogniser. However, the flexibility of a voice manager system is quite high, as it can be used to control any application as long as the graphical enviroment programming standards are adhered to.

3.3.1.2 Consumer electronics

By *speech aware consumer electronics* we understand mobile telephones, video recorders, TVs, car radios, and similar devices, which can be controlled by speech input. At the time of writing, most systems of this type are still under development and only a few are available on the market. In these cases the recognition system is an 'embedded system', i.e. it is integrated into the other functions of the product, which means that it is difficult to separate the functionality of the recogniser from that of the product itself.

3.3.1.3 Professional embedded applications for hands-busy operation

This category is similar to the consumer electronics category, except that for professional systems integration is often carried out especially for the user. An example would be a meat factory, where workers need to use their hands for meat-checking and use spoken commands in order to control processing devices. Other application domains include medical operations, where control over the positioning of the patient's bed can be controlled by simple voice commands, or a jet fighter cockpit where all non-critical flight operations can be voice commands. The main aim implementing voice-control is that spoken commands increase efficiency.

Because often the ASR system is specifically designed and implemented for a particular situation it is sometimes possible to capture the output of the recogniser directly, which can make assessment easier.

3.3.2 Typical issues

There are some issues which are particularly characteristic of the class of command and control systems, which will be discussed in this section.

3.3.2.1 Performance measures

The performance measures which can be used for evaluation are:[1]

RECOGNITION ACCURACY The word recognition accuracy for a word recognition system is defined as the number of correctly recognised words divided by the number of words in the test (see Gibbon et al. 1997, Chapter 10). If the number of words in the test is N, and the number of missed (deleted) words d, the number of inserted words i, and the number of substituted words s, the word accuracy a of the system is defined as

$$a = 1 - \frac{s + i + d}{N} = 1 - w. \tag{3.1}$$

Word accuracy is often expressed as a percentage. From an academic point of view, the word error rate w is a better measure. This is just the complement of the accuracy, i.e. one (or 100 %) minus the accuracy. If the word error rate w is known, an estimate of the standard deviation s_w can be found from the number of words in the test, N (van Leeuwen and Steeneken 1997):

$$s_w = \sqrt{\frac{w(1-w)}{N}}. \tag{3.2}$$

The standard deviation for the word error rate is the same as for the accuracy.

OOV-REJECTION An out-of-vocabulary word (OOV word) is a word that is spoken by a user that cannot be recognised by the system, because it is not in the system's active vocabulary. It might, for example, not be intended for the system but for a colleague of the user.

ERROR RECOVERY Both the system and the user are bound to make errors once in a while. A good system allows the user to undo actions triggered by previous spoken commands.

[1] Note also the extended terminology, in particular the terms *precision* and *recall*, *false negatives* (misses) and *false positives*, used in the related case where *sets* of competing recognition outputs are considered rather than just one. This typically occurs in an experimental system development situation (but see below, evaluation of services). – Ed.

A	= set of reference events
B	= set of event hypotheses
CP	= set of correct positive event hypotheses, $A \cap B$
FP	= set of false positive event hypotheses, $B - CP$
FN	= set of false negative event hypotheses, $A - CP$
The following measures are defined, often as percentages:	
Recall	= $\lvert CP \rvert$ / $\lvert A \rvert$
Precision	= $\lvert CP \rvert$ / $\lvert B \rvert$
R&P	= $2 \times Precision \times Recall$ / $(Precision + Recall)$
	= $2 \times \lvert CP \rvert$ / $(\lvert A \rvert + \lvert B \rvert)$

RESPONSE TIME Important for usability is the time it takes to respond to a spoken command, i.e. system reaction time. It is defined as the time from the end of the command utterance to the start of the action. Both the average time and the distribution of the response time are important parameters.

SITUATIONAL AWARENESS Users that give commands to a system have certain expectations about what they can say. This might depend on the internal state of the system ('active vocabulary'), but if the user is not aware of that state, for whatever reason, it is said that he has lost his *situational awareness*. This measure is difficult to quantify because it involves essentially subjective impressions of both the test subject and the experimenter.

3.3.2.2 Speech recognition parameters

It is essential to know the underlying technology of a command and control system. Core parameters are:

ISOLATED/CONNECTED/CONTINUOUS SPEECH Some speech recognition manufacturers tend to misclassify the technology of the product. An isolated word speech recogniser is often used for command and control. This is often a good idea, because the system can then take an action after every word. Connected word and continuous speech recognition systems allow for complex commands to be given as a single utterance. It is very unlikely that these systems react before the end of the utterance is detected.

SPEAKER DEPENDENCE A speaker dependent system typically has some kind of training phase. For command and control, this usually means that the list of all possible commands must be trained several times (up to 3–5 times per word is normal). A speaker independent system should be able to do without any training, but it may still need gender information about the user and it may need to adapt to the microphone and speech level. A speaker adaptive system may include an enrolment procedure in a different part of the product, for instance in an accompanying dictation system.

VOCABULARY DESIGN There are several ways of handling the vocabulary. The vocabulary can be restricted to all possible commands in the application, as is likely to be the case with *linked applications* and embedded systems. The vocabulary may also be dynamic, meaning that the words are dynamically read from the application. In this case, it might be necessary to train words explicitly for which the recogniser does not have pronunciation information.

Recommendation 17
Try to identify the underlying technology of the speech recognition system. It can be of importance to the design of the test. Also, verify that all words in the application are in the vocabulary, i.e. can be recognised.

3.3.2.3 Level of integration in environment

For some performance measures such as recognition accuracy it may be beneficial if the output of the recogniser can be captured directly. This may have to be done through analysing a communication line, by looking at a screen, or by analysing a log file. Usually the rule applies that the more the recognition system is integrated into the environment, the less likely it is that the direct

recognition output can be seen. For embedded consumer products the level of integration is high, as is also the case for computer applications linked to a speech recogniser.
A high level of integration may also limit control over the various parameters. For instance, a complete reset of all acoustic models with retraining might not be implemented, in which case a structured assessment of different conditions might be difficult.

3.3.2.4 Types of feedback

Very important is the type of feedback that a command and control system gives. Feedback is important both for the situational awareness of the user, for error recovery, and for the evaluation of the accuracy of the recognition system. Forms of feedback include:

ACOUSTIC SIGNALS A beep can be used as an acoustic feedback signal that an utterance has been recognised. Differences in pitch and pitch pattern can code whether or not the recognition was successful.

SYNTHESISED SPEECH A system that has no visual feedback mechanism might repeat the command which was recognised ('speak back').

GRAPHICAL A graphical indication can show up at the users terminal, or the word that has been recognised can be displayed on the screen ('read back').

BY ACTION TAKEN Feedback might be completely left out of the system; in a design of this type the recognised word can only be deduced indirectly from the action taken.

Some systems give a small acoustic signal (beep) to prompt the user to speak, although this is not usual for command and control systems. Such behaviour should not be confusable with signals that indicate reception of a command.

> *Recommendation 18*
> Analyse the type of feedback the system gives after recognising a command word or string.

3.3.3 Evaluation design

When an inventory of the typical issues involved with the command and control system has been made, the design of the evaluation can be carried out. In this section general aspects of evaluation design will be treated; some specific examples will be discussed in the following section.
First, a decision should be made on the type of evaluation, either subjective or objective. Then the performance measures have to be chosen.

3.3.3.1 Performance measures

Individual performance measures in the test need to be separated. It is better to have a separate section in the test protocol that evaluates OOV rejection,, for example, than to deduce this from other test results.

> *Recommendation 19*
> Separate the individual performance measures in the test.

The various performance measures need their own methodology

Accuracy In what way can it be deduced which word has been recognised:

- Textual visual feedback
- Textual logging to a computer file or communication port
- Graphical visual feedback
- Audio feedback
- Feedback by action

Only some of these feedback types are usable for automatic recognition scoring, and consequently the experimenter frequently has to evaluate scores for individual words manually.

Error recovery Make a log of

- the severity of recognition/errors in all situations
- procedures for 'undoing' misrecognition

A special test for 'high risk' commands ('Yes'/'No,' 'OK'/'Cancel' etc) needs to be included. Because the allowed error rate for such words is low, many instances are needed in the total test in order to measure such a low error rate. Alternatively, one might decide to test these words for robustness under harder conditions by adding noise.

OOV-rejection Note

- the probability that a user might not speak to the system in normal use.
- the provision of 'sleep'/'wake up' commands in system, push-to-talk switch, or microphone selection tools.

A test should be included for the rejection of OOV-words, based on the inventory made, or for frequently used words that are not always active.

Feedback Analyse the feedback, if any given, and measure quality, if applicable:

- audio: intelligibility.
- visual: readability, conspicuity
- by action: awareness of what has been recognised

Administering a questionnaire to the subject at the end of the session can provide this kind of information.

Response time System response time is usually only relevant if it is too high. If the impression arises during initial testing that the response time is not a limiting factor, measurement can be omitted. Instead of a measurement, a subjective impression of response time can be considered. Sometimes the recognition system itself can output the response time, in which case verification of this information is necessary.

Situational awareness This can be expressed as the number of times a test subject uttered a command in a context where it was not allowed. A subjective impression by the tester or the subject can also be used as a measure. Suitable questions could be:

- Is the list of possible commands always clear to subject?
- Are special skills required (learning effect)?
- Is on-line help available?

3.3.3.2 Overall measure

In a subjective test, the ultimate overall measure is: "Can the task be completed?" This is a measure that includes recognition, error recovery, situational

awareness, and feedback. In this sense, the time required in order to complete the entire test might also be indicative of the quality of the system.
General impressions of test subjects can also be indicative of how the system performs.

3.3.3.3 Benchmarks

It is always a good idea to include a benchmark test. For command and control, this means "how do people perform *without* the speech input?". Here, the recognition rate cannot be used as a measure but other measures, such as time to complete the task or subjective impressions can be used. Also, a 'Wizard of Oz' (WOZ) experiment can be set up. In this case, the commands are not interpreted by a recognition system but by a hidden human participant ('the man behind the curtain') who feeds all spoken commands to the system. Using this technique the system can be evaluated without the speech input, and results can be compared to the performance with speech input. Note that a WOZ experiment can take a fair amount of effort; cf. Gibbon et al. (1997) for discussion in the contexts of corpus collection and system evaluation.

3.3.4 Examples

In this section we give two examples of tests that have been carried out on command and control systems.

3.3.4.1 Evaluation of a voice manager for an Advanced Crew Terminal

The project SPACT[2] for the European Space Agency involved adding speech input/output to an existing operating environment, the Advanced Crew Terminal (ACT).[3] The ACT is a collection of tools that can help an astronaut in his daily work, providing electronic time schedules, procedure checking, experiment control, and data acquisition. It was implemented in a Microsoft Windows operating environment as a collection of application programs.
Speech input was implemented with a consumer off-the-shelf (COTS) recognition system, of which mainly the voice commanding part (the 'voice manager') was used. The voice manager is a clever tool that can track the contents of the currently active application (the window that has the focus), and dynamically adapt its active vocabulary. All normal Windows widgets can be read, such as menus, buttons, check boxes, radio buttons, pull down menus etc.

3.3.4.1.1 Evaluation goal

The specific goal of the evaluation was to determine the word error rate of the speech recogniser in the command and control mode, specifically for the words occuring in ACT. The ASR product is built for dictation, so normally evaluations are carried out in dictation mode. The problem of assessing the recogniser in command and control mode is that the voice manager itself cannot be controlled, i.e. the active vocabulary cannot be set externally.

[2]ESA contract number ESA C11695/95/NL/JG. The study was managed by ESA/Estec, Directorate of Technical and Operational support.
[3]ESA contract number ESA C10524/93/NL/JG

3.3.4.1.2 Setup of the experiment

First, an analysis of all possible command phrases in the ACT framework was made. This led to an inventory of the overall vocabulary and many active vocabularies, depending on the context. The voice manager was taught all out-of-vocabulary (OOV) words by giving an approximate spelling and an acoustic example. This extension of the vocabulary was done by one test subject, because this particular ASR product only stored the phonetic representation of an expression.
Then a special Windows application EVAL was built which allows the assessment of speech recognisers in the command and control mode. This program reads 25 expressions from a text file and places the command texts inside 25 buttons. A 26th expression is read from the file and put in a large 'target button' (see Figure 3.1). Subjects are requested to say the word in the target button. The voice manager dynamically reads all expressions found in the buttons, and places them in its active vocabulary. Thus, it will recognise one of the 26 alternatives. After recognition, the voice manager 'presses' the recognised button. EVAL records button 'pressed' into a logfile, and loads 26 new words from the text file into the buttons. Thus the active vocabulary can be controlled. When the voice manager does not recognise any word, causing a miss or deletion, this has to be recorded by hand.

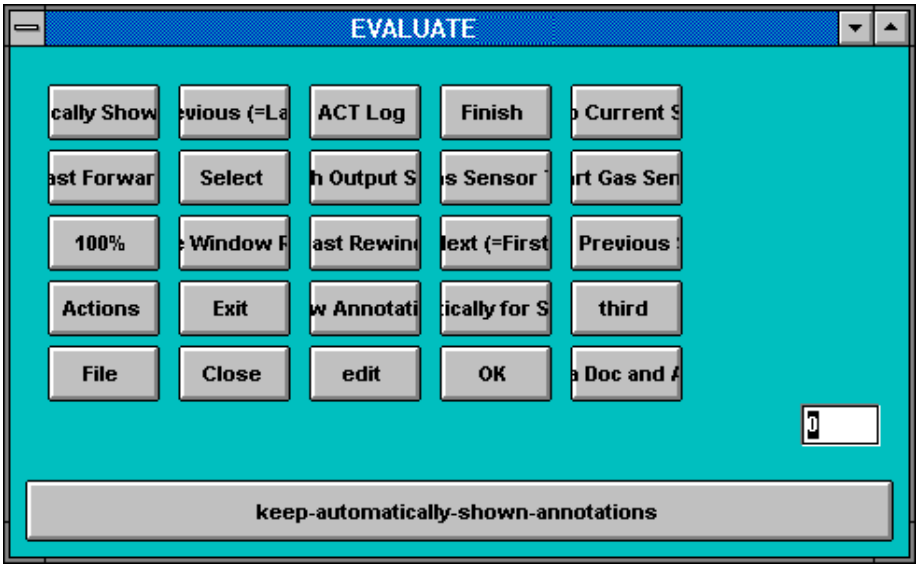

Figure 3.1: The diagnostic program EVAL

The EVAL interface shown in Figure 3.1 displays 26 buttons, the contents of which are read by the ASR system, operating with a well defined dynamic vocabulary. In the case shown in the figure, the test subject is requested to utter 'keep automatically shown annotations', one of the possible commands in the ACT framework. Other words are shown in the smaller buttons, which are

not intended for the user; the speech recogniser is able to read the words and add them to the active vocabulary.
There were 221 words in the ACT vocabulary. All of these were displayed exactly once in the test. Thus, the input test file consisted of $221 \times 26 = 5746$ expressions. The test perplexity is 26, and the 25 confounding words were chosen from the words that are normally active in the ACT application where the test word occurs. Care was taken that the 'always active words', such as the command "close window" were indeed always available in the command buttons, giving a representative situation of the confounding (confusable) word set, and as a safety measure against premature ending of the experiment.

3.3.4.1.3 Experiment

The acoustic conditions in space can be severe. In the Mir space station the noise levels have been determined to be 74 dB(A). This noise was regenerated in the laboratory, shaped to match the noise measured on board the Mir.
Because the enrolment of the dictation system takes a long time for a test subject, only a limited number of subjects was used. Two subjects participated in both quiet and noise conditions. The enrolment was carried out without noise.
In Table 3.5 the results of the assessment are given. Standard deviations are estimated from the word error rate and the number of words through the binomial expression, equation 3.2 on page 218.

Table 3.5: Recognition results, for three subjects. There is one condition where the enrolment speaker and the test speaker were not the same. The speaker who trained the exception vocabulary was always pp1. The test consisted of all 221 expressions defined in the word lists for ACT. The last column indicates the word error rate (WER), which does not include misses.

Subject	Enrolment	Noise	Correct	Error	misses	WER
pp1	pp1	office	217	4	6	$1.8 \pm 0.9\,\%$
pp2	pp2	office	217	4	19	$1.8 \pm 0.9\,\%$
pp3	pp3	office	217	4	5	$1.8 \pm 0.9\,\%$
pp3	pp3	office	220	1	5	$0.5 \pm 0.5\,\%$
pp1	pp2	office	211	10	40	$4.5 \pm 1.4\,\%$
pp1	pp1	Mir noise	218	3	7	$1.4 \pm 0.8\,\%$
pp2	pp2	Mir noise	217	4	24	$1.8 \pm 0.9\,\%$

3.3.4.2 Fast jet cockpit control

A project commissioned by the Royal Netherlands Airforce has been completed in which the cockpit of a fast fighter jet was extended with a speech input modality (Cockpit 1996). The target system was the cockpit of the jet aircraft after a technical modernisation update.
The project consisted of 3 parts:

- COTS recognition system selection,
- implementation in an F16 simulator,
- evaluaton with test pilots as subjects.

In the first part, a connected word recognition system that was implemented as a PC plugin card was selected. It was verified that the system was capable of recognising 100 dB(A) under the high noise levels ambient in a fighter jet cockpit. This involved the use of a noise-cancelling microphone mounted under the oxygen mask of the pilot.
For this application, voice control over all non-critical functions of the modernised aircraft was implemented. These included radar, navigation and radio control. Because the recogniser was a connected word system, complex commands could be issued, such as 'switching VHF-channel one four enter' or 'radar range up.' In total, a vocabulary of 281 words was defined, and a complex grammer allowed for a wide range of possible commands. The perplexity of the grammar was about 17.

3.3.4.2.1 Laboratory evaluation

In the laboratoy the recognition system was evaluated with recordings of read command strings. The utterances were made by pilots who had not been introduced to the possibilities and command syntax of the aircraft. The strings were read from a computer screen, making the talking style read speech.
Evaluation of the recogniser performance was performed by straight forward comparison of recognised words and prompt texts. Based on a test of 3 persons, the avarage word error rate was 3 %, and the utterance error rate 5 %.

3.3.4.2.2 Simulator evaluation

For the evaluation of the total system, the speech recognition system integrated with the cockpit controls, a subjective test with professional air force pilots was conducted in a simulator. In the simulator the spoken commands had the same effect as the normal manual controls. Three subjects participated in the test. This number was limited by the availability of the pilots.
First, the subjects were instructed how to use the spoken commands. Then they uttered all of the 281 words three times to train the recogniser, wearing the oxygen mask. After that, they flew a number of sorties (flights) in the simulator. During the sorties, certain tasks had to be carried out. Missions were simulated that involved air-to-air and air-to-ground fights. In the mission the pilots were requested to use voice commands for operational actions.
The first sorties were flown in order to give the pilots the chance to get used to the new system. During adaptation, the pilots were reminded how to issue the spoken commands by a skilled experimenter. For the final evaluation sorties, a new recogniser training session was carried out in order to correct changes in speaking style due to experience.
In total, 17 sorties were flown in the evaluation. All utterances were recorded on digital audio tape and the recogniser output was logged to a computer file. The recogniser used a press-to-talk switch in order to avoid recognising normal communication as spoken commands. The times at which the press-to-talk

switch was pressed were also recorded. The recordings were used in order to be able to analyse the recognition results later in the laboratory.

> *Recommendation 20*
> Always record all spoken utterances during a field test. This will help you later in analysing the data.

During the sorties the spoken commands were directly transcribed by an experienced experimenter using a commercial word dictation system, by repeating all commands in a close-talking microphone. In this way, a coarse indication of the word error rate was obtained for use in debriefing the subject. In some occasions, particular words were re-trained before the next sortie was flown.
The direct feedback given to the subject was through a 4-letter abbreviation of the recognised word in the Head-up Display (HUD). This feedback was not really used by the pilots.
Subjective assessment measures were obtained in the de-briefing after the sorties were flown. Word and command string accuracy were determined in the laboratory afterwards, on the basis of a time-alignment of the recognition log and a transcription of the recording.
In the evaluation it turned out that there were many instances where the recogniser was at a different 'node' in the syntax than the context which the test subject expected: there was a loss of 'situational awareness' of the recogniser state. Also other 'mistakes' of the pilot were encountered, for instance, the press-to-talk switch was sometimes used with slightly incorrect timing, giving rise to partly cut-of or missed words. Although the above errors are representative of the implementation as a whole, it gave a misleading impression of the performance of the speech recognition subsystem.
In order to measure the 'pure' recognition performance, excluding effects due to loss of situational awareness or incorrect use of the press-to-talk switch, a second evaluation was performed in the laboratory afterwards. Of all utterances recorded, only the ones were selected that were correct according to the syntax. These were replayed in the laboratory with the same recogniser under identical training conditions.
Finally, the recorded database of spontaneous speech was used in order to assess new speech recognition systems, and comparing the results to the original recognition system. In Table 3.6 the average word error rates for the three conditions dicussed above are given.

Table 3.6: The recognition performance during the sorties, after cleaning up the utterances, and with a different recognition system.

Condition	Recogniser	Word Error Rate
during sorties	standard	0.73
only correct utterances	standard	0.78
only correct utterances	new system	0.85

3.4 Document generation

3.4.1 Typical systems

A document generation system is a system which allows entry of data or documents, e.g. a database front end or a word processor. Speech input can be added to such a system in order to achieve faster entry of data or text. Speech input systems are the automatic version of the dictaphone, where professional typists process the remarks recorded on tape. It is therefore not surprising that traditional niche markets such as radiology reports were first entered by the speech recognition companies. Nowadays, commercial versions of dictation systems exist for a wide range of domains, running on personal computer hardware.

3.4.1.1 PC large vocabulary dictation systems

Several companies now offer dictation for PC hardware. The recognition occurs entirely in software; it tends to require very high processor capacity and has quite large memory requirements.[4] Speech is captured via a standard sound card. These dictation system products often have separate editions that differ in vocabulary size (20,000, 50,000, 120,000 words) and the domain the vocabulary has been optimised for (journalism, medical, legal, etc.). Originally, these systems were isolated word recognition systems, but now continuous speech versions are also available. The products are available in a number of languages, including English (American and British pronunciation), the major European languages, Chinese, and Arabic.

3.4.1.2 Domain-specific professional dictation systems

The market for professional dicatation is traditionally determined by radiologists and lawyers. The task involves many highly domain-specific words, and relatively standard sentences. A professional system is different from a commercial PC system in that is has been developed specifically for the profession concerned, and may therefore run on special purpose hardware. The training may have been carried out using dictaphone tapes and written transcripts. The grammars that these dictation systems use are based on these collections of written transcripts.

3.4.1.3 Data entry systems

A wide variety of data entry applications can be made speech aware by integration with a speech recognition system. The type of data entered may vary from digits and letters for car licence plates, to charactarisations of items under inspection. A limited vocabulary and reasonably strict syntax are typical for such implementations.

[4] This statement has no absolute meaning, as economically available processor and memory capacity increase rapidly from year to year. It must also be seen relative to the resource requirements of other interactive computer applications, such as office-suite programs.

3.4.2 Typical issues

3.4.2.1 Enrolment of speaker dependent/adaptive systems

Long and professional usage of a speech input system makes it profitable to invest some time training the sytem to the specific user's voice. In large vocabulary dictation such a process is called 'enrolment'. There are commercial considerations to keep the enrolment procedure as short as possible and perhaps even let the user start with a speaker independent system. If a dictation product is sold in the shop, a new buyer wants to see results immediately, without a lengthy enrolment procedure, indeed if possible with no enrolment procedure at all. However, large vocabulary dictation often benefits from speaker adaptation. It is therefore recommended to include the enrolment procedure in the evaluation test.

3.4.2.2 Learning curve of the test subjects

The enrolment procedure not only has the function of adaptation of the system to the user, but also to enable the novice user to learn how to use the dictation system. The user will learn to use words for punctuation, and commands for correcting errors. The manner of speaking will change over a period of time, i.e. the user's skills follow a learning curve. This effect is relevant when test subjects are used in order to assess a dictation system.

3.4.2.3 Uncontrolled adaptation strategies for dictation systems

Systems are often not only adaptive to the speaker's voice, but also with respect to acoustic channel, vocabulary, grammar and word models. The suppliers do not usually refer explicitly to all of these. One should be aware in designing a test that the active vocabulary and the language model might change during use. Dictation systems generally have a strategy of learning from errors in recognition, but this requires that the system knows which dictated texts have been corrected by the user and which have not. Therefore there is usually an explicit action required from the user to indicate that a piece of dictated text has been corrected. It is worth trying to analyse the product in order to identify the conditions under which the language model and vocabulary are updated.

3.4.2.4 Different error correction strategies

Related to the various adaptation strategies is the issue of error correction. Error correction can be implemented

1. During the dictation session: A special command (e.g. 'Oops!') is issued by the user to indicate that the last word that has been recognised must be corrected. These methods need word-by-word on-line recognition.
2. Directly after the session: Some systems do not require visual feedback during the dictation session but allow the user to read the dictated text after the session, and correct individual words or phrases. This method needs to store the original waveform in order to be able to play back the words that are apparently wrong.
3. Some time after the session: It is also possible that the error correction is carried out by someone else at a later stage. In this case the whole dictation status (uncorrected text and original waveform) can be saved for later operation.

The speed and efficiency at which errors can be corrected may vary; usually there is a list of 'top-N alternatives' for words.

3.4.3 Evaluation design

Because the state of the art in document generation systems develops rapidly, it is hard to give a general outline of the evaluation design. However, a number of items can be given to be taken into account.

Domain Define a domain of speech for which the system is tested. This should match the expected application of the product. For PC dictation systems, a number of topics should be chosen. For instance, if the domain of the product is 'Journalism' (a common package for PC dictation systems), one could use an article from the current daily newspaper as a test text. Several different newspapers should be used as sources, and several topics should be selected.

Adaptation control Make a distinction in the test between where adaptation is allowed and where it is not. In the latter case, make sure that no actions will be taken by the test subject that will influence the acoustic or language model of the dictation system. In the case where adaptation is allowed, a sufficiently long learning period should be used, after which the evaluation test can take place. As an example, the dictation task can be completed twice. The first time, the system should start out fresh. Then error correction can be completed (including adding new words to the vocabulary), and the same text can be dictated a second time. If there is language model adaptation, results should be a lot better.

Accuracy Performance is usually expressed as the percentage of correctly recognised words, misses (deletions), insertions, and substitutions (confusions).

Dictation speed Record the number of words per minute.

Error correction strategies A good measure for the easiness of error correction is the average time spent per correction. This is a parameter that is difficult to measure accurately: ideally, the number of corrections necessary for each test subject and dictation system is more or less the same. Special highly confusable test sentences may be used for this test.

Dictation speech In representative usage the dictation system is used in 'dictation speech' mode. In the test a predefined text is almost certainly used. Try to include a test in which the test subject is asked to dictate a letter to a friend or relative about a neutral subject, for instance the subject's last holiday.

Benchmark The dictation system can not only be compared to another system but also to humans. For such a 'human benchmark,' professional secretaries should be employed who can use recordings of the first dictation session as input. Control over the playback device should be given to the secretary, as is the case with dictaphones. Error rate and dictation speed are the most obvious performance measures for the human benchmark.

3.4.4 Examples

3.4.4.1 Comparative test in *C'T Magazin*

The German computer magazine *C'T Magazin* conducted a comparative test between five PC dictation systems of two different manufacturers (Malaske 1998). The systems were both isolated word and continuous speech recognisers, and used various underlying operating systems. The performances were tested on the same hardware.

3.4.4.1.1 Test protocol

The test consisted of several steps, for each recognition system, after the usual installation of the software product had taken place:

1. Without any enrolment or adaptation (or if this was not possible, with a minimum amount of training, 'fast training'), a text of 100 words and 20 punctuation markers was read. From this a speaker independent score was obtained, as well as an indication of how well the standard vocabulary covers the text.
2. The enrolment was performed.
3. Five short texts were dictated, and error correction was carried out immediately. This step was then repeated, until the recognition rate did not improve.
4. A new text was read to the recognition system, and the final recognition rate was determined from this result. Another measure was dictation speed.

Apart from the general protocol, specific quirks and features were mentioned in a separate section in the report for all of the systems. These were typically statements about

- specific words that gave rise to problems in recognition, such as inter-punctuation symbols,
- capitalisation of words,
- adding words to the dictionary,
- mode of operation: integration in word processor versus own dictation window,
- the enrolment text, the time it took to do the enrolment, etc.,
- spelling modes,
- influence of noise,
- possibility of voice macros,
- organisation of the speech models on hard disk,
- choice of dialect,
- synthesis feedback possibility.

3.4.4.1.2 Results

The outcome of the test that is most comparing, is a final table 'checklist,' summarising the technical specifications and some subjective judgements on the quality of a number of aspects. The judgements are rated on a five-point scale ($--$, $-$, 0, $+$, $++$). Most results are more descriptive by nature, and can be found in separate sections for all of the five products.
It is not mentioned in the article how many test subjects conducted the test. There must have been be more than one because some numbers are averages.
It is mentioned that the accuracy percentages are only indicative, and the figures are not used in the final table. Three percentages are mentioned in describing paragraphs: the accuracy dicrectly after installation (none/minimal training), the average accuracy of the vocabulary training cycle, and the final percentage for the new, unseen text. Also mentioned is the number of times the vocabulary training had to be repeated before the word error rate stabilised. As an example, the results are given in Table 3.7.

Table 3.7: Accuracy results in *C'T Magazin* test. (n.m. = not mentioned)

System	A	B	C	D	E
accuracy before training (%)	80	70	70	75	90
accuracy during cycle (%)	96	96	98	96	98
accuracy new text (%)	93	93	90	90	95
number of cycles	2	1.5	3	3	3
enrolment time (h:m)	n.m.	0:45	n.m.	0:40	n.m.

3.4.4.1.3 Discussion

The evaluation performed in *C'T Magazin* represents a good consumer oriented attempt to compare commercial products that have a similar function (PC dictation systems), but have quite different implementations. Both isolated word and continuous speech recognisers were tested on various software platforms. The problem of comparing systems that need explicit enrolment and systems that rely more on adaptation is approached by having both a full enrolment and an adaptation cycle. The evaluation is mostly qualitative, and the accuracy numbers mentioned did not undergo any statistical test or comparison. Details about the number of subjects, the order of testing etc. are not given, but one has to bear in mind that the evaluation is a consumer test and not scientific research.

3.4.4.2 Human benchmark for PC dictation systems

At the University of Munich, Germany, a comparative test between two German PC dictation systems was carried out, which included a comparison with a professional secretary (Burger and Tillmann 1997). The PC dictation systems were evaluated by a single test subject by dictating texts from a business consulting environment. Three experiments were carried out:

ADAPTATION The recognition rate for the first six dictated pages was measured. After each page the word error rate was determined so that the adaptation effect could be studied.

LEARNING A text of 392 words was dictated seven times, in order to study the effect of adaptation of the system by learning from error correction.

DICTATION SPEED The time to accomplish dictating the text used in the previous experiment was measured, including error corrections. Also a professional secretary typed the text, and made corrections. The time it took her to do this was used as a reference.

The result of the comparison of systems with a human subject was that the dictation speed varied from 12 words/minute in the first iteration to 22 words per minute after the fourth iteration. The professional secretary typed the text (including corrections) at a speed of 26 words per minute. There was no apparent difference between the two systems.

3.4.4.3 Multilingual comparison of a PC dictation system

Another method of comparing dictation systems is to compare the same system in several languages, rather than to compare different systems. A leading dictation software company has performed such a test (Barnett et al. 1995). Their current PC dictation system (an isolated word, large vocabulary recogniser) was available in 5 different languages. The company performs regular evaluations of their own product. The comparison between languages was published.

3.4.4.3.1 Test protocol

In order to compare languages, the test documents were chosen from widely translated authors, Hegel, Verne, Saint-Exupéry, Grisham, and a section from the user manual of a speech recogniser. Test subjects were native speakers of the languages tested, which were English, French, German, Spanish and Italian. The number of subjects per language was four (two male, two female).
The dictated speech signals of all test subjects were recorded, and stored with the reference transcriptions, permitting the evaluation of the recogniser to be performed in 'batch mode' and repeated several times with the same acoustic data. In this way, two measures of recognition accuracy were determined:

WITHOUT ADAPTATION The word error rate of one of the five texts was determined with the recogniser cleanly 'out of the box' without any adaptation. This was repeated for all five texts, and for all subjects in all languages.

WITH ADAPTATION The word error rate of one of the five texts was determined after the recogniser had been adapted to the speech of the other four texts spoken by the same speaker.

The recordings also allow evaluate of the product itself, as it evolves in time.

3.4.4.3.2 Analysis

In this evaluation, more than the word error rate alone was determined. In order to give an impression, the following items were studied:

THROUGHPUT This is a measure of the word error rate, which includes the effort it takes to correct an error. The measure is 100 % for perfect recognition. With errors, the contribution is such that simply recoverable errors (e.g. ones that appear in the alternatives list) weigh less than errors that are difficult to correct (e.g. words that must be spelled out).

HOMOPHONES Homophones are words that have the same phonological form but differ with respect to their orthographic form. Errors of this type cannot be blamed on the acoustic analysis of the recogniser, but are due to the language model. The number of homophones differs across languages.

IN-VOCABULARY ERRORS Many of the errors made are due to the fact that the uttered words were not in the (active) vocabulary. These errors are a direct consequence of the vocabulary coverage of the language. For instance, with the same vocabulary size, the coverage of German is smaller than that of Engish by a factor of 3–5 because of the presence of inflected word forms. By only looking at the errors of words that were *in* the vocabulary, the influence of this effect is corrected.

WITHOUT LANGUAGE MODEL The bigram language model was removed in order to see the influence of the language model on the performance.

Some of these studies can only be made by the developers of the system because they require knowledge of the internals of the system, which normal commercial users do not have. The experimental design was forced to include different subjects for different language conditions. The analysis of difference was carried out using F-ratios of the variance of the word error rates within languages and between languages.

3.4.4.3.3 Results

Some of the results are summarised in Table 3.8. For this dictation product, the effect of speaker adaptation is clearly visible. The effect of the language on the word error rate is significant at the $p = 0.01$ level. More results can be found in the article (Barnett et al. 1995).

Table 3.8: Some results of the experiment, showing language dependence of the word accuracy, expressed in %. The last line shows the homophone error rate.

Language	German	Spanish	Italian	French	English
without adaptation	82	86	89	84	87
with adaptation	87	89	92	87	91
idem, in vocaulary only	91	92	94	88	91
homophone error rate	22	25	17	73	25

3.5 Services and telephone applications

3.5.1 Typical systems

Services and telephone applications are closely related: a large market segment for speech technology lies in telephone services. A service that is not related to telephone applications would be an automatic kiosk information center that communicates via speech, or an automatic teller (cashpoint, bankomat). Examples of other systems are:

SWITCH SERVICES Telephone companies can use speech technology in the switch to provide services like voice dialing (calling a person by uttering his name), voice mail ('answering machine' service), and short message services (SMS, reading out a short message or an electronic mail message). The technology runs inside the switch, and can be tested only by calling the service.

TRAVEL INFORMATION In many countries, a large effort is in progress to automate travel information services by using speech technology. In this case the technology is implemented in a call centre instead of the telephone switch itself. However, the access procedure is very similar to the previous example.

PAROLE MONITORING In the United States a system is in operation where the location of someone who is on parole can be verified by checking the caller's telephone number against his voice pattern (speaker verification).

Bank transactions Another example of speaker verification technology is found in banking services. In order to authorise bank transactions, the voice pattern of the client is used.

3.5.2 Typical issues

The interaction procedure for a telephone service is quite different from that of a command and control and dictation system. The only feedback is usually a synthesised voice. Related to the services, there are a number of issues involved in assessment:

Dialogue control When the system is tested, a dialogue must be started with the system. For instance, in order to evaluate recognition, a path in the dialogue must be chosen that permits words of the test set to be uttered. However, the position in test dialogue depends on the recognition result. In case of correct recognition, the path is followed as expected, but when a recognition error occurs, the path back must be found in order to continue the test.

Assessment of recognised word Similarly, it is difficult to determine if a word has been recognised correctly. This can be done only indirectly by analysing the response of the system. Because the systems often respond with speech it is very difficult to make an automatic assessment. The tester must do the scoring of the words by hand.

Synthesis Because of spoken feedback, the assessment of the synthesis system employed becomes an important factor in the evaluation of the service as a whole.

Speaker verification As indicated in the examples above, speaker verification is a technique that will be used in many services. Special tests are needed to assess the quality of that part of the system.

Communication channel Telephone applications have a limited bandwidth by definition. The standard bandwidth is 300–3400 Hz, but this may vary between different connections. The male voice in particular loses the fundamental frequency in this bandwidth, and this may have consequences for the intelligibility and the voice quality of synthesised voices. Apart from this, the signal-to-noise ratio of the connection may vary, as well as the absolute value of the signal. However, nowadays more and more telephone switches have become digital, which reduces the amount of variation due to the communication channel.

Noise Services in kiosks are most likely to be used in places where there are many people. This means that there is often background noise (people, outside street noise) and also the acoustics may have special properties (for example a large hall).

Stability of system Service providers constantly update their systems in order to improve the performance. For instance, if a speech recognition system is employed, a service provider may use recorded information of earlier calls to improve the quality of recognition in the future. If such a system update occurs during the evaluation, the results may be strange and invalid.

> *Recommendation 21*
> When testing a service, make sure that the system itself stays the same over the period of the evaluation. It might be wise to do the test in cooperation with the service provider.

3.5.3 Evaluation design

In the evaluation of a service, one usually has to evaluate the service as a whole. It may be impossible to select only the speech input side of the system, which

means that the dialogue structure has to be studied, and that the evaluation procedure has to be designed around the dialogue. The following steps are important in the design and it is recommended that they should be followed:

Dialogue structure Experiment with the service, and try to get a complete overview of the dialogue. This involves making an inventory of the menu structure, but also finding out what kind of exceptions can occur. For instance a response like "Sorry, I didn't understand you, please repeat" might be possible in many situations.

Task Define tasks to be completed. For a voice dialing system this can mean programming several names and trying to call people by saying their name.

Measures Depending on the service, several performance measures are possible:

- Success rate What is the percentage of the times the task was successfully completed?
- Speed What is the average time it takes to complete a transaction or to retrieve an information item?
- Impression What is the subjective impression of the system?
- Voice quality What is the quality of the synthesis?
- False negatives How many attempts are necessary for a user to identify himself successfully to the speaker verification system?
- False positives The number of successful attempts at unauthorised access.

Item list Make a list of items to be assessed. For instance, for the speech recognition part of the service a number of keywords can be selected. These can include general speaker independent words, and specific speaker dependent words.

Design Try to design a test protocol that will evaluate all of the test items in the list. Sometimes it may be difficult to reach a certain point in the dialogue structure, in such cases consider not to test the specific item.

Human benchmark with normal telephone operator For the performance measures 'success rate' and 'speed' one might include a service that is operated by human telephone operators (or in the case of a kiosk service: a normal counter).

3.5.4 Examples

3.5.4.1 Comparative study of two voice dialing services

For a European car magazine a comparative study of two voice dialing services was made (Roks 1997). Two mobile phone network providers had recently started this service, and the question was how well they worked in a car that was equipped with hands free GSM telephones.

3.5.4.1.1 Experimental design

The service provided is calling somebody by speaking his name. This is implemented by a speech recogniser in the telephone switch (unlike some GSM handsets that have speech recognition in the telephone itself). Both providers required some training of command words, and the training of all names in the list of callable persons. The service was provided only for GSM networks, and the test domain was the use of the mobile phone in the car. The research

question was to determine the quality of the recognition system under various conditions.
The main parameters measured were the recognition rate and the call setup time (the time between the utterance of the name, and the moment the phone rings). The recognition rate was expressed by defining a penalty for each recognition error. Different penalties were used for various errors. In Table 3.9 a summary of the possible errors is given.

Table 3.9: The penalty values for various errors

Error	Penalty	Example
Confirmation	1	"Program"—"'Program,' is that correct?"
Deletion	2	"Yes"—"Please say yes or no"
Substitution	5	"Program"—"Setup menu. ..."

The experimental design further concerned the following points:

Acoustic environment The evaluation was performed by seating test subjects in a laboratory car mock up, with simulated car noise according to ITU car spectra. By changing the level of the noise, different car driving speeds were simulated.

Conditions There were eight test conditions in total. The services were tested with two car speeds (80 and 110 km/h), two GSM telephones in combination with a car kit, and two different microphones (standard 'clip-on' microphone, and directed swan neck microphone).

Balancing 18 test subjects took part. Each of them tested all conditions. The conditions were perfectly balanced in order (over the first 16 subjects), so that learning effects would be averaged out between test subjects.

Paradigm Subjects read instructions from a service provider and listened to an introduction. They chose five names of persons they knew. Then these five names were trained according to the protocol of the service provider. Finally they tested each of the names by uttering them one by one. The telphone number stored for all the names was the number of the laboratory, and was keyed in rather than spoken.

Scoring The experimenter followed the dialogue outside the test room. Using a test protocol form, he could indicate the success of the various steps taken. For instance, for each of the five names to be trained a note of successful or failed training was made, as well as the necessary steps to reach the programming mode: the words "Program," "Add name," etc.

3.5.4.1.2 Results

For each subject, a system was tested for 20 calls (5 names, two telephones and two driving speeds). The average penalty found, over all subjects, was 3.1 and 5.1 for the two systems. The difference was significant at $p < 0.05$. No significant differences were found between the telephone brands, the driving speed, or the gender of the test subjects.
There also was a very significant difference in the average call setup time for the two systems, 24.7 and 17.4 seconds respectively. It must be noted here, however, that the system showing the longer call setup time had an extra prompt for the

novice user, before actually dialing the number: "The system will now call ..., say 'abort' to abort." This prompt is left out after a certain number of calls, but because the system was reset for each new condition. This state was never reached in the test.

3.5.4.2 Elsnet Olympics

At the EUROSPEECH'97 conference in Rhodes, ELSNET organised an evaluation of several spoken language systems (den Os and Bloothoofd 1998). In a special call room, 10 different systems could be called. Participants of the conference could call the systems, and fill out a questionnaire afterwards.
In total, over 250 questionnaires were returned to the organisers. A statistical analysis showed that five factors explained 75 % of the variance in the answers. They were:

1. general appreciation,
2. functional capabilities,
3. intelligibility of speech output,
4. the user's proficiency in the spoken language
5. the user's familiarity with spoken language systems.

The last two factors are user-oriented because this was an international conference and many of the subjects were non-native.

3.5.4.3 The MASK Kiosk

In the ESPRIT project MASK (Multimodal-multimedia Automated Service Kiosk), a prototype for an information kiosk with multimodal input and output is being developed (Gauvain et al. 1997). The input modalities comprise touch screen and speech recognition with a language understanding system.
Evaluation of the kiosk functionalities is carried out continuously during the project. Test subjects operate the kiosk in order to get a subjective and objective evaluation figures, and the responses are used in order to improve the performance of the system.
The speech recognition part is evaluated in terms of speech recognition accuracy and speed. The language understanding part is evaluated using written transcriptions of the speech. In this way, both parts can be evaluated independently.
Apart from this object evaluation, users filled out questionnaires in order to get subjective assessment of the system. Questions were categorised in terms of

1. user friendliness,
2. reliability, and
3. ease of use.

One of the results is that with improving system performance subjects speak more easily to the system, and use longer and more varied sentences. The prototype was used for collecting representative speech and language data during a field test at St. Lazare train station in Paris. Using the new information the system's performance was improved.

3.6 Conclusion and summary of recommendations

The field of consumer off-the-shelf (COTS) product testing for spoken language technology products is a young and rapidly developing field, with practitioners in areas as varied as research institutions, industrial application development, and consumer magazines. In the present contribution, basic criteria for product classification and types of evaluation procedure were discussed, and the evaluation of a number of different products was discussed in detail.
The recommendations for product testing are summarised below.

1. Be aware that there are speech recognition manufacturers who claim that their product can deal with 'continuous speech' while in fact the systems are isolated word recognisers which still need tiny pauses, or which can only deal with limited vocabularies (e.g. digits) in continuous speech.
2. Decide what is the minimum difference in performance which can be considered an interesting difference.
   Choose a level of significance, e.g. $p < 0.05$.
3. If possible, make an estimation of the mean and variance of the performance measure, and base the number of subjects needed on these estimates and the minimum interesting performance difference.
4. Try to use at least 4 test subjects, 2 male and 2 female.
5. Try to balance the testing order of the systems among the subjects.
6. Before automation of an evaluation test setup into a testbed, assure yourself that the investment effort will be returned.
7. If a questionnaire is to be filled out, reserve a spot in the test protocol where there is no intrusion in the test itself. Also, the test subject should not be put under time stress.
8. Do not ask the subjects to answer more questions than necessary. Do not ask the same set of questions more often than necessary.
9. For subjective measures, use a five-point scale for the answer.
10. Make an inventory of the acoustic environment in which the product or service is typically used.
11. Be aware that for noise conditions higher than 60 dB(A), the Lombard effect may change the level and intonation of a speaker.
12. Do not expose test subjects to ambient noise levels higher than 80 dB(A), because otherwise a permanent hearing loss could result from the experiment. For experiments at higher levels, consult a human hearing specialist first, and have a proposal for the experiment checked by your local scientific ethics committee.
13. For PC-based systems, check that the microphone is recording without distortion. If necessary, use a separate microphone amplifier and use the line input.
14. Make sure that the electrical signals do not overload the system.
15. Make an inventory of all the capabilities for all the systems that are evaluated. In comparative test, only the common subset of capabilities can be compared quantitatively.
16. Unique capabilities may be very important to the general quality of a product or service.
17. Try to identify the underlying technology of the speech recognition system. It can be of importance to the design of the test. Also verify that all words in the application are in the vocabulary, i.e., can be recognised.

18. Analyse the type of feedback the system gives after recognising a command word or string.
19. Separate the individual performance measures in the test.
20. Always record all spoken utterances during a field test. This will help you later in analysing the data.
21. When testing a service, make sure that the system itself stays the same over the period of the evaluation. It might be wise to do the test in cooperation with the service provider.

4 Terminology for spoken language systems

4.1 Introduction

4.1.1 Terminology standards

One of the goals of the EAGLES project is to start to document the previously neglected area of terminology for the human language technologies; strictly speaking, this includes the terminology of the discipline of terminology itself. Although there are many 'dictionaries of linguistics', glossaries in technical handbooks, and the like, there has been little attempt to establish terminological standards in this area.
This chapter is concerned with documenting the main factors involved in terminological standardisation in the area of spoken language systems. In order to do this, two parallel strategies are followed:

1. The principles of the discipline of terminology, of the construction of a 'terminology', and of the creation and use of terminological databases (termbanks, termbases) are documented and discussed insofar as they are relevant to this task.[1]
2. An experimental prototype of a simple but practical internet–based terminological database (EAGLET – EAGLES Term Database) was developed in order to illustrate the practicability and potential of this endeavour.

Spoken language technology is a sub–field of human language technologies, along with a number of others:

- Spoken language technology (the older term 'speech technology' is still very frequently used) is concerned with the automatic analysis, synthesis and identification of spoken utterances.
- Natural language processing (NLP) is the discipline concerned with the automatic analysis and generation of written texts.
- Computational linguistics is concerned with formal theories of the representation and processing of language, with a non–exclusive tendency to concentrate on written language.
- Newer interdisciplinary ventures address areas such as gestural accompaniments of verbal communication, autonomous gestural communications, and multimodal communication (i.e. communication in various sensory and motor modalities).

A long term goal for an enterprise such as this would be to provide overall terminologies for all of these disciplines, within the bounds set by the need to do justice to innovation and development in the field. A further long term goal would be the development of multilingual terminology; this would require massive effort, however, and is not feasible in the present context.
It will be convenient to start with an informal working definition of the term 'terminology':

[1] The scientific background of this chapter, however, is that of lexicography and (computer) linguistics. This is reflected in the way of argumenting and also in the terminology used. A terminologist would have described many things differently. This the reader should bear in mind.

> Terminology is the vocabulary of specialised technical or scientific sublanguages.

According to the ISO standards, the aspect of standardisation is not implied in the definition of the term. In practice, however, the vocabulary of scientific sublanguages is subject to standardisation. The degree of standardisation may range from fairly systematic conventions at one end of the scale through the term systems of manufacturing companies to officially agreed national (such as BSI, ANSI, DIN) or international (such as ISO) definitions with legal or quasi–legal status.
At each of these levels of formality, terminology standards are needed in order to support a number of communicative requirements such as the following:

1. in scientific work, the theoretical cohesion of the subject matter and methodology;
2. in manufacturing industries, the support of planning and production processes and inter–departmental communication;
3. in industry, commerce and academia, the basis for consistent documentation;
4. also in industry, commerce and academia, the foundation for accurate technical translation.

Beyond this, the globalisation of political negotiations, the emergence of tightly knit global economic dependencies, and intricate international patterns of trade constantly create a need both in existing and newly developing areas for terminological standardisation, motivated by cost–efficiency at the material level, and the minimisation of misunderstandings at the level of communication. Terminology management has become a crucial tool for these purposes (Sager and Nkwenti-Azeh 1989). In general terms, then, terminology fulfils functions of supporting *information interchange*, *coordination*, and *re–usability of resources* in different contexts.
There are several standards laid down by the International Standards Organisation (ISO) which apply to terminology management, computational aids in terminology, layout, etc. The following list is taken from Schmitz (1998).

- ISO 639 (1988): Code for the representation of names of languages
- ISO 639-2 (1996): Alpha-3 code for the representation of names of languages (draft, will be finished soon)
- ISO 704 (1987): Principles and methods of terminology (revision will be finished soon)
- ISO 860 (1996): Terminology work – Harmonization of concepts and terms
- ISO 1087 (1990): Terminology – Vocabulary (is being revised)
- 1087–2.2 (1997): Terminology work – Vocabulary – Part 2: Computer applications (draft, will be finished soon)
- ISO 1951 (1951): Lexicographical symbols and typographical conventions for use in terminography (is being revised).
- ISO 6156 (1987): Magnetic tape exchange format for terminological / lexicographical records (MATER) (is to be replaced by ISO 12200)
- ISO 10241 (1992): International terminology standards – Preparation and layout
- ISO 12200 (1997): Terminology – Computer applications – Machine-readable terminolgy interchange format (MARTIF) (draft, will be finished soon)
- ISO 12616 (1995): Translation-oriented terminography (draft)

- ISO/TR 12618 (1994): Computational aids in terminology – Creation and use of terminological databases and text corpora
- ISO 12620 (1997): Terminology – Computer applications – Data categories (draft, will be finished soon)

The ISO standard 12620, which serves as the basis for the specification of a spoken language terminological database, gives precise descriptions of the *data categories* (i.e. the types of lexical information) and data items which can be selected for a terminological database. It does not, however, specify the structure of an individual term entry. With respect to spoken language technology some of these data categories need refinement while others may be coalesced into a single category (see Section 4.4.2.2).
The goal of terminological standardisation is always to provide a point of reference for calibrating vocabularies used in technical information interchange. In the present case, the EAGLET termbank is intended as an initial point of reference and a starting point for future development.

4.1.2 Termbank users

The potential users of terminological databases range from students in training and scientists and engineers in research and development consortia through the day to day business of communicating about technical objects and processes to the information and decision–making requirements of company and political management. In the case of spoken language terminology, potential users range from linguists and phoneticians to computer scientists and engineers involved in developing systems for automatic speech recognition, speaker verification, speech synthesis, and dialogue systems with multi–modal input and output subsystems.
Companies, particularly large companies, tend to develop their own terminology for purposes such as those outlined above. In research and development contexts, different criteria may apply: as the field develops, new concepts are introduced and old concepts mutate as results are evaluated and re–assessed. Many are sceptical, for different reasons, about the value of standardisation, including terminological standardisation:

- Companies, even in the same field, tend to develop their own terminologies according to company policies and structure.
- Scientists insist on terminological flexibility in order to support new approaches and minimise the danger of domination by particular scientific paradigms.

But whatever the differences, a minimal goal of terminological standardisation, for example in the operational form of terminological databases, is to make terms and their meanings as explicit as possible in order to facilitate coordination of processes, the interchange of information, and the re–usability of resources. More far–reaching goals involve the harmonisation of terminologies across boundaries of different traditions — such as electrical engineering and computer science or linguistics and phonetics — or different companies or different language areas.

4.1.3 Chapter outline

This chapter addresses various points to be taken into consideration when designing a terminological database for spoken language technology.
In the first section, some of the central notions of terminological theory are introduced (Section 4.2.1).
Second, relevant considerations in the area of spoken language terminology are discussed (Section 4.4), including differences in terminological, terminographic and lexicographic procedures.
Third, Section 4.5 addresses relational databases, the most common type of database.
Fourth, Section 4.7 introduces the EAGLET Term Database for spoken language systems, developed in the EAGLES Phase II project. EAGLET currently contains about 1250 term entries and is constantly being extended and revised in consultation with experts in the field. The core set of EAGLET terms are taken from the EAGLES Phase I Spoken Language Working Group *Handbook of Standards and Resources of Spoken Language Systems* (Gibbon et al. 1997).

4.2 Terminological basics

4.2.1 Central notions in terminological theory

There have been various definitions of the term *terminology*. In 1990, Sager extracted three readings of the term: "the set of practices and methods used for the collection, description and presentation of terms", "a theory, i.e. the set of premises, arguments and conclusions required for explaining the relationships between concepts and terms...", and "a vocabulary of a special subject field" (Sager 1990, p. 3). The current definition as given in the ISO standards, however, matches only Sager's third definition. According to ISO 1087, terminology is not the discipline, method or science, but "the set of designations belonging to one special language". The "science studying the structure, formation, development, usage and management of terminologies in various subject fields" is referred to as *terminology science*. (ISO CD 1087-1: 1997)
Termbanks (or *termbases*) are databases containing the vocabulary of a special subject field.
In most of the literature, the theory of terminology is squarely based on a simple model of the sign, that is, terminological theory is based on semiotics, the theory of signs. In the simplest case, the underlying semiotic structure is the relation between a term and a concept (in more mentalistic approaches), or between a term and an object (in more realistic approaches), or between a term as the 'keyword' or *definiendum* in a definition, and a textual description as its 'reading' or *definiens* (in more nominalistic approaches).
These three relations are generally combined in recent approaches: *concept*, *term*, *object* and *definition* are regarded as being equally important; the interrelations are exemplified by the concept pyramid shown in Figure 4.1 (cf. Suonuuti 1997). An extension to the area of multilingual terminology is visualised in Figure 4.2. In this idealised case, the concepts, objects and definitions remain constant; translation of definitions, of course, brings a new dimension of problems in this respect, which cannot be dealt with here. In current linguistic theory, a somewhat different sign model is used, as illustrated in Figure 4.3; see

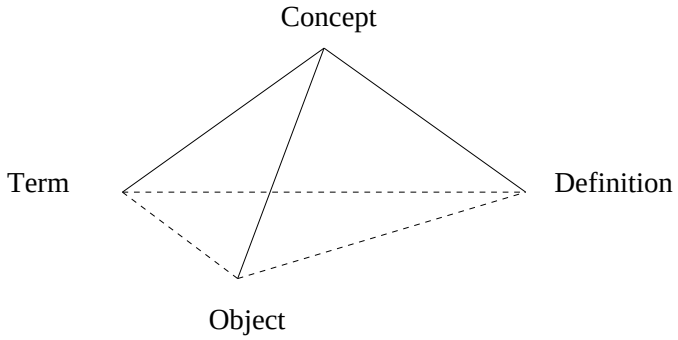

Figure 4.1: Extended semiotic triangle

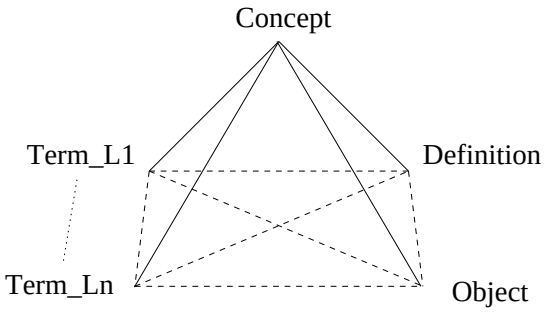

Figure 4.2: Semiotic pyramid for multilingual termbases

Gibbon (1999, forthcoming) for discussion.
Standard terminological theory assumes the existence of objects in the real world,[2] concepts are seen as mental images of such objects. It should be emphasised that these are heuristic assumptions, and not philosophical statements about the ontology of signs and objects; in practice, standard terminological theory thus adopts a mentalistic viewpoint.
However, many linguistic approaches to the study of vocabulary in general are non-mentalistic. For example, in the related fields of lexical and logical semantics the invocation of mental concepts is generally avoided. Objects are conceived simply as classes (families, fields, etc.), whose elements are individual object instances within space–time coordinates. An overview of lexicological and lexicographic terminology for spoken language systems is given in Chapter 6 of Gibbon et al. (1997). The terminology in this area may be clarified as follows:

1. Lexicography is the branch of applied linguistics concerned with the design

[2]The notion of *real world* is problematic because there are many concepts which refer to imaginary objects such as unicorns. Therefore it can be argued that the *real world* includes objects of fictional worlds as well.

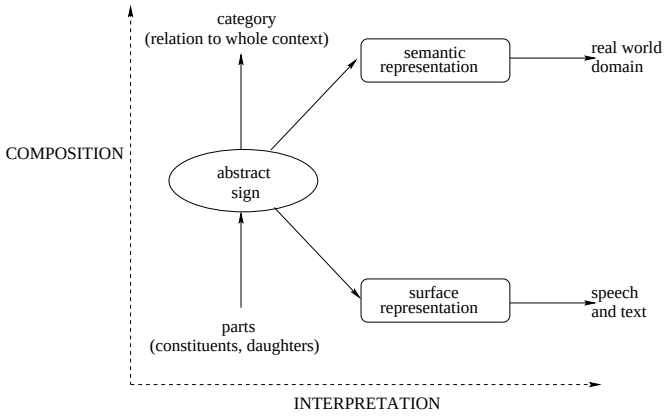

Figure 4.3: Sign model

and construction of electronic or paper lexica for practical use.

2. Lexicology is the branch of descriptive linguistics concerned with the theory and practice of describing types of lexical information, with the emphasis on lexical semantics, and collocational and idiomatic restrictions.
3. Lexicon theory is the branch of formal and computational linguistics and of psycholinguistics which is concerned with theoretical explanation, with formal modelling of lexical representation and processing.
4. Terminology science is the study of technical terms, i.e. the vocabulary of a technical sublanguage, in relation to objects in the real world and to concepts.[3]
5. Terminography is the applied discipline concerned with the creation of terminological reference works such as terminological dictionaries and databases (termbanks, termbases).[4]

The reasons for regarding lexicology and terminology, or lexicography and terminography as distinct disciplines are partly historical, partly in the difference between the mainly descriptive goals of lexicology and lexicography, and the mainly normative goals of terminology and terminography, and partly in the procedural emphasis on a form–based orientation in the former versus a concept–based orientation in the latter. An integrated overview of current work and terminology in the first three areas with respect to the human language technologies is given in Van Eynde and Gibbon (1999, forthcoming).
In practice, the terminographer in fact proceeds largely like the descriptive lexicographer when it comes to documenting existing terms, but with an additional dimension of systematisation and normalisation on the basis of conceptual hierarchies.

[3] According to ISO CD 1987–1: 1997, *terminology science* is the "science studying the structure, formation, development, usage and management of terminologies in various subject fields".

[4] According to ISO 1087–1, *terminography* is "part of terminology work concerned with the recording and presentation of terminological data", where *terminology work* is "work concerned with the systematic collection, description, processing and presentation of concepts and their designations".

In terminological theory, a *term* is regarded as the verbal representation of a concept, and a one-to-one correspondence between concepts and terms is assumed in the ideal case. The *definition* of a term is simultaneously the *verbal description* of a concept.
There are many different kinds of definition, any of which could in principle be used in different contexts for terminological definitions:[5]

- Ostensive definition: set of actual object instances in the class represented by a term.
- Definition by example: list of contexts containing uses of the term.
- Contextual definition: paraphrase of a term in use in a given context.
- Definition by prototype: provision of a model of a typical representative of the object class, perhaps as a visual (graphic) model.
- Definition by analogy: reference to similar known objects of the same general type.
- Definition by *genus proximum et differentia specifica*, i.e. by the nearest more general type and specific differences from other objects of the same general type.

The classical way of defining a concept is to give the *genus proximum*, i.e. the nearest general kind relating to the concept in question, and the *differentia specifica*, i.e. the properties that distinguish a concept from other items of the same general kind.
However, in most cases terminologies do not follow this strict pattern, but, as in lexicology, admit several kinds of definition, in terminology and lexicography, such as synonyms, paraphrases, examples, drawings, photographs, etc. (Sager (1990), p. 42–43, and Sager and L'Homme (1994)), which can easily be related to the characterisations given above.
Sager and L'Homme (1994) propose seven elements constituting the semantic specification of a concept:

1. the subject-field, or *domain*, e.g. 'Spoken Language Technology', 'mathematics', etc.
2. the concept class, e.g. 'material entity', 'abstract entity', 'activity', etc.
3. the *genus proximum*, which may stand in a *type-of relation* (also called *ISA relation*) or in a *PARTOF relation* to the defined concept, e.g. 'frequency' is a type of 'acoustic measure', or 'phonetics' is part of 'linguistics'.
4. the concept class which constitutes the *genus proximum*
5. the relationship between the defined concept and *genus proximum*, i.e. whether it is of the *ISA* ('type-of') or *PARTOF* kind.
6. the *differentia specifica*
7. any nonessential characteristic.

These elements are incorporated into the *microstructure* (see Sections 4.3.2 and 4.4.2) proposed for the EAGLET spoken language terminology specification introduced in Section 4.4.2.

[5] In terminology science two main kinds of definition are distinguished: the intensional definition and the extensional definition. The intensional definition describes the intension of a concept by stating the superordinate concept and the delimiting characteristics which differentiate the given concept from related concepts. The extensional definition enumerates all the specific features of a given concept. This reading of *intension* and *extension* differs from that used in linguistics, where the term *extension* refers to the object(s) in the real world while the *intension* refers to the semantic content of a lexical unit.

4.2.2 Relations between terms

The starting point for a terminological analysis is, ideally, the real world and its objects. A subject field is divided into subfields and these into subsubfields etc. until no further distinction can be made, and each field can be assigned to a concept. After the classification and hierarchy of concepts has been worked out, terms are assigned to these concepts.
In creating a lexical database, a lexicographer, on the other hand, adopts a different strategy by starting from terms and their definitions, usually providing detailed phonological, morphological and syntactic information, and arrives at placing these terms into a hierarchy characterised by either ISA (type-of) or PARTOF relations. In lexical semantics the ISA (type-of) relation is called *taxonomic relation* (or *taxonymic* relation), the PARTOF relation is called *mereonomic relation* (or *meronomic / meronymic relation*):

- Taxonomy: A hierarchy defined by the relation of generalisation and its inverse, specialisation; referred to in artificial intelligence as the *ISA* hierarchy. The *ISA* relation is perhaps the fundamental lexical relation. The term is rather general, and covers relations which have been referred to in other formalisms and theoretical frameworks with terms such as: *paradigmatic relation*, *classification*, *field*, *family*, *similarity*, *set partition*, *subset–set inclusion*, *element–set membership*, *generalisation*, *property*, *implication*, *inheritance*. Typical *ISA* relations define, in phonology, the natural classes characterised by distinctive feature vectors or by distributional classes based on syllable or word positions; in morphology, affix and stem classes; in phrasal syntax, parts of speech and constituent categories; in semantics, synonym, antonym and hyponym sets, or semantic fields.
- Mereonomy (meronomy): A hierarchy defined by the relation of parts to wholes, and parts to parts; often referred to as the *PARTOF* hierarchy. The *PARTOF* relation is the fundamental syntactic or combinatorial relation. Like *ISA*, the term is also rather general, and a wide range of different relations are covered by it in different approaches to linguistics in general and lexicography in particular: *syntagmatic relations*, *mereological (merological) / mereonomic (meronomic) relations*, *part–whole relations*, *part–part relations*, *(immediate) constituency / domination*, *command relations* (e.g. *c–command*), *dependency relations*, *government relations*, *argument structure*, *thematic role structure*, *subcategorisation frames*, *case frames*, *valency*, *anaphoric binding relations*, *categorial functor–argument application*, *concatenation*, *linear ordering*, *prosodic (autosegmental) association and precedence relations*, *child–child (sister) relations*, *parent–child (mother–daughter) relations*.

In terminological work, a different metaterminology is used:

- *logical concept hierarchy* (also called *generic concept hierarchy*), which constitutes the hierarchy of concepts holding an *ISA* relation, i.e. a taxonomy.
- *ontological hierarchy* (also called *partitive hierarchy*), constituting the hierarchy of concepts in a *PARTOF* relation, i.e. a mereonomy.

In artificial intelligence one speaks of *inheritance* when referring to taxonomies: subordinates share all attributes with their superordinates and consequently do not have to be specified for these properties but can *inherit* their shared properties from their immediate superordinate class, thus reducing redundancy

and saving storage room, and also ensuring the preservation of consistent information within the class. In predicate logic the relation corresponds to the operation of *implication*, in particular *material implication*.
Examples of a logical concept hierarchy (taxonomy) and an ontological hierarchy (mereonomy) are given in Figures 4.4 and 4.5.

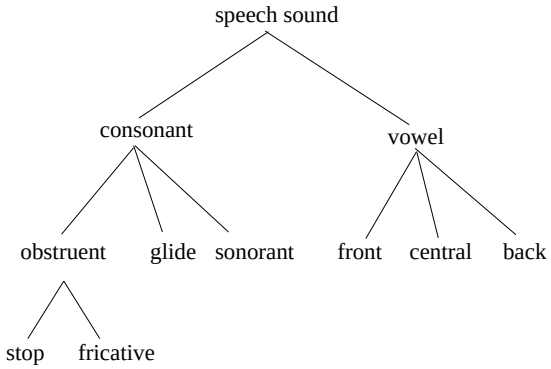

Figure 4.4: Example of a logical concept hierarchy

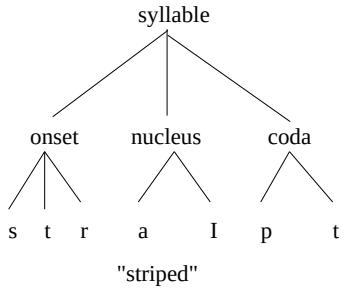

Figure 4.5: Example of an ontological hierarchy

The term 'fricative' represents a concept which is subordinate to the concepts represented by the terms 'obstruent' and 'consonant' in the logical concept hierarchy because *A fricative is a type of obstruent, an obstruent is a type of consonant, and therefore (by the transitivity of the implication operation) a fricative is a type of consonant.* is an acceptable sentence (see Cruse (1986) for this kind of lexical semantic test technique). Analogously, the term 'onset' is subordinate to the term 'syllable' in the ontological hierarchy, because *An onset is part of a syllable* is an acceptable sentence. These general criteria apply to all areas of terminology, of course, not just to the spoken language term models illustrated here.
A problem with the single hierarchy model illustrated here is that terms may be simultaneously located within several hierarchies, depending on *sorts of prop-*

erty, often represented in terms of *attributes* which take their values from groups of related properties. For example, 'consonant' is located not only in the *manner of articulation* hierarchy illustrated in the Figure, but also in at least one *place of articulation* or *feature geometry* hierarchy. Likewise, 'vowel' is located in a conceptual space consisting of at least four independent (with minor exceptions) place of articulation dimensions: front–back, open–close, rounded–unrounded, and nasal–oral. The ontological hierarchy is also not as simple as the Figure shows, because, in addition to the temporal *precedence* relation a temporal *overlap* or *parallelism* relation is required (to describe co–occurrence of features, and prosody).

A distinction was drawn above between the terminological and the lexicological strategies for creating termbanks/lexical databases: termbanks tend to be concept oriented, lexical databases headword-oriented. In real applications, however, one finds both forms combined resulting in termbanks including lexicographical information linked to concept entries. These termbanks have been called *lex-termbases* (Melby and Wright 1998). The design of such a database is the goal of the MARCLIF project (MAchine-Readable Concept- and Lexicographically oriented Interchange Format). The EAGLET term database can also be regarded as a lex-termbase as it is very much headword-oriented (in its present state) and includes detailed morphological and phonological information. Also the labels of the data categories used for describing terms reflect the linguistic, lexical semantic background. EAGLET is organised primarily by the orthographic forms of terms and not by the general concept hierarchies for spoken language technology.

The preceding discussion has shown that the description of terminology has a number of dimensions, some of which are relatively independent of lexicology and lexical semantics on the one hand, and lexicography on the other. The characteristic features which are more central in terminology and terminography than in the related linguistic disciplines will be taken to be the following:

1. Term systematisation, normalisation, and standardisation.
2. Domain and task–oriented term creation.
3. Systematic word–creation based on morphological analysis (word formation, including derivation and compounding; to a lesser extent inflection).
4. Use of lexicological analysis including colligation (the cooccurrence restrictions on parts of speech, POS), collocation (the cooccurrence restrictions on individual words, whether semantic or purely idiomatic).
5. Use of lexicographic methods (including computational corpus analysis, lexical database construction and access).

The first three points characterise the *normative* or *prescriptive* aspects of terminology, while the last three (morphology overlaps both areas) characterise its *descriptive* aspects.

4.3 The organisation of terminology

4.3.1 The onomasiological and semasiological perspectives

The types of lexical information are more complex and varied than simple dyadic sign models suggest. Ignoring practical aspects such as documentation of the term gathering process, and concentrating on the lexicographic component of

terminographic work, lexical objects have in general terms at least the following kinds of property:

1. *Lemmata* (lexical access keys, as opposed to *abstract lemma*, which may be numerical)
2. Words (including terms)
3. Morphemes, morph variants
4. Word classes (semantic, syntactic, morphological fields/types)
5. Representations:
    1. Phonological
    2. Orthographic
    3. Semantic
    4. Conceptual

The organisation of a lexicon or terminology can in principle have any of these objects as its primary search category. However, traditionally just two main structural principles are generally recognised in lexicography and *a fortiori* in the lexicographic aspects of terminology:

1. The semasiological principle. Semasiological organisation is based on the forms of words, that is, either on their orthography (the most usual choice) or on their phonology. In many languages there are complex relations between the orthographic variants of a word, many of which are mediated by the morphology of the language; the same applies to phonological variation. Standard alphabetic dictionaries are the most straightforward example of semasiological organisation; the choice of lemma or headword, and sub–lemmata, is determined on morphological grounds.
2. The onomasiological principle. The grouping of words in terms of their meaning, i.e. on conceptual or semantic grounds, is the second main principle of lexicon organisation. Onomasiological organisation is based on semantic fields or on shared properties of meaning. These semantic fields are hierarchically organised in terms of the two basic types of relation 'taxonomy' and 'meronomy' mentioned above:
    1. Taxonomy: the *type-of* or *ISA* relation of semantic networks, defining elements of sets, and subsets of supersets.[6][7]
    2. Mereonomy (meronomy): the *PARTOF* relation of semantic networks, defining parts of wholes, and sub–aggregates of aggregates.

    Within the onomasiological principle, strictly speaking two further criteria can be distinguished:

    1. Conceptual. Semantic objects are defined as language–independent.
    2. Relational. Semantic objects are defined as language–specific.

    For terminological work, the first assumption, that semantic objects are language–independent, is a useful idealisation; the second assumption is, however, often closer to actual practice and to the complex meanings of words in different languages.
    For multi–lingual terminology, a set of relations based on the frequently cited *Vauquois triangle* (see Figure 4.6) can be defined: the peak of the triangle

[6] Closely related to the *ISA* relation is, in conventional artificial intelligence parlance, the *HASPROP* relation, which assigns the shared properties which characterise sets of objects which enter into *ISA* relations.

[7] For a discussion of a relational semantic network of linguistic terminology see Lehmann (1996).

represents universal concepts; the levels below this represent language–specific notions which need to be connected via explicit transfer relations (transfer rules).

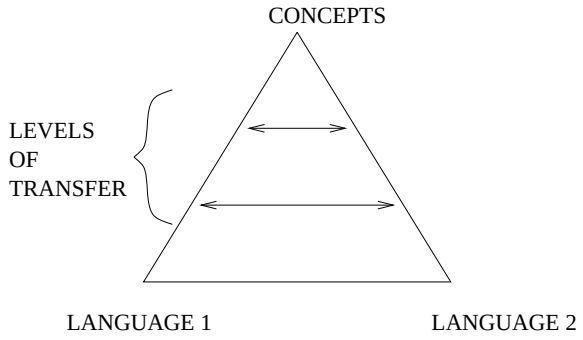

Figure 4.6: Vauquois triangle for terminology

The Vauquois triangle model can easily be related to the sign models discussed in Section 4.2; it corresponds to one side of the extended pyramid model of Figure 4.2.
The triangular model is also useful as a visualisation of the relation between *special languages* or *technical sublanguages* as opposed to *general everyday language* as used in informal speech and writing, and also as a visualisation of the relation between the languages of different disciplines which refer to the same objects in the real world or to the same concepts, but in a different terminological framework. A good example of the second kind of relation is the *finite state machine (FSM)* of computer science, most generally referred to in contemporary computational linguistics as a *finite state transducer (FST)*; a probabilistic variety of FSM is referred to in pattern recognition and spoken language engineering as a *Hidden Markov Model (HMM)*. The formal basis for formalising the objects referred to by these terms is the same; the differences are differences in the terminological traditions of different disciplines.
In deciding on a terminological model, fundamental decisions on these points have to be made.

4.3.2 Terminological macrostructures and microstructures

As mentioned above, there are two basic semiotically determined strategies for organising access to a termbank:

1. semasiological (according to the form of terms, i.e. via alphabetically ordered orthography), or
2. onomasiological (according to the meaning, concepts, or models of real–world objects).

A semasiological approach has the advantage that it is easy to automatise the process of ordering. The disadvantage can be seen in the complexity of definitions required for families of related terms.

An individual entry has to be structured according to a set of types of lexical information, or data categories: the *microstrucure*. A lexical microstructure is a specification of the attributes according to which a lexical entry in a terminological or other lexicon can be partially or fully specified. Formally, a microstructure (or any combination of the attributes it embodies) is a partial function assigning properties to lexical entries. More concretely, the microstructure determines the content of a lexical entry and thereby provides a definition for the lexical entry in terms of its properties. In its implementation as a database, a microstructure may be seen as a vector of information types, corresponding to a database record structure. The microstructure is dependent on the *macrostructure* in the sense of needing more specific information within the context of an onomasiological macrostructure for the whole termbank and its subject field. The types of lexical information may include definitions, examples of usage, formulas, pictures, spelling, morphological and syntactic categories and structure.
A possible source of terminological confusion lies in the term 'data' itself:

1. Lexicologists, linguists, speech engineers habitually refer to texts and speech recordings, sometimes also to transcriptions, as data.
2. Lexicographers (and terminologists) habitually refer to the readings of lexical or term entries as 'data'.
3. Database specialists and users refer to any contents of database tables and any inputs to programmes as 'data'.

The lexicographic and terminological usage is more closely related to the database usage than to the linguistic, lexicological and speech engineering usage.
For lexicologists, linguists and speech engineers, the readings of lexical or term entries are far removed from text or speech data, and are the result of implicit or explicit theory–construction. A lexicon, in this context, is part of an overall theory or of an implemented model for a theory.

4.4 Spoken Language terminology

4.4.1 The hybrid character of SL terminology

Terminology for human language engineering and related fields such as linguistics and phonetics, both for written text and for speech, is determined both by theoretical research and by pre–competitive development considerations. There is currently no effective standardisation of any practical consequence, except in the sub–fields which are related to electrical engineering, signal processing, and acoustics.
It is clearly out of the question to try to harmonise theoretical terminologies; for very good logical reasons theoretical terms cannot be isolated and defined in the absence of their theoretical environment. A practical procedure is to analyse the actual vocabulary used by experts in applications oriented fields. However, experts from many different disciplines are involved: from speech sciences (including clinical phonetics on the one hand, and orthographic transcription and signal labelling on the other) through descriptive linguistics (phonology, morphology, syntax, semantics and pragmatics), and computational linguistics (linguistic data acquisition, lexicon construction, parsing) to electrical engineering and product evaluation methodology. This heterogeneity forebodes

non–consistent terminological models from the different disciplines which cannot be naively tossed together into a single pool without the danger of confusion and contradiction.
However, since the mid–nineties, an increasing number of sophisticated spoken language consumer products have appeared on the mass market, including dictation systems (notably those of Dragon Systems, IBM, Kurzweil and Philips), and numerous readback (Text–To–Speech, TTS) systems. These developments are likely, in the mid–term, to drive the standardisation of terminology as in other fields with mature product development and marketing.

4.4.2 Toward a microstructure for SL terminology

4.4.2.1 Types of lexical information

A general outline of the kinds of terminological and lexicographic information which may be required in a termbank is shown in the following microstructure. Note that non–conventional information types are included, for example pronunciation information in case the termbank is used with readback access software by blind or otherwise handicapped users.

1. Surface properties:
    1. Spelling (orthographic representation)
    2. Pronunciation (phonological representation)
    3. Inflectional paradigm class (inflectional morphology)
2. Structural properties:
    1. Part of Speech (specifications of varying granularity)
    2. Word formation type (derivation, compound, phrasal compound)
    3. Word formation structure (stems, affixes, roots)
    4. Specification of the degree of lexicalisation in terms of
        1. partial compositionality ('frozenness'),
        2. contextual restrictions (colligation (subcategorisation), collocation, idiomaticity).
3. Content properties:
    1. Definition by *genus proximum et differentia specifica*
    2. Characterisation by semantic relations to other entries
    3. Specification in terms of semantic components
    4. Formal term definition as a structure or equation
    5. Interpretation function in a conceptual/material model
    6. Idiosyncratic constraints
    7. Extensional prototypes:
        1. Graphical representation of semantic relations
        2. Graphical representation of conceptual or material model
        3. Audio prototype for sound (noise or pronunciation)
4. Usage
    1. Prototypic examples
    2. Attested corpus source occurrences
5. Status
    1. Identity of lexicographer
    2. Version history

3. Standardisation status
4. Recommendations

4.4.2.2 More generic microstructures

A synthesis of data categories corresponding to fields in records of a relational terminological database, and based on ISO 1087 and ISO 12620, is introduced in Schmitz et al. (1994) and has been adapted for present purposes.
The categories have been re–grouped and further categorised in order to provide more structure for designing appropriate database record structures.
The overall microstructure vector is presented below as field–specific subvectors, also referred to as microstructures (strictly: submicrostructures). In formal and computational linguistic terms, a microstructure can clearly be formalised as an attribute–value structure, and therefore used in the context of current unification–oriented systems on the one hand, or represented with SGML markup on the other (see also Section 4.6.5).
The following descriptions are taken from ISO 12620; they are reprinted in Melby and Wright (1998).

4.4.2.2.1 TERMS substructure

TERM: Designation of a defined concept in a special language by a linguistic expression.

MAIN ENTRY TERM, HEAD TERM: Any designation of a concept heading a terminological record.

4.4.2.2.2 MEANING substructure

SYNONYM: Term that represents the same concept as the main entry term in a term entry.

QUASI–SYNONYM: Term that represents the same concept as another term in the same language, but for which interchangeability is limited to some contexts and inapplicable in others.

INTERNATIONAL SCIENTIFIC TERM: Term is part of an international scientific nomenclature as adopted by an appropriate scientific body.

TERM/CONCEPT RELATION: Characteristic of the term/concept assignment indicating its relative degree of ambiguity.

DEGREE OF EQUIVALENCE: Extent to which the concept associated with a designation in L1[8] covers the same characteristics as the concept associated with a designation L2.[9] The degree of equivalence can include:

- smaller $<$
- equivalent $=$
- approximately equivalent $\approx$
- larger $>$
- non–equivalent $\neq$

DEFINITION: Statement that describes a concept and permits its differentiation from other concepts within a system of concepts.

[8] L1 = language 1
[9] L2 = language 2

EXPLANATION: Statement that describes a concept, but does not adequately permit its differentiation from other concepts within its system of concepts.

CONTEXT, EXAMPLE (DEPRECATED): Text or part of a text in which a term occurs.

CONCEPT RELATION: Link between two or more concepts. Types of concept relation can include:

- generic relation
- partitive relation
- sequential relation
- temporal relation
- spatial relation
- pragmatic relation
- ...

DOMAIN, SUBJECT FIELD, SUBJECT LABEL: Field of human knowledge to which a terminological record is assigned.

CLASSIFICATION: Arrangement of concepts into classes and their subdivisions to express the relations between them; the classes are represented by means of a notation. Types of classification elements can include:

- class
- notation
- thesaurus descriptor
- non–descriptor
- keyword
- indexing term

THESAURUS DESCRIPTOR: Term in a thesaurus that may be used to represent a concept in a document or in a request for retrieval.

KEYWORD: Word or group of words, possibly in lexicographically standardized form, taken out of a title or of the text of a document characterizing its content and enabling its retrieval.

INDEXING TERM, INDEX WORD: Term used to designate a concept in an index.

4.4.2.2.3 FORM substructure

VARIANT: One of the different forms of a term, including spelling variants, morphological variants and syntactical variants.

SYMBOL: Designation of a concept by letters, numerals, pictogrammes or a combination thereof.

ABBREVIATED FORM: Term resulting from the omission of any part of a term while designating the same concept. Types of abbreviated forms can include:

- short form
- abbreviation
- initialism
- acronym
- clipped term

GRAMMAR: Grammatical information about a term. Types of grammar can include:

- part of speech
- gender
- grammatical number
- singular
- dual

- plural
- voice
- principal parts
- inflection
- animate

PHRASEOLOGICAL UNIT: Any group of two or more words that form a sense unit. Types of phraseological unit can include:

- collocation
- set phrase
- standard text

TERM FORMATION: Classification of a term according to the methodology employed in creating the term. Types of term formation can include:

- borrowed term
- loan term
- barbarism
- loan translation
- false calque
- paraphrase
- neologism

4.4.2.2.4 USAGE substructure

TERMINOLOGY ACCEPTABILITY RATING, TERM STATUS: One of a set of codes indicating the usage status of a term. Types of terminology acceptability rating can include:

- standardized item
- preferred term
- admitted term
- deprecated term
- superseded term
- rare term
- recommended term
- suggested term
- non–standardized term
- new term

REGISTER: Classification indicating the relative level of language individually assigned to a lexeme or term or to a text type. Types of register can include:

- stilted register
- formal register
- technical register
- neutral (standard) register
- colloquial register
- slang register
- vulgar register
- in–house register
- bench–level register
- intimate register
- literary register

RESTRICTION: Category of factors that limit the usage of a term. Types of restrictions can include:

- trademark
- trade name

GEOGRAPHICAL USAGE: Term usage reflecting regional differences.

TIME RESTRICTION: Indication of a period of time during which a term was subject to special usage.

TRANSFER COMMENT: Note included in a term entry indicating the degree of equivalence, directionality or other special features affecting equivalence between an L1 term and an L2 term.

DIRECTIONALITY: Property of multilingual equivalent terms indicating whether a similar degree of equivalence exists when moving from L1 to L2 as when moving from L2 to L1.

RELIABILITY CODE: Code assigned to a data element of record indicating adjudged accuracy and completeness.

4.4.2.2.5 OTHER substructure

NOTE, REMARK, COMMENT: Statement that provides further information on any part of the terminological record. Types of notes can include:

- example
- suggestion
- usage note
- table
- figure
- formula
- unit
- range

4.4.2.2.6 PRODUCTION substructure

OWNER SUBSET: Code used to identify a terminology entry as associated with a specific terminologist.

APPLICATION SUBSET: Code used to identify a terminology entry as associated with a specific application.

ENVIRONMENT SUBSET: Code used to identify a terminology entry as associated with a specific application.

CUSTOMER SUBSET: Code used to identify a terminological record as associated with a specific customer.

PROJECT SUBSET: Code assigned to a specific project with which a term, record or entry is associated.

PRODUCT SUBSET: Code assigned to a product to which a term is related.

SECURITY SUBSET: In–house security classification of a term.

TRANSACTION EVENT: Occurrence assoicated with the management of a database. Types of transaction can include:

- creation
- update
- check
- approval
- withdrawal
- standardization

DATE: Point of time at which a transaction or event takes place. Types of date can include:

- creation date
- update date
- check date
- approval date
- withdrawal date
- standardization date

DATE RESPONSIBILITY: Code for identifying individual entering, checking or signing off on a field or record. Types of responsibility can include:

- creator
- updater
- checker
- approver
- user
- subset owner

STANDARDIZATION STATUS: Status of a term submitted or selected for standardization. Types of standardization can include:

- committee status
- organizational status
- process status

ENTRY STATUS: Code indicating the level of completeness and accuracy of an entry within a terminological database.

4.4.2.2.7 SOURCES substructure

BIBLIOGRAPHIC DATA: Data that describe and uniquely identify documents.

BIBLIOGRAPHICAL DATA ITEM: Standard data element type included in a bibliographic entry. Types of bibliographical data items include:

- author
- editor
- title
- place of publication
- publisher
- location of document
- volume
- issue
- edition
- date of publication
- page
- ISBN number
- ISSN number

4.4.2.2.8 MACROSTRUCTURE substructure

CROSS–REFERENCE, CROSS–REFERENCE TYPE: Category or pointer field or record used in a database for navigation (chaining or jumping) to another related location, e.g. another record. Types of cross-references can include:

- concept system cross–reference
- term status cross–reference
- synonym cross–reference
- see cross–reference
- see also cross–reference
- abbreviated form cross-reference
- full–form cross–reference
- equivalent cross–reference
- antonym cross–reference
- bibliographic cross–reference
- terminological source code
- responsibility cross–reference

4.4.3 Recommendations on termbank development

The recommendations are structured according to the three main phases of a classical software engineering model.

Requirement specifications:

1. Describe the potential users of the termbank (engineers, researchers and developers in other fields, schooling personnel and trainees, marketing specialists, management, translators and documentation specialists).
2. Describe the uses of the termbank in relation to user groups and their environments (available equipment, educational, commercial or industrial institution, integration with other lexica, translation memories).
3. Define the domain of the termbank, and thus its content, in terms of hierarchies of concepts to be covered.
4. Determine whether a monolingual, bilingual or multilingual termbank is required, and ensure the availability of the appropriate morphological information about each language (inflection, word formation).
5. Adhere to the ISO standards insofar as they are relevant to your domain and application.

Design:

1. Select the appropriate macrostructure, i.e. with onomasiological organisation (access via concepts, term forms as search targets), or with semasiological organisation (access via forms, concepts as search targets), or a lex–termbase or other variety of hybrid termbase.
2. Construct the *ISA* and *PARTOF* hierarchies with the domain–relevant concepts.
3. Design the appropriate microstructure, with types of lexical information, domain information, and version information.
4. Design the navigation (access) strategies in terms of lookup, search and (in a terminological hyperlexicon) cross–referencing link structures.
5. Plan data (in the sense of termbank entry) acquisition and maintenance logistics.
6. Design ergonomic user interface and networking features in relation to the user groups and team structures.

Implementation:

1. Decide whether to use a commercially available termbank or to adapt an existing database management system to your requirements.

2. If you decide on a commercially available termbank, ensure that your computing environment will support it adequately, that maintenance support is available for the duration of the termbank life–cycle, and that qualified staff are available.
3. Employ staff who are qualified both in terms of the domain and in terms of the linguistic and terminological basics of the languages concerned, and whose access rights for modifying the termbank are well–defined and supervised.
4. In adding entries to the termbank, ensure that appropriate information sources are available (text corpora, other termbanks, experts for consultation), and state the sources of terms.
5. Ensure that the termbank provides support, for instance in the following way: "Termbank software should provide a facility for prompting terminologists when building up terminological records. The greater breadth of data required for each entry and the construction of a complex network of conceptual relationships means that some form of expert system is required to control the work of terminologists." (Sager 1990, p. 154).
6. Ensure that the termbank provides overview functions and visualisations of structures and interrelations in the database, as well as optimal reaction times in interaction with the terminologist user and the end user.
7. The termbank should provide different modes of access and flexible implementation of navigation strategies for onomasiological, semasiological and other search strategies.
8. With respect to shared references and links, avoid redundancies for these reasons (Melby and Wright 1998):

   - Economy of storage
   - Consistency: Updating and correcting of database entries is less time consuming and consistency is maintained. Otherwise all occurrences of an entry would have to be updated or corrected, which is much more error prone and easily leads to inconsistencies.

9. Consult specialists in the field at all stages of the process — they are also your clients.

4.4.4 Recommendations for further reading

Introductory works on terminology: Felber and Budin (1989), Arntz and Picht (1989), Helmut Felber (1979), Sager (1990), Wüster (1991). For further reading on terminology consult Schmitz et al. (1994), Dutz (1985), Budin (1996), Suonuuti (1997), Wright and Budin (1997), Schmitz (1998) and Schmitz (1997). The second volume of the *Handbook of Terminology management* by Wright and Budin is going to be published in 1999. For further reading on lexical semantics the reader is referred to Lyons (1977) and Cruse (1986).
A detailed analysis of the terminology of linguistics in terms of ISA relations can be found in Lehmann (1996). This approach focusses on the logic status of each term and on the ISA relations identified for linguistic terminology. Lehmann discusses 10 subordinating relations ("x is a y", "x is a class of y", "x is an aspect/a feature of y", etc.) and two cross referencing relations ("x is related to y", "Lemma of x is y"). Meronymic or PARTOF relations are dealt with to a minor extent.

4.5 Relational databases

If a term database for spoken language is to provide both onomasiological and the semasiological query perspectives, the database system has to fulfil basic constraints with respect to at least the following two general criteria:

- database model: the choice of microstructure vector or vectors;
- representation of terms and term relations: the specification of data types for entries in the microstructure vector positions, i.e. in the database fields.

All terms should be uniquely accessible through a commonly shared key attribute. In general, the key attribute is defined as the orthographic representation; however, conceptually the two attributes are distinct. In cases of homography, for example, two entries may be required, each with the same orthography; if the orthography were simultaneously the key, then the key would not uniquely distinguish the entries.
The key is defined more generally as a *key root* which is identical to the orthographic representation, and a *key extension*, which is taken from a set of numerals (usually consecutive numerals starting with '001').
A conventional representation format for a terminological database or termbank is the *relational database*. The *relation* is a table whose structure (columns) represents the microstructure of the terminology, and whose records (rows) represent terms and the property values for each data category.

4.5.1 Components of a relational database

1. Architectural model: Definition of the database type, in the case of the EAGLET database Codd's (see Codd 1970) relational database model, and its relations.
2. Database engine: Definition of the algorithms for access to the relation, as a specification for the software that implements the model.
3. Front end tools: The view of the database and database access which is available to the user (client).
4. Normalisation rules: The data (in the sense of 'database entries') are entered according to a specific syntactic definition of the database relation.

4.5.2 Structures in the relational model

The relational database model contains the following specific relations:

1. Table (File, Relation)
2. Record (Row, Tuple)
3. Field (Column, Attribute)

A relational database can be visualised as a table (see Table 4.1).
The relations between the records ('*vcards*' for '*virtual term record cards*') are expressed by *key mapping operations*:

- *candidate key:* any unique identifier for a record.
- *primary key:* the candidate key you choose to use.
- *compound key:* a key made up of more than one field
- *foreign key:* a field in one table that points to a record in a different table.

The different interlinked relations in a complex relational database can be represented by means of an 'entity–relationship diagramme'.

Table 4.1: Relational database visualised as a table

	Col_{key}	Col_1	Col_2	...	Col_n
$Record_1$	...	...	...	...	...
$Record_2$	...	...	...	...	...
...	...	...	...	...	...
$Record_m$	...	...	...	...	...

4.5.3 Codd's definition of a relational database system

A set of qualitative criteria for the specification and evaluation of terminological databases is provided by a simplified version of Codd's definition of a relational database system (see Codd 1970).

- All information must be represented explicitly in one and only one way: as values in tables.
- Each and every datum in the database must be accessible by specifying a table name, a column name, and a primary key.
- Null Values must be treated systematically.
- At least one character-based language must be provided which can be used to modify the structure and contents of the database. This language must be able to do all of the following:
  - data definition
  - view definition
  - data manipulation
  - integrity constraints
- At least two integrity constraints need be supported:
  - No primary key can have a null value, and no primary key can be duplicated.
  - For every non-null foreign key there must exist a matching primary key.
- Although non-relational tools are allowed, it must be possible to manage the database using only relational tools, and the non-relational tools must not be allowed to bypass the integrity constraints imposed by the relational tools.

4.5.4 Query language

In a database system the phrase *query* subsumes all actions manipulating the DB (i.e. any method or function like adding, deleting, searching, etc.).
The most common language to send queries to a relational database as defined by Codd is the *Standard Query Language (SQL)* described in the *ISO/IEC 9075:1992* Norm. This standard is implemented in nearly every relational database application today.

4.5.5 Software implementations

For academic purposes a popular relational database (database server) is the *mSQL Server* from *Hughes Technologies*, Australia.

Though it only provides a subset of *ISO/ANSI SQL* (a kind of SQL light since it tries to cover the whole range of everyday uses) is free of charge for universities and researchers. For development purposes, miniSQL is suitable; however, for dissemination in an industrial context this software is clearly not suitable.
Distributed use of a central terminological database via the WWW has several advantages, centered on the *version integrity criterion*: rather than maintaining several copies of the database, a single copy accessible via the web provides a straightforward guarantee of authenticity and version integrity simply by means of token–identity at the central server for all clients.
However, the use of a single server for a multi-user multi-query system can affect performance. The effects depend on actual demand and can only be empirically determined.

4.5.6 Distribution of data generation over time

When setting up a database system one of the most important questions next to structuring the tables and records is how the raw data material is represented, i.e. stored, in the database.
Usually some DB attributes are of a static nature (e.g. text strings like the orthographic form), some are more abstract (e.g. the hyperlink relations in-between the term set or with entities locally or somewhere out there in the WWW) and others may be of a highly dynamical type (e.g. video or sound samples).
The provision of static information is in principle easy (though often expensive) to generate *a priori*, for example by using filters to scan given sources). Other types of hyperlink relation or multimedial event may have to be generated on demand as a response to the client's query (often called '*on-the-fly*' generation). The database system consequently has to provide several different methods for generating, updating and storage of term information.

4.5.7 Distribution of data generation over resources

Generating data can be very time and resource intensive. Two strategies may be considered for handling this. First, database entry acquisition may be distributed, as far as possible, not only over time but also between the involved parties. Second, acquisition can be fully or partially automated, and may include generic search and processing facilities for term and term usage extraction.
Building user interfaces for the World Wide Web means using the *Hypertext Markup Language (HTML)* to write the needed WWW Pages (Hyperdocuments). But *HTML* documents are of a fairly static nature. You can not implement any kind of really interactive interface between the user (client) and the DB System (server) by only using *HTML*.
The mechanism of encoding '*electronic forms*' in *HTML* only permits sending back *feedback information* in the sense of unvalidated client input. The syntax validation and semantic interpretation of this feedback has to be performed by the server.
This implies that there are at least three interactions between client and server to correct a syntactically or semantically wrong client query:

1. Server sending the electronic form.

2. Client sending back the user's input for that form.
3. Server validating client's feedback.

In case of errors start over at 1.

To overcome this WWW standard problem (i.e. validation of user input) *Netscape* has invented the supplementary WWW language *'JavaScript'* that provides an easy access to *HTML* forms while being resident on client computers.
With it the interface designer can distribute all input validation mechanisms to the client machine. In an ideal world scenario the feedback will then be 100 % correct when sent back to the server.
This reduces the server load in two ways:

1. The server does not have to compute the input validation (apart from some worst case traps).
2. It does not have to send back a correction request (including the interface again).

4.5.8 Required system components

As discussed in the previous sections the following modules are needed to build up a relational database system with an ergonomical client interface:

1. Database server
2. Filters to import external data
3. Database query language
4. Interface application
5. Interface programming language
6. Interface validation language

4.6 Terminology Management Systems (TMSs), databases, and interchange formats

Specific tools for terminology management on PCs have been on the market since about 1985. This development started in the 1960s with Main Frame Term Banks (TEAM, Siemens AG Germany; LEXIS, Federal Language Office, Germany; EURODICAUTOM, EU Commission) and Mini Computer Term Software (Danterm, Copenhagen University/ Business School). Today most programs designed for PC networks are based on Windows, for example TermStar, MultiTerm and Danterm.

4.6.1 MultiTerm

With the software MultiTerm produced and sold by Trados a well established and easy to handle Terminology Management System (TMS) exists. The current version is *MultiTerm ´95 Plus!* with interesting features:

- Data categories can be defined according to users' needs.
- Crosslinking is possible by hyperlinks. This is done by inserting a function 'hyperlink' at the appearance of the synonym or term in the entry. A click on the hyperlink will result in a new search for the highlighted term in the database.

- Trados have special discounts for universities and other academic institutions.
- A Web interface (*MuWi) is available for MultiTerm enabeling access to MultiTerm-Databases via the WorldWideWeb. Access to MultiTerm Databases via the Web interface is restricted to database queries; editing, insertion and maintainance is not intended.*
- *In a future version MultiTerm will support MARTIF (ISO 12200) im- and export (see Section 4.6.5, working on an SGML database basis.*
- *Search routines are easy to implement; some search routines are already available.*
- *Data import will be possible from other systems.*
- *"MultiTerm '95 Plus is a free-format text database useful for terminology databases, address databases, and document management systems." Each entry can consist of up to 32,000 characters and contains up to 500 fields each of which can be 4096 characters long.*
- *It is concept oriented; one entry corresponds to one concept.*
- *A concept can be represented by terms in 20 languages. Search is possible for any of these.*
- *Besides the term entry, MultiTerm allows additional free-text information for every term in so-called text-fields, e.g. definition, context, examples, etc.*
- *Another type of information is added to a term entry via attribute field. Here the user can choose items from a pick list, e.g. subject field (domain), source, clients, etc.*
- *Graphics can be linked to an entry.*
- *Wildcard can be included in the search.*
- *Links to related terms are available.*
- *Problems:*
  - *MultiTerm does not work platform independently; it runs only with Microsoft Windows 95 (and NT 4.0).*
  - *Although there is WWW accessibility to MultiTerm databases there is no insertion interface. This is not necessarily a major problem because terms and definitions from outsiders need reviewing before insertion. This could easily be done by forms using CGI (Common Gateway Interface), producing an easy to import format for a MultiTerm interface.*

4.6.2 ITU Telecommunication Terminology Database: TERMITE

TERMITE is a multilingual (English, French, Spanish, Russian) database provided by the ITU (International Telecommunication Union), Geneva, Switzerland (see TERMITE 1999). Within this organisation governments and private companies coordinate telecom networks and services worldwide.
The Terminology, References and Computer Aids to Translation Section updates and maintains the database mainly on the basis of ITU printed glossaries published since 1980. Adding to those terms entries which relate to more recent activities of ITU are included.
The main target group of TERMITE are translators and users working in the field of telecommunications.
TERMITE has the following features:

Each of the four languages can be chosen as source language. The search entries includes the full term, a keyword included in the term itself, wildcards followed

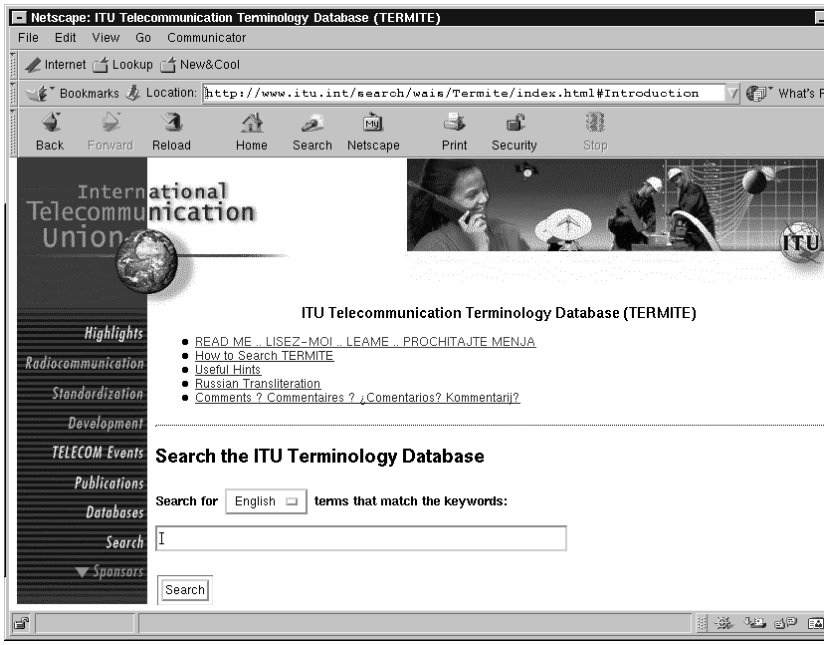

Figure 4.7: User interface of the TERMITE database

by an asterisk, abbreviations, and acronyms. The order of the keywords is not important for the search.
TERMITE generates a numbered list of terms. It includes administrative information about the year and month of the last update, the serial number of the term, and the languages in which the term is presented to the user. The entry which fits best the search keyword is listed at the top.
After the selection of one term, the full entry is displayed and structured as listed below:

- serial number
- last update
- term in source language
- abbreviation
- source
- synonyms
- term in other languages (alphabetical order: French, Spanish, Russian)
- abbreviation
- source
- synonyms

The entries in Russian are transliterated. This transliteration has to be used for searches if the source language is Russian. ITU provides an online transliteration table.
Critique:
As the user group consists of people who have not the same knowledge of the

scientific field of communication definitions and the context in which a term is used should be included in the entry for a better understanding.

4.6.3 TERMIUM – Canadian Linguistic Data Bank

TERMIUM is a terminology database installed and maintained by the Translation Bureau (see TERMIUM 1999). This agency primarily provides services (translation and linguistic services, interpretation services and terminology services) to the Canadian federal government. TERMIUM is an English - French, French - English electronic dictionary available on CD-ROM for Windows, DOS and Macintosh. A Windows demo-version can be downloaded ("`http://www.translationbureau.gc.ca/demo-e.htm`").
It consists of over 3 million terms, their definitions, contexts, examples of usages, and administrative data. Additional to these basic types of information an entry can also include grammatical or stylistic information, equivalents for federal government abbreviations, titles of documents, acts and regulations, and other information important in a governmental context.
The terms belong to a wide range of subject fields (28: for example: Arts, Electricity, Informatics, Mathematics and Physics, Natural Sciences, Telecommunications and Postal Service) a list of which is included in the online description of the database.
Therefore and because of the various data categories for each entry it is a very valuable source of information for translators, terminologists, as well as for people who write in English or French.
The information displayed in the entry is made up of two databases: a linguistic and a sources database. The linguistic one consists of the terms, synonyms, variants, abbreviations, definitions, contexts, proper names, and translation problems. The sources database includes the documentary records or bibliograhpic information.

4.6.3.1 The structure of a TERMIUM entry

In a table consisting of two columns the English information is provided on the left hand side, accordingly the French information is given on the right hand side.
Each entry consists of the data categories listed below:

1. Subject field(s)
2. Entry block: term, grammatical information, synonyms, variants, abbreviations
3. Textual support: definitions, examples, phraseology
4. Source block: sources of information

Not every entry contains all the listed data categories. In general, the following information is presented:

- the subject field or field to which this [the term meaning]meaning relates,
- the terms that denote this meaning,
- the texts that explain the meaning examined on the record,
- the (decodable) codes identifying the documents used to prepare the record.

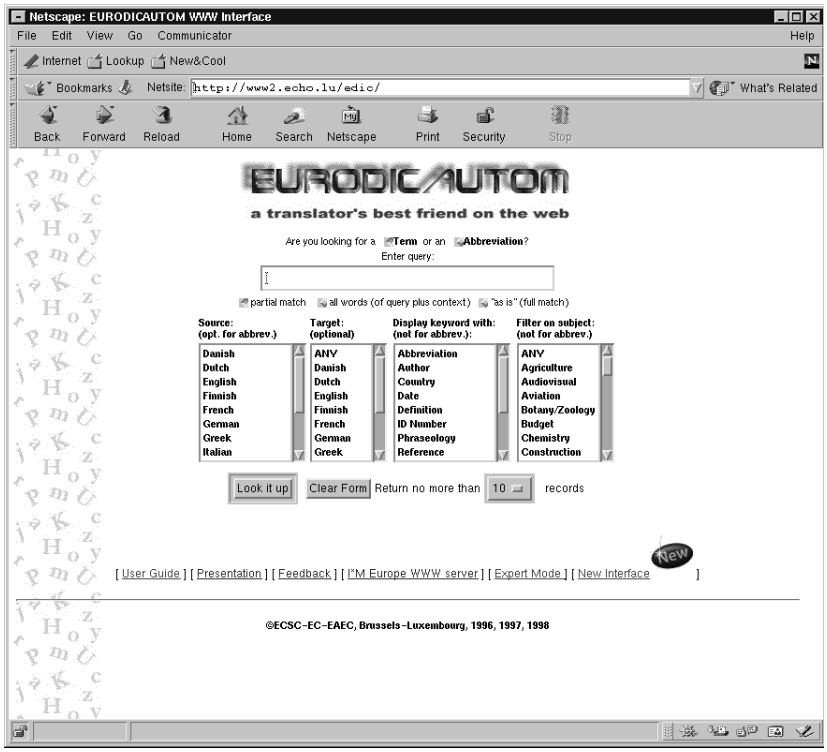

Figure 4.8: Interface of the EURODICAUTOM database

4.6.4 EURODICAUTOM

EURODICAUTOM is a multilingual terminological database of the Commission's Translation Service, which was initially developed to assist in-house translators. However, today it is consulted by EU officials and language professionals throughout the world. It is drafted in twelve European languages (including Latin) and covers a broad spectrum of human knowledge, such as administration, economy, geology, commerce, building, nuclear, arts, etc., but the main interest lies in European Union topics. The database contains about 5 million entries of technical terms, abbreviations, acronyms and phraseology. A typical entry displays two sorts of data (EURODICAUTOM 1998):

- general information:
  - Reliability code,
  - Date (when an entry was created or modified),
  - ID Number of the entry within a collection,
  - Type (a particular terminological collection),
  - Subject code (indicating the specialised field),
  - Terminological bureau (office responsible for storage of the information),

  and

- terminological information:
  - Keyword (a term in the broadest sense),
  - Phraseology (a phrase or sentence showing the context),
  - Definition of the concept,
  - Reference (the source of the information),
  - Technical note (or explanatory note).

4.6.5 MARTIF terminology interchange format (ISO 12200)

Developing a terminology database is an expensive and time consuming activity. In order to avoid double work and to prepare tools for data interchange and the re-use of ressources, a standard form for terminology interchange has been established. Of course one could assume the possibility of terminology interchange without any structural markup (e.g. a text file), but this results in the necessity of reformatting the unstructured data manually — a time and cost intensive undertaking. Therefore this possibility can be neglected.
The possibility of organising terminology in different database formats makes it seem unlikely to assume that for terminology interchange one could agree on a certain relational database format, such as some SQL formats. Nevertheless in order to enable terminology interchange a powerful tool was developed in cooperation with the *Text Encoding Initiative* (TEI). The goal was to produce a format that is platform independent and publicly available. The resulting format, the *Machine-Readable Terminology Interchange Format* (MARTIF), also known as ISO (FDIS) 12200, is based on *Standard Generalized Markup Language* (SGML, ISO 8879).[10] 150 data categories are defined for MARTIF in ISO (FDIS) 12620. The huge number of data categories results from the different needs and approaches of different working groups. Unfortunately MARTIF does not meet the needs of non-concept oriented approaches to terminology, e.g. lexicographic and NLP approaches, because MARTIF is *concept oriented* rather than *headword oriented*.
SGML code, and consequently MARTIF, is not easy to read for humans, however it is not intended to be read by humans. As there are HTML browsers (such as Netscape Communicator) for the presentation of HTML documents on the WWW, MARTIF can be presented by the means of userfriendly interfaces. A MARTIF parser could be included in a 'black box' guiding the terminologist inserting data — or even the user searching for specific information — to insert only data conforming to the MARTIF specification as well as presenting only relevant information to the user. This of course would require a stricter definition of data categories as is intended at the moment. The concept of MARTIF allows to include new data categories as well by including data categories as attributes to *Generic Identifiers* (GIs), i.e. a list of definitions of SGML tags. The advantage here is to keep the system as open as possible. On the one hand, placing data categories as GIs would mean to write them to the *Document Type Definiton* (DTD), which is a description of the formal content of an SGML document. The DTD of a specific format is written once to specify the format. To make interchange possible it is necessary that all users refer to the same DTD.

[10] The well known *Hypertext Markup Language* (HTML) used for WorldWideWeb documents is an example for SGML.

Making it possible to include data categories as attributes to GIs on the other hand opens the way to wider flexibility.
Another advantage of the MARTIF format is the possibility of targeting external links from within the document. References as well as all possible hypertext links can be included easily.
The first implementation of MARTIF still requires that programmers take a look at the sources (i.e. files to convert to MARTIF as well as to convert from MARTIF) before implementing tools for import into new systems. To develop a so called *blind* access, with the possibility of MARTIF interchange between any two systems, further standardisation is needed, especially for data categories, specific subject fields, etc.
Below, an example of a MARTIF file taken from the MARTIF test suite (available on the WWW pages of the Translation Research Group at Brigham Young University) is given. It has been modified for inclusion, only the `<body>` is given with only two languages represented.
For further information on MARTIF consult the WWW pages of the Translation Research Group at Brigham Young University (currently at "http://www.TTT.org").

Appendix: MARTIF test suite file

```
<body>

<termEntry>
  <descripGrp>
    <descrip type='subjectField'> appearance of materials </descrip>
     <note> treated in DIN under paper and cardboard </note>
   </descripGrp>

<note> The in-house working group for Optics is slated to finalize
this entry by 1995-12-15. </note>

   <ntig lang=en>
    <termGrp>
     <term> opacity </term>
     <termNote type='pos'> n </termNote>
    </termGrp>

    <descripGrp>
     <descrip type='definition'> degree of obstruction to the
      transmission of visible light </descrip>
     <ptr type='sourceIdentifier' target='ASTM.E284'>
    </descripGrp>

    <descripGrp>
     <descrip type='figure'> Degrees of Opacity </descrip>
     <note> The chart provides graphic images illustrating various
     degrees of opacity. </note>
     <ptr type='figure' target='f357'>
    </descripGrp>
```

```
    <adminGrp>
     <admin type='responsibility'> ASTM E12 </admin>
    </adminGrp>
   </ntig>

   <ntig lang=de>

    <termGrp>
     <term> Opazit&auml;t </term>
     <termNote type='pos'> n </termNote>
     <termNote type='gender'> f </termNote>
    </termGrp>

    <descripGrp>
     <descrip type='definition'> Ma&szlig; f&uuml;r die
      Lichtundurchl&auml;ssigkeit </descrip>
     <ref type='sourceIdentifier' target='DIN-6730-1992-08'> p. 5 </ref>
    </descripGrp>
    <adminGrp>
    <admin type='responsibility'> Normenaussch&szlig; Papier und
     Pappe (NPa) im DIN Deutsches Institut f&uuml;r Normung e.V. </admin>
    </adminGrp>
   </ntig>

</termEntry>

</body>
```

4.7 The EAGLET Term Database: an SL termbank

EAGLET is an enterprise providing standard terminology in the field of spoken language systems. It has been developed within the framework of the EAGLES Phase II project, LE 3-4244 (Telematics Applications Programme — Sector D/12: Language Engineering).

4.7.1 A hypergraph–based approach

In order to take the Scylla of heterogeneity in the field of SL terminology into account, and avoid, on the other hand, the Charybdis of a completely *ad hoc* hybrid description, a new approach is proposed which

- combines the traditional semasiological and onomasiological approaches to terminology characterisation,
- re–uses existing computer–readable terminological documentation and relevant text,
- develops a notion of a terminological hypergraph model and applies this in the construction of a terminological hyperlexicon.

With this goal in mind, the traditional device of *conceptual graphs* in the onomasiological characterisation of terminology is replaced by an explicitly defined

macrostructure with substructures which are designed to be realised as a *terminological hypergraph*, which in turn serves as a specification for the design of a *hyperlexicon* for implementation on CD–ROM and World Wide Web contexts (EAGLET HyperLexicon). The 'leaves' of the hypergraph are the terms; terms and their vector of defining properties are used to specify the data categories and records of the terminological relational database (EAGLET DB); EAGLET DB is operational with provisional functionality, and EAGLET HyperLexicon will remain for the medium term future as a specification.

4.7.2 Conceptual parts

1. *Architectural model*: For EAGLET, a single relation, the microstructure, is defined. The architectural model is the specification for the implementation in database software.
2. *Database engine*: In EAGLET development, the engine used is that of the software package miniSQL. To provide a maximum of platform independence and to prevent controversies resulting from usage of different versions the database is stored on one machine and accessed via the WordWideWeb.
3. *Front end tools*: In the EAGLET implementation, JavaScript menu control is used. This script language is implemented in most modern WWW browsers and is platform independent.
4. *Normalisation rules*: An HTML form input mask is used. Only the categories and data fields that are implemented in the form are possible to enter and to display.

4.7.3 Information storage

The following three main types of field are currently envisaged for the EAGLET relation:

1. *static*: the entity (here: term attribute value) is stored in an *ASCII* coded format (e.g. simple text, *SAMPA* Notation, LaTeX code)
2. *hyperlink*: term relations that depend on the query's context are built up as *URLs*.
3. *media event*: data structures are coded (if neccessary in an appropriate data format '*on-the-fly*'). They are stored outside the DB and are referenced by *URLs*.

4.7.4 System components

1. Database server: Currently an SQL Database Server is in use (*Hughes Technologies' mSQL* — Version 2.0.3)
2. Filters to import external data: UNIX Scripts — written under Solaris 2.5.1 and Linux are available to manipulate external data into a suitable format for mSQL building import functions via the script language *lite* — Version 2.0.3, which is in principle a modification of C.
3. Database query language: *mSQL's* CGI interpreter and script language *lite* — Version 2.0.3
4. Interface application: any WWW browser — e.g. *Netscape Communicator 4.0* with the ability of interpreting Hypertext Markup Language (HTML) version 3.2 or higher and *JavaScript* version 1.1 or higher.
5. Interface programming language: *Hypertext Markup Language (HTML)* – SGML Public Indentifier "-//W3C//DTD HTML 3.2//EN"

6. Interface validation language: *JavaScript* — Version 1.1

4.7.5 Structure

The overall structure of EAGLET is shown in Figure 4.9.

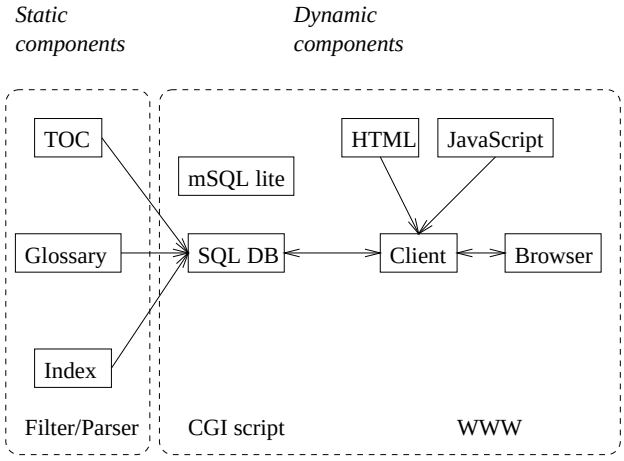

Figure 4.9: Structural overview of EAGLET

As it is seen, the database with its structured entries (taken from the table of contents, glossary and index of the *Handbook of Standards and Resources of Spoken Language Systems*) is of a fairly static nature with the possibility of evaluation and manipulation via filters and parsers. The SQL database machine serves as the interface via the scripting language *mSQL lite* to the dynamic, interactive components of the client starting a query using an HTML form (evaluated by a JavaScript applet) displayed by a browser. This is of dynamic nature: every interface page is generated with the up to date data of the database, and queries can be reduced to the user's needs.

4.7.6 EAGLET macrostructure for SL terminology

In view of the complexities involved — and the very large number of degrees of freedom — a pragmatic approach has been taken in the development of the EAGLET concept. The approach involved developing a macromodel for spoken language terminology based on the macrostructure of the Parts and Chapters of the *Handbook of Standards and Resources for Spoken Language Systems*. For this purpose, the following textual components of the Handbook will be incorporated into a hypergraph design for a terminological hyperlexicon:

1. Table of contents (TOC). The TOC represents a possible onomasiological structure for the content, and provides an elementary variety of onomasiological indices into the text.
2. Body of text. The body of the handbook provides expert–developed contexts in which terminology is authentically attested; the body of text in the chapters therefore defines an authentic corpus of attested forms in context.

3. Glossary. The Glossary is effectively a semasiological dictionary with headword and definitions, usually of the *genus proximum et differentia specifica* type.
4. Index. The Index indirectly provides a semasiological concordance, with headword and pointers into corpus of attested forms in context.

4.7.6.1 Sub–taxonomies

The text source for the terminology hyperlexicon provides taxonomies with a greater degree of granularity than that outlined so far, based on a subtree of the table of contents of the *Handbook of Standards and Resources for Spoken Language Systems*. For convenience in representation, the taxonomy is divided into the sub–taxonomies:

- System Design,
- Corpus Design (see Figure 4.10,
- Lexicon Development,
- Language Models,
- Physical Characterisation,
- Assessment methodology,
- Recogniser assessment,
- Speaker Verification assessment,
- Synthesis assessment (see Figure 4.11), and
- Interactive Dialogue System Assessment.

The structure is modified from the basic text organisation of the *Handbook of Standards and Resources for Spoken Language Systems*, and is intended to represent, in each case, a first starting point for a pragmatic applications orientated basic system design taxonomy.
Taken together, the sub–taxonomies constitute a comprehensive taxonomic hierarchy of fine granularity; the sub–taxonomies have been curtailed at a coarse–grained level, but as the textual structure of the *Handbook of Standards and Resources for Spoken Language Systems* shows, much finer grain is available. In later versions of the work on spoken language terminology, this will be used for graphically oriented access to term definitions.

4.7.6.2 Graphical representations of sub–taxonomies

Figures 4.10 and 4.11 show two of the sub–taxonomies mentioned above.
The sub–taxonomies for different areas show very different kinds of structure, as to be expected. The differences encompass the following topological and semantic features of the graphs which are, with few exceptions, tree graphs:

- Number of nodes
- Depth of branching
- Breadth of branching
- Differences in node interpretation, e.g. in terms of *formalism* (notation, terminology, nomenclature), *empirical method*, or *sub–domain* (field, subject)
- Differences in edge interpretation, e.g. as *ISA* or strict taxonomic interpretation, vs. *PARTOF* or mereonomic interpretation.

However, the explicit graphical representation of the taxonomies provides a useful basis for future work, in which similarities between the different sub–taxonomies can be examined in more detail and, in some cases, merged.

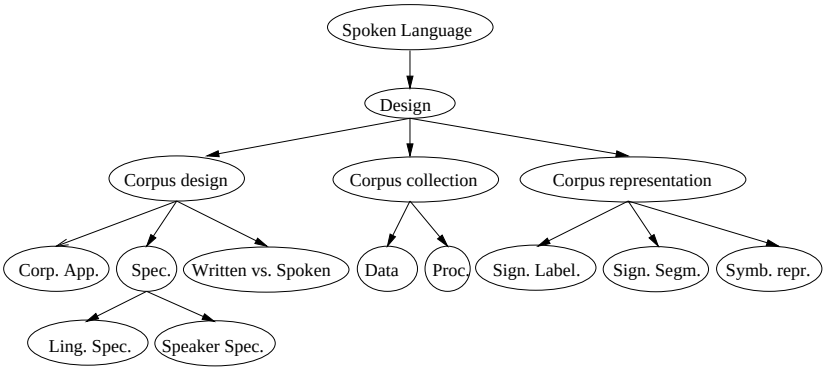

Figure 4.10: A basic corpus design taxonomy

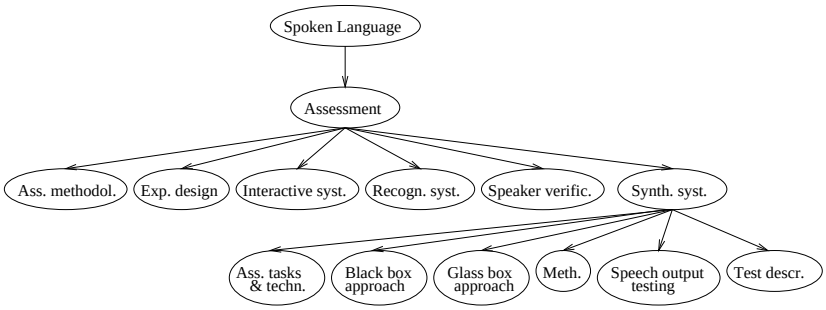

Figure 4.11: A basic speech synthesis taxonomy

4.7.7 EAGLET microstructure for SL terminology

The currently implemented EAGLET microstructure contains the following data categories as fields in the database:

1. Form: Orthography
   A representation of the term in standard British English orthography.
2. Form: Pronunciation
   The phonemic transcription of the term in SAMPA notation a revised version of which is presented in Gibbon et al. (1997).
3. Form: Part of Speech
   The structure of compounds is given in attribute–value notation.
   Example: The term 'text-to-speech system' is analysed as '[N: [N: text][PREP: to][N: speech][N: system]]'.
4. Form: Inflections
   As nearly all terms in EAGLET are nouns, this category basically indicates the plural form(s) of terms. The possible values are: -s ('badger' – 'badgers'), -es ('search' – 'searches'), -0 ('sheep' – 'sheep'), none ('Bayesian decision theory'); for nonregular forms and the '-ies' plural in words like 'frequencies' the plural form is given in full.
5. Semantics: Domain

'Domain' refers to the individual chapter of the *Handbook of Standards and Resources of Spoken Language Systems* the term can be assigned to. The default value 'Spoken Language Technology' has been entered for all terms, and, where possible, the more specific subject field such as 'physical characterisation, 'corpora', 'lexicon', 'interactive dialogue systems' is added.
Example: For 'Hidden Markov Model' the value is 'Spoken Language Technology: language modelling'. Many terms, however, are difficult to place because they are very general, such as 'orthographic transcription', a term that occurs in nearly all handbook chapters and, like many others, is not restricted to the domain of spoken language technology.

6. Semantics: Hyperonyms
   The data category 'hyperonyms' corresponds to the classical genera proxima in terminological theory. A *hyperonym* is the verbal representation of the superordinate concept of a term in a taxonomy.[11]
   Examples: *morph* is a hyperonym of *bound morph* because '*A bound morph is a type of morph*' is acceptable.
   Accordingly, *unidirectional microphone* is a hyperonym of *cardioid microphone.*
7. Semantics: Hyponyms
   A *hyponym* is the verbal representation of the subordinate concept of the term in question.
   Examples: A *bound morph* is a hyponym of *morph* because *A bound morph is a kind of morph* is an acceptable sentence.
   The hyponyms of *microphone* are *unidirectional microphone*, *bidirectional microphone*, *omnidirectional microphone*, *ultradirectional microphone*, *pressure zone microphone*, *headset microphone*; *headmounted microphone*, *table-top microphone*, *handheld microphone*, *room microphone*; *dynamic microphone*, *condenser microphone.*
8. Semantics: Synonyms
   A synonym is a term that represents the same concept as the main entry term in a term entry. In EAGLET, no distinction is made between genuine synonyms and quasi–synonyms. Quasi synonyms are terms that represent the same concept in the same language, but for which interchangeability is limited to some contexts and inapplicable in others.
   Example: *wolf* is a synonym of *skilled impostor.*
9. Semantics: Antonyms
   This data category covers terms denoting all types of lexical opposite. Complementaries, i.e. terms that "divide some conceptual domain into two mutually exclusive compartments" (Cruse 1986, p. 198) are treated as a subset of antonyms.
   Example: *cardioid microphone* and *hypercardioid microphone* are antonyms of *supercardioid microphone.*
10. Semantics: Definitions
    As in most standard general dictionaries, EAGLET not only contains analytical definitions, i.e. definitions which give a noun phrase providing the meaning of the term in question (Sager and L'Homme 1994), but also definitions that contain nonessential characteristics and information that would be classified as 'world knowledge'. In many cases also the source of the definition is given. Example: The unidirectional type of microphone is most sensitive to

[11] Cruse (1986) distinguishes the semantic relation of taxonomy (type-of relation) from that of hyponymy (kind-of relation), which is a less restrictive relation. In Cruse's approach the set of taxonyms of a term is a subset of the set of hyponyms of the term. For EAGLET, however, no such distinction has been made.

sound arriving from one direction and more or less attenuates incident sound from other directions. Thus, unidirectional microphones will suppress intended sound when pointed at the wanted sound source, i.e. the speaker. (Gibbon et al., p. 303)

11. Semantics: Meronymic superordinates
    Terms that are superordinates in a PARTOF hierarchy. Example: *syllable* is a meronymic superordinate of *onset* because *The/An onset is part of a syllable* is an acceptable sentence.
12. Semantics: Meronymic subordinates
    Terms that are subordinates in a PARTOF hierarchy. Example: *onset* is a meronym of *syllable*, because *An onset is a part of a syllable.* is an acceptable sentence.

    In EAGLET no distinction is made between facultative and non-facultative parts, and no information is given as to whether constituents occur in a certain order or not: for example, the order onset–nucleus–coda (= parts of a syllable) is not expressed in EAGLET. Note that two or more meronymic hierarchies may co-exist depending on the classificatory criterion.
13. Context: Examples
    A term and its definition is exemplified.
    Example: 'un' and 'able' in 'unbearable' are affixes.
14. Context: Graphic models
    This data category is reserved for visual representations of a concept.
15. Context: Audio models
    This data category is reserved for auditory representations of a concept.
16. Context: Formulas
    Here formulas are given that might replace a textual definition.
17. Context: Concordance links
    Here the occurrences of a term in the WWW edition of Gibbon et al. (1997) is given. At the moment this information is not accessible.
18. Administration: Date
    This administrational category shows the date of the last change of the record.
19. Administration: Author
    The administrators who performed the changes to the record are given.

4.7.8 Using the EAGLET Term Database

4.7.8.1 Access to the EAGLET Term Database and system requirements

The following URL is provided for access to the database:
"`http://coral.lili.uni-bielefeld.de/EAGLES/SLWG/TERMBANK/interface.shtml`"
In order to enhance the client–server interactivity, EAGLET makes extensive use of JavaScript, which must be supported by the web browser in order to access the EAGLET query interface.
Instead of prompting for abstract Standard Query Language (SQL) expressions EAGLET uses advanced interaction elements (buttons, select boxes, etc.) available on mostly all current operating systems (e.g. Windows 3.xx/95, X-Windows, MacOS). These GUIs (Graphical User Interfaces) allow the user to perform queries by simply clicking on the relevant items. Therefore it is recommended to use GUIs for a comfortable interaction with the EAGLET database.

4.7.8.2 Selecting terms and attributes

Figure 4.12 is a screenshot of the EAGLET user interface.

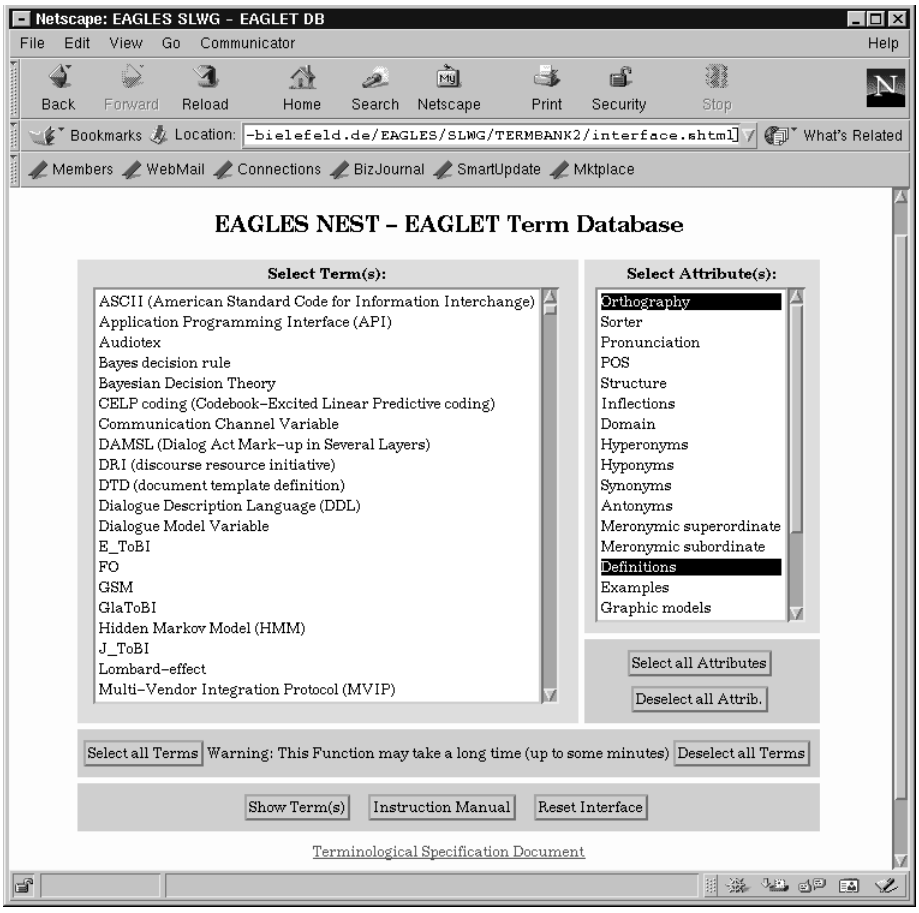

Figure 4.12: EAGLET Term Database interface

The interface basically consists of two picklists. One is a list of terms included in the EAGLET Term Database, the other shows the data categories (attributes) used for analysing and describing the terms. The user can select (by mouse click) one or more terms and attributes that are of interest for him. Five attributes are preselected: 'Orthography', 'Hyperonyms', 'Meronymic superordinate', 'Definitions', 'Examples'. Alternatively, he can press the button "Select All Attributes" for a thorough description of a term.
The marked terms and attributes are highlighted (by a black bar or other).

If the cursor is in the term or attribute field, the user may type in the term/attribute on the keyboard instead of scrolling down the list. This helps saving time, considering that there are currently about 700 terms available.

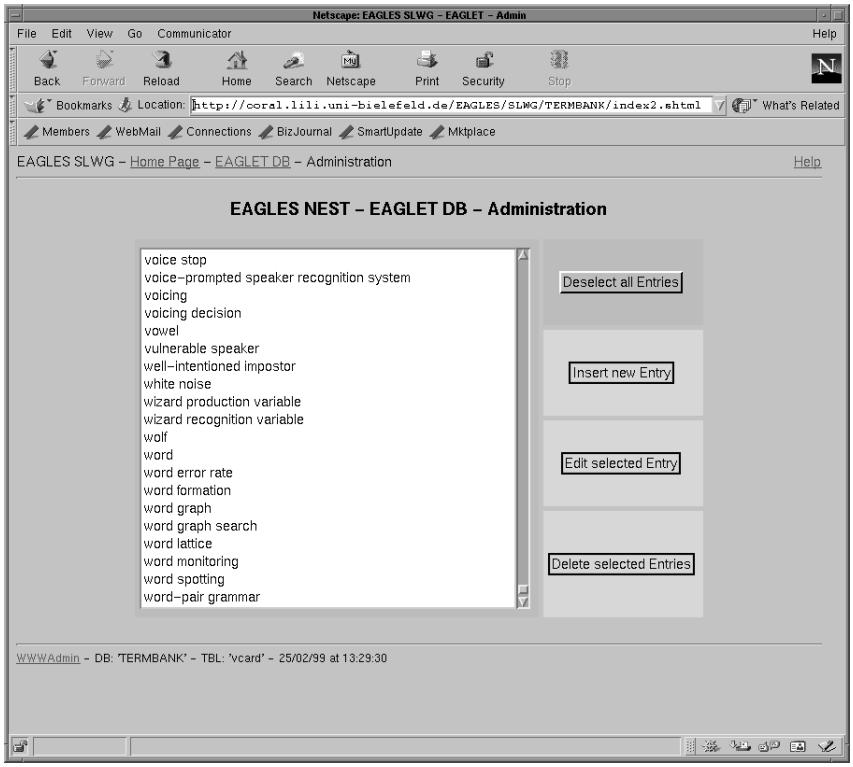

Figure 4.13: EAGLET administration interface

4.7.8.3 Deselecting attributes and terms

A term or attribute is deselected by clicking on the marked item again. All marked attributes can be demarked by making use of the 'deselect all attributes' button.
If the user wishes to demark all marked items at once, he may use the 'reset' button.

4.7.8.4 Submit the query

The 'show terms' button submits your query to the database server. The user is shown the list of terms described by the attributes he selected.

4.7.8.5 Further advice / help

Online help is provided ('instruction manual' button).

4.7.8.6 The EAGLET administration interface

The EAGLET administration interface is shown in Figure 4.13. EAGLET provides the functions 'Insert New Entry', 'Edit Selected Entry', and 'Delete Selected Entries'. An entry which has been marked for deleting or editing can be demarked by pressing the 'Deselect All Entries' button.

Inserting a new term is done by pressing the 'Insert New Entry' button, which opens an interface where the orthographic representation of the new term can be typed in. After giving the insert command the input interface with all data categories described above is presented and the data can be typed in.
For editing, only one term must be marked. Marking is done as explained in Section 4.7.8.2.

4.7.9 Future work

1. Currently EAGLET contains about 1250 term entries. It is envisaged to enhance the database to up to 1500 terms in the near future.
2. So far access is basically semasiological, i.e. access is very much headword oriented. It is intended to represent the taxonomies outlined above in graph form so as to permit onomasiological access as well.This is especially interesting for users who need a brief overview of the concepts of spoken language technology and who wish to learn about the relevant classifications before they turn to the details of individual terms.
3. The subfields of spoken language technology as given by the Handbook and Supplement chapters are not equally well represented with respect to the number of terms included, and the depth of representation of the conceptual fields is not uniform. Some work will have to be invested in making the representation more uniform and detailed.
4. Since the sources of terms are the Handbook and the Supplement, concepts/related concepts may be missing because they are simply not mentioned. These gaps must be detected and filled with the respective terms.
5. It is not yet intended to expand EAGLET into a multilingual termbase.

5 Reference materials

5.1 Introduction

What is so special about spoken language? Speech differs from written language in many ways (Tillmann 1997):

- Speech is a signal over time, whereas written language is a symbolic representation, made up of categories.
- Speech is volatile – once it has been produced, it is gone.
- Speech is produced and processed in real time.
- In speech, errors are compensated automatically by speaker and listener.
- Spoken words may take on almost any form ranging from carefully articulated to very reduced.
- Technically, speech is stored as sampled speech signals, whereas written language is represented as strings over a given alphabet.
- Speech and multi-media data requires large storage capacities, whereas written language can be stored in a compact format.

Clearly, speech deserves special treatment. Speech recognition, speech synthesis, and speech encoding are the major areas of SLP (Spoken Language Processing). Any development in SLP requires SLP resources, and at the same time provides new SLP resources, for example, the development of a speech recogniser requires substantial speech data for training; once the recogniser is available, it can be used for future automatic orthographic transcriptions of speech.
These resources need to be shared – for scientific exchange as well as commercial purposes. This exchange is only possible if a minimum standard of quality is assured. The field of SLP is now mature enough to propose such minimum standards for the technical formats, content, and procedures related to SLP resources.
This compilation of reference materials contains references to information considered to be important for the creation and exploitation of SLP resources. It is a source of information complementary to the discussions in the Usenet newsgroups and the WWW information archives.
The material is divided into four major sections:

- Organisations and Infrastructure
- "SLP at work"
- SLP procedures, tools, and formats
- Technology

Each section contains subsections with a short explanatory introduction followed by a list of reference items. Each item consists of a brief description (often copied from the original source of information), keywords, and an Internet address or a publication reference (if applicable).
To make this compilation a useful source of information, references generally point to institutions or *stable* WWW sites: agencies, organisations, university departments, academic or industrial research labs, WWW search engines, and important newsgroups and their frequently asked questions (FAQ) archives. These are places to start searching for more information.

5.2 Organisations and infrastructure

5.2.1 Speech resources, agencies, and associations

Speech resources, i.e. corpora, tools, applications, lexica, etc., are developed in academic and commercial SLP laboratories all over the world. To make these resources available, and to concertate the development of new resources, *agencies* have been established.
National agencies often arose from existing language or speech related institutions. They cover the political or geographical extension of a country or the distribution region of a single language. These agencies often also cover further language related topics, e.g. diachronic linguistics, normative orthography (and its reforms), etc. Because of their importance for the cultural identity of a nation they are often publicly funded.
Special interest group agencies are established by professional associations, e.g. the acoustical societies that exist in many countries. The best known agencies dedicated to SLP resources are the LDC (Linguistic Data Consortium) in the US, ELRA (European Linguistic Resources Association) in Europe. Providers and consumers of SLP resources can become members of these agencies to gain access to SLP resources, and to influence the development of new resources, e.g. through competitions and projects. In general, these agencies provide resources for more than one language.
In projects, a *consortium* defines a project task and applies for funding. The consortium often consists of academic and industrial partners. Because of the high cost of the creation of SLP resources, even competitors can cooperate in a consortium – the resources are created together, but exploited individually. In publicly funded projects, the resources created become publicly available – not necessarily free of charge – after a given time span.
Funding organisations range from national government institutions to industrial companies and university departments. On a national level there exist government institutions in most countries that provide funds for projects. Europe is a special case because besides the national governments there is the EU (European Union). The EU commission issues calls for project proposals within key areas of technology. To obtain funding, a collaboration of academia and industry is often mandatory.

5.2.1.1 EAGLES

EAGLES (Expert Advisory Group on Language Engineering Standards) is an EU funded action to report on and promote the use of standards in the areas of spoken language processing, text, and terminology.
In 1997 the Spoken Language Group of EAGLES published the *Handbook of Standards and Resources for Spoken Language Systems* (Gibbon et al. 1997).

Source, Availability

"`http://www.ilc.pi.cnr.it/EAGLES/home.html`"

5.2.1.2 Acoustical societies

This list of acoustical societies around the world was compiled by Metin Erdogan ("erdogan@fiesta.me.metu.edu.tr") of the Middle East Technical University of Ankara, Turkey (for printing, the list was revised by the author).

Argentina	Asociacion de Acusticos Argentinos Laboratorio de Acustica AR 1897 – Gonnet
Australia	Australian Acoustical Society Darlinghurst, NSW 2010
Austria	Austrian Acoustics Association Technische Universität Wien Institut für Allgemeine Physik A–1140 Wien
Belgium	Belgian Acoutics Assosciation (ABAV) Av. P. Holoffe 21 B-1342 Limelette
Brazil	Sociedade Brasileira de Acustica Universidade Federal de Santa Catarina Departamento de Engenharia Mecanica Florianopolis – SC
Canada	Canadian Acoustical Association PO Box 1351, Station F Toronto, M4Y 2V9 Canada
Chile	Sociedad Chilena de Acustica San Francisco # 1138 Santiago de Chile
China	Acoustical Society of China 17 Zhongguancun St. Beijing 100080 China
Czech Republic	Czech Acoustical Society Technicka 2 CZ–166 27 Prague 6
Denmark	Acoustical Society of Denmark Technical University of Denmark DK–2800 Lyngby
Finland	Acoustical Society of Finland c/o Helsinki University of Technology FIN–02150 Espoo
France	Societé Française d'Acoustique 23 avenue Brunetiere F–75017 Paris
Germany	Deutsche Gesellschaft für Akustik University of Oldenburg D-26111 Oldenburg

Greece	Hellenic Acoustical Society Patision 147 GR–112 51 Athens
Hungary	Scientific Society for Optics, Acoustics (OPAKFI) Fo utca 68 H–1027 Budapest
India	Acoustical Society of India CEERI Centre, CSIR Complex New Delhi – 110012
Italy	Associazione Italiana di Acustica via Cassia 1216 I–00189 Roma
Japan	Acoustical Society of Japan Nippon Onkyo Gakkai 4th Floor 2–7–7 Yoyogi, Shibuya-ku Tokyo
Korean Republic	Acoustical Society of Korea Korean Federation of Science and Technology 635-4, Yeoksam-dong Kangnam-gu Seoul 135-080
Mexico	Instituto Mexicano de Acustica P.O. BOX 75805 Col. Lindavista 07300 Mexico, D.F.
Netherlands	Nederlands Akoestisch Genootschap Postbus 162 NL–2600 AD Delft
New Zealand	New Zealand Acoustical Society CPO Box 1181 Auckland, New Zealand
Norway	Norsk Akustisk Selskap c/o Lydteknisk senter-NTH Sintef Delab N–7034 Trondheim
Peru	Sociedad Peruana de Acustica Garcilazo de la Vega 163 Salamanca de Monterrico Lima 3
Poland	Polskie Towarzystow Akustyki Instytut Akustyki Uniwersytet Adama Mikiewicz ul J.Matejki 48/49 PL–60-769 Poznan
Portugal	Portuguese Acoustical Society SPA - CAPS/Instituto Superior Tecnico Av. Rovisco Pais P–1096 Lisboa CODEX

Romania	Societatea Romana de Acustica Universitatea Politehnica Bucuresti Independentei nr. 313 ROM 77206 Bucuresti
Russia	Russian Acoustical Society 4 Shvernik ul Moscow 117036 Russia
Singapore	Singapore Acoustics Society Acoustical Services Pte Ltd 209–212 Nanyang Ave Singapore 2263
Slovakia	Slovak Acoustical Society Racianska 75 PO Box 95 830 08 Bratislava 38 Slovakia
South Africa	South African Acoustics Institute P.O. Box 912-169 Silverton South Africa, 0127
Spain	Sociedad Espanola de Acustica Serrano 144 E–28006 Madrid
Sweden	Svenska Akustiska Sallskapet Ingemansson AB Box 47 321 S–100 Stockholm
Switzerland	Schweizerische Gesellschaft für Akustik Postfach 251 CH–8600 Dübendorf
Turkey	Turkish Acoustical Society - TAS Y.T.U. Mimarlik Fakultesi Yildiz 80750 Istanbul
UK	Institute of Acoustics 5 Holywell Hill, St Albans, Herts AL1 1EU
USA	Acoustical Society of America 500 Sunnyside Blvd. Woodbury, NY 11797

Source, Availability

"`http://www.me.metu.edu.tr/courses/ME432/addressc.html`"

5.2.1.3 DARPA

DARPA (Defense Advanced Research Projects Agency) is a US agency. Its mission is to "develop imaginative, innovative and often high risk research ideas offering a significant technological impact that will go well beyond the normal evolutionary developmental approaches; and, to pursue these ideas from the demonstration of technical feasibility through the development of prototype systems."
In the SLP arena, DARPA is best-known for its competitions, e.g. speech recognition of the Switchboard corpus. These competitions are held in collaboration with the NIST (National Institute of Standards and Technology).

Source, Availability

"`http://www.arpa.gov/`"

5.2.1.4 NIST

The American National Institute of Standards and Technology has an active SLP group. This group contributes to the advancement of the state-of-the art of SLP so that spoken language can reliably serve as an alternative modality for the human–computer interface.
NIST develops measurement methods, provides reference material, e.g. speech corpora, organises benchmark tests within the SLP community, and builds prototype systems.
NIST has proposed a widely used standard header for audio signals, the NIST-SPHERE header. It consists of a simple ASCII formatted text information describing the signal following the header; the header options can be user-specified. Software for creating and manipulating NIST-SPHERE files is available at the NIST web site.

Source, Availability

"`http://www.nist.gov/speech`" for informations on the SLP group,
"`ftp://jaguar.ncsl.nist.gov/pub/`" for the Speech File Manipulation Software (SPHERE) Package Version

5.2.1.5 ACL

The ACL (Association of Computational Linguistics) is the scientific and professional society for people working on problems involving natural language and computation. The ACL journal, *Computational Linguistics*, continues to be the primary forum for research in computational linguistics and natural language processing.

Source, Availability

"`http://www.cs.columbia.edu/ acl/home.html`"

5.2.1.6 BAS

The BAS (Bavarian Archive for Speech Signals) was founded as a public institution in January 1995 and is hosted by the University of Munich, presently at the Department of Phonetics (Institut für Phonetik und Sprachliche Kommunikation – IPSK).
The BAS is dedicated to provide databases of spoken German in a well-structured form to the speech science community as well as to speech engineering.
The BAS features an extensive online catalogue with access to speech and annotation samples.

Bavarian Archive for Speech Signals
c/o Institut für Phonetik und Sprachliche Kommunikation
Universität München
Schellingstr. 3 / II
D-80799 München
Tel.: +49-89-2180-2758
Fax: +49-89-2800362
Email: "bas@phonetik.uni-muenchen.de"

Source, Availability

"http://www.phonetik.uni-muenchen.de/Bas"

5.2.1.7 EACL

The EACL is the European chapter of the ACL. It is hosted by the University of Geneva.

Source, Availability

"http://issco-www.unige.ch/eacl/eacl.html"

5.2.1.8 IEEE

The IEEE (Institute of Electrical and Electronics Engineers) is a professional organisation. The Signal Processing chapter has a section devoted to speech. The IEEE publishes several journals and organises many workshops and conferences, e.g. ICASSP.
Many national engineering professional organisations have established close collaborations with the IEEE.

Source, Availability

"http://www.ieee.org"

5.2.1.9 COCOSDA

The International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques for Speech Input/Output, COCOSDA, has been established to encourage and promote international interaction and cooperation in the foundation areas of SLP. The importance of collaboration which transcends national boundaries is increasingly recognised. This is both because of the practical and scientific value attached to systematic work which encompasses a range of languages and analytic approaches and also because of the practical need to establish common methods of performance description and quantitative comparison.

Source, Availability

"`http://www.itl.atr.co.jp/cocosda/`"

5.2.1.10 ELRA: European Linguistic Resources Association

The ELRA (European Language Resources Association) was established in Luxembourg in February 1995, with the goal of founding an organisation to promote the creation, verification, and distribution of language resources in Europe. Being a non-profit organisation, ELRA aims to serve as a focal point for information related to language resources in Europe. It will collect, market, distribute, and license European language resources.
ELRA will help users and developers of language resources, government agencies, and other interested parties exploit language resources for a wide variety of uses. Eventually, ELRA will serve as the European repository for EU-funded language resources and interact with similar bodies in other parts of the world. ELRA is currently located in Paris, France:

ELRA/ELDA
55-57, rue Brillat Savarin
F-75013 PARIS
Tel: +33 1 43 13 33 33
Fax: +33 1 43 13 33 30

Source, Availability

"`http://www.icp.grenet.fr/ELRA/home.html`"

5.2.1.11 ELSNET

ELSNET is the European Network in Language and Speech. The long-term technological goal is to build multilingual speech and natural language systems with unrestricted coverage of both spoken and written language. ELSNET, which has over a hundred European academic and industrial institutions as members, is one of over a dozen Networks of Excellence established by the European Commission's ESPRIT Division for Basic Research.

Source, Availability

"http://www.elsnet.org"

5.2.1.12 ESCA (European Speech Communication Association)

The main goal of the Association is "to promote Speech Communication Science and Technology in a European context, both in the industrial and Academic areas", covering all aspects of speech communication (acoustics, phonetics, phonology, linguistics, natural language processing (NLP), artificial intelligence (AI), cognitive science, signal processing, pattern recognition, etc.).
ESCA is the organiser of the Eurospeech conference series, and supports many SLP workshops and summer schools for academic and industrial audiences.

Source, Availability

"http://ophale.icp.inpg.fr/esca/esca.html"

5.2.1.13 FRANCIL

FRANCIL is a francophone network of scientific research. Its aim is to establish collaborations between research laboratories inside and outside France, especially in the southern hemisphere, and to encourage scientific production in French.
FRANCIL organises competitions on selected research areas, and holds workshops and conferences. FRANCIL currently is hosted by Limsi in France.

Source, Availability

"http://www.limsi.fr/Recherche/FRANCIL/frcl.html"

5.2.1.14 Institut für deutsche Sprache

The "Institut für deutsche Sprache" (IDS) in Mannheim was founded in 1964. It is the central non-universitary institution for the research and documentation of the contemporary German language.

Source, Availability

"http://www.ids-mannheim.de"

5.2.1.15 European Commission (Language Engineering and Human Language Technology)

The European Commission sets up so-called 'Frameworks' that identify the key technologies that receive funding by the EU. A Framework spans five years. In the Fourth Framework (1994–1998) SLP related developments were concentrated in the Language Engineering (LE) sector, in the Fifth Framework (1999–2003) it will be in the Human Language Technology (HLT) sector.
The aim of Language Engineering is to facilitate the use of telematics applications and to increase the possibilities for communication in and between European languages. Work focuses on pilot projects that integrate language technologies into information and communications applications and services. A key

objective is to improve their ease of use and functionality and broaden their scope across different languages.
The Euromap project within LE has published two surveys on Language and Speech Technology in Europe, and the European Commission Directorate General XIII/E has produced "A World of Understanding", a CD-ROM that contains all key projects funded by LE during the Fourth Framework.
The surveys contain an overview of language related activities by country, links to industrial applications, a list of EU-funded projects and their coordinators.
The reports and the CD can be obtained from

European Commission, DG XIII/E
Rue Alcide de Gasperi
L–2920 Luxembourg
"hltlux.dg13.cec.be"

The Directorate General XIII/E also maintains a mailing list.

Source, Availability

"http://www.echo.lu/le"

5.2.1.16 LDC: the Linguistic Data Consortium

The LDC (Linguistic Data Consortium) is an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases, lexica, and other resources for research and development purposes.
The LDC was founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), and is partly supported by the National Science Foundation. It is hosted by the University of Pennsylvania.
Most LDC services and corpora are available to members and non-members. Non-members in general pay substantially higher license fees, and some services or corpora may not be available to them at all. Evaluation licenses are available for some corpora.
LDC features a well organised WWW site with online access to the large catalogue and search in selected corpora.

Linguistic Data Consortium
University of Pennsylvania
3615 Market Street, Suite 200
Philadelphia, PA 19104-2608

Tel (215) 898-0464
Fax (215) 573-2175

Source, Availability

"http://www.ldc.upenn.edu/"

5.2.1.17 LIMSI

The LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur), a member of the French CNRS (Centre National de la Recherche Scientifique) network, is active in SLP within its Human–Machine communication department. The major research topics are spoken language processing, language and cognition, interaction and multi-modalities, and human cognition.

LIMSI-CNRS
Université de Paris-Sud
F 91403 ORSAY
France

Tel: +33 (0)1 69 85 80 80
Fax: +33 (0)1 69 85 80 88
WWW: "`http://www.limsi.fr`"

5.2.1.18 NSF

The NSF is a US American government agency. It initiates and supports scientific and engineering research through grants and contracts, to award graduate fellowships, and to foster the interchange of scientific information among the scientific communities.
The SLP related activities of the NSF are part of the Information and Intelligent Systems group.

The National Science Foundation
4201 Wilson Boulevard
Arlington, Virginia 22230
USA

Tel: +1-703-306-1234

Source, Availability

"`http://www.nsf.gov`"

5.2.2 Archives, general information

Independent of agencies and projects many voluntary actions and commercial providers offer valuable resources. One major source of information is the WWW (World Wide Web), where *archives* and *search engines* provide efficient access to up to date information. A second source are the *FAQ* (Frequently Asked Questions) lists which are compilations of information relevant to a Usenet newsgroup.
WWW archives store information in a database available at WWW sites distributed all over the world. Information enters these archives via search robots or explicit insertion. Search robots are applications that access or *visit* WWW sites and retrieve the documents accessible at these sites; these documents are

then indexed and added to the database. Most WWW archives provide a mechanism to enter information explicitly by requesting a *URL* (Uniform Resource Locator), i.e. an Internet address, which is then used as a starting point for a search robot.
There are two major types of access to the information stored in WWW archives:

- full text search, or
- search by category.

In both cases, a user enters a query string – usually a simple boolean expression – and the search engine then returns all matching documents. In full text search, the entire text base is indexed automatically, and this index is scanned to find the requested search string. In search by category only text descriptors and keywords are searched. This requires that every document entered into the database be categorised – and this requires knowledge about the type and content of the document. Full text search finds only documents that contain the search string explicitly, but it does not find related texts if they do not contain the search string. Search by category finds documents with matching keywords and descriptors, even if the documents do not contain the exact search string.
The result is sorted by relevance by the search engine. This order of relevance is determined by proprietary measures, e.g. number of occurrences of search string in the document, number of links referring to the document, number of visits to the document, etc.
The large number of different search engines has led to the development of *meta search engines* that search different search engines in parallel and attempt to sort the result by a predefined or user-specified ranking scheme. Meta search engines can be found either in the WWW, or be part of the operating system on a local machine, e.g. Sherlock in the MacOS. Local meta search engines may also index the local file system.
In the Internet there exist discussion groups (for historical reasons they are called Usenet groups or *newsgroups*) for almost any topic. A newsgroup is focused on one common subject, e.g. `comp.speech.users` and `comp.speech.research` for SLP. A contribution to a newsgroup is distributed to all news servers in the Internet and may be read by the subscribers to this newsgroup (many WWW archives subscribe to all newsgroups and add relevant postings to their archive). The hierarchy of newsgroups is self-organising. New discussion groups or subgroups can easily be created if there is sufficient support by subscribers.
FAQ lists are digests of the ongoing discussion in a newsgroup; they contain information that is of fundamental importance to the newsgroup and that is requested over and over again.

5.2.2.1 Audio File Formats FAQ

Audio File Formats FAQ contains a general overview of most audio file formats and features links to "official" audio file format descriptions, e.g. AIFF, RIFF, or NIST. Software for signal editing and file format conversion is also presented. The Audio File Formats FAQ is maintained by Chris Bagwell at "`cbagwell@sprynet.com`"; it was established by Guido van Rossum in 1991.

Table 5.1: SLP related newsgroups

Area	newsgroup
Applications and Products	`comp.speech.users`
Speech Synthesis and Recognition	`comp.speech.research`
Natural Language Processing Group	`comp.ai.nat-lang`
Neural Networks	`comp.ai.neural-nets`
Telecommunications (unmoderated)	`alt.dcom.telecom`
Telecommunications (moderated)	`comp.dcom.telecom`
Telecommunications Technology	`comp.dcom.telecom.tech`
Digital Signal Processing	`comp.dsp`
Language	`sci.lang`

Source, Availability

"`http://home.sprynet.com/sprynet/cbagwell/audio.html`"

5.2.2.2 SLP related newsgroups

Table 5.1 lists the newsgroups most relevant to SLP.

5.2.2.3 comp.speech.FAQ

The `comp.speech.FAQ` (maintained by Andrew Hunt of SUN) is the most comprehensive speech related archive on the WWW.
The FAQ contains many links to speech technology related pages, and is divided into six sections: General Information on Speech Technology, Signal Processing for Speech, Speech Coding and Compression, Natural Language Processing, Speech Synthesis, Speech Recognition
A hypertext version of the FAQ is provided by the Speech Group at Carnegie Mellon University, Pittsburgh.

Source, Availability

"`http://www.speech.cs.cmu.edu/comp.speech`"

5.2.2.4 Usenet Frequently Asked Questions (FAQ) List

This ftp site contains the Frequently Asked Questions (FAQ) list of many usenet newsgroups.

Source, Availability

"`ftp://rtfm.mit.edu/pub/usenet/`"

5.2.3 Education and conferences

The field of SLP requires expertise in the areas of phonetics, physics, computer science, linguistics, psychology, and physiology – among others. In many educational institutions, courses on SLP are offered in the context of these related areas.

Only recently dedicated SLP courses have been offered, mainly in the form of tutorials at speech related conferences, summer schools, or workshops. It can be expected that out of these courses full university curriculae will develop.
The most important international conferences for SLP are the ICSLP, the ICASSP, and Eurospeech. In 1998, LREC, the first international conference on Language Resources and Evaluation was held. At this conference, which gave a good overview of existing resources and ongoing projects, both speech and language processing were present.

5.2.3.1 European Student Journal on Language and Speech

The European Student Journal on Language and Speech is an online publication. It is a common initiative by EACL, ESCA and ELSNET.
The Journal explicitly encourages graduate students and students in postgraduate master courses to submit manuscripts.

Source, Availability

"`http://web-sls.essex.ac.uk/web-sls/`"

5.2.3.2 Eurospeech

Bi-annual international conference on spoken language processing and phonetics organised by ESCA. Eurospeech is focused on applied SLP research and development and features extensive SLP technology demonstrations, e.g. the Eurospeech '97 Olympics for telephone operated dialogue systems.

5.2.3.3 ICASSP

International Conference on Acoustics, Speech, and Signal Processing – annual conference organised by the IEEE. The main focus of ICASSP is signal processing in general, but many sessions and presentations are devoted to speech and SLP.

5.2.3.4 ICPhS

International Congress of the Phonetic Sciences held every four years. ICPhS is focused on phonetics and basic research, but it has sessions on technology as well.

5.2.3.5 ICSLP

Bi-annual International Conference on Spoken Language Processing. Like Eurospeech, ICSLP is focused on the applied research in SLP and SLP product development.

5.2.3.6 Audio–Visual Speech Processing

AVSP is a conference on audio-visual speech processing. Auditory-visual speech production and perception by human and machine is an interdisciplinary and cross-linguistic field.
AVSP is an annual satellite conference to the major speech related conferences.

Table 5.2: SLP related journals

Name	ISSN	Publisher
International Journal of Speech Technology	1381-2416	Kluwer Academic Publishers
Speech Communication		Elsevier
Asia Pacific Journal of Speech, Language, and Hearing		Allen Press
Language and Speech	0023 8309	Kingston Press
Journal of Phonetics	0095 4470	Academic Press
Phonetica	0031 8388	Karger
Computational Linguistics	08912017	MIT Press

5.2.3.7 Journals

Table 5.2 is only a very short list of the most important SLP related journals. A more extensive list can be found in the `comp.speech` FAQ.

5.2.3.8 Survey of the state of the art in Human Language Technology (1996)

The goal of the survey on Human Language Technology is to provide an overview of the main areas of work, the capabilities and limitations of current technology, and the technical challenges that must be overcome to realise the vision of graceful human–computer interaction using natural communication skills. The survey is available online.
The HLT survey was supported by the National Science Foundation, the European Union, CSLU of the Oregon Graduate Institute, and the University of Pisa.

Source, Availability

"`http://cslu.cse.ogi.edu/HLTsurvey`"

5.2.3.9 Spoken language processing: A primer

A concise overview of SLP resources on the WWW, by Mark Liberman of LDC and Ron Cole of the CSLU at the Oregon Graduate Institute (OGI).

Source, Availability

"`http://www.ldc.upenn.edu/myl/LR_background.html`"

5.2.3.10 SLP courses at OGI

CSLU offers laboratory short courses on building spoken dialogue systems and text-to-speech synthesis. These courses, offered during the summer months, provide students with both theoretical background in areas of language technology, and hands-on experience developing spoken language systems using the CSLU Toolkit.

Source, Availability

"`http://cslu.cse.ogi.edu/courses/shortcourses.html`"

5.2.3.11 ELSNET summer schools

The European Network in Language and Speech organises summer schools that focus on selected SLP topics, e.g. lexicon development for language and speech processing, or multimodality in language and speech systems.
These summer schools in general last two weeks per course; emphasis is placed on active student participation, i.e. by having student presentation sessions, and tutorial sessions where students can work under the supervision of the lecturer.

Source, Availability

"`http://www.elsnet.org`"

5.2.3.12 Summer Institute of Linguistics

The Summer Institute of Linguistics (SIL) is hosted by the International Linguistics Center in Dallas, Texas. The SIL organises courses and workshops and provides a rich source of SLP tools.

Source, Availability

"`http://www.sil.org`"

5.3 "SLP at Work"

SLP technology currently (1999) is being used primarily for man–machine interfaces and telecommunications applications, and as an enabling technology for new services and speech research and other research areas.

5.3.1 Speech interfaces

Man–machine interfaces traditionally consist of keyboards, pointing devices, buttons or switches. Speech interfaces may provide a more natural and comfortable means of communicating with a device, and may replace or complement the other interface modalities. For blind or motorically impaired persons, speech may be the only interface modality available.
In a speech interface a speech recogniser analyses human speech and maps it to an internal representation that triggers the execution of some action by the machine. A speech synthesiser generates speech which is then output to the user.
Speech recognisers can be categorised by the size of the vocabulary, speaking style, signal bandwidth, and speaker dependency.
The recognition performance of a speech recogniser is usually expressed by the *word error rate* (WER), i.e. the percentage of incorrectly recognised words in an utterance. It depends on the technical quality of sound input (high bandwidth vs. telephone vs. mobile phone quality), the type of speech (isolated words vs. connected speech; formal vs. casual style, standard vs. dialect pronunciation),

environment noise (background speech and noise vs. quiet environment; close-talk vs. table-top microphone), the complexity of the task (small vs. large vocabulary) and of the interaction (master/slave vs. interactive communication), and others.
Today, speech recognisers are mostly based on Hidden Markov Models (HMMs), i.e. a statistical approach; neural nets are less common. HMM recognisers for large vocabularies are usually based on phonemes, whereas HMM recognisers for small vocabularies are based on entire words. The speech signal is split into overlapping frames, and signal parameters are computed for each frame. These frames are then analysed by the HMM which assigns a label to each frame. An *alignment algorithm* (Viterbi algorithm) maps the label sequence to the entries of a dictionary. For disambiguation, higher-level knowledge, e.g. a statistical language model, morphological knowledge, or a grammar, is used.
In general small vocabulary speaker dependent speech recognisers have a low WER of about 2%; they can be successfully used to operate single-task devices, e.g. equipment in an operation theatre in a hospital or form-based technical supervisions. Large vocabulary and speaker independent speech recognisers have a WER of about 25% under regular conditions; dictation systems achieve a WER of better than 10% for limited vocabularies, e.g. medical or juridical, and with close-talk microphones.
Speech synthesisers generate speech either completely synthetically or by concatenating fragments of prerecorded human speech. Synthetic speech synthesis uses a formal and abstract model of human speech production. This model is controlled by parameters. Early text-to-speech systems were of this type; speech is generated directly from a text representation by converting the text into a phonemic representation which is then used to set the parameters of the synthesiser. In the last years, speech synthesisers that use prerecorded human speech have been developed. To overcome the finite size of a prerecorded vocabulary, the speech is cut into segments which are then concatenated for output. The goal is to find signal fragments that can be concatenated with as little mismatch as possible.
For speech synthesisers to sound natural simply producing speech from text strings is not sufficient. Stress and intonation variations are as important, and hence sophisticated speech synthesisers apply intonation contours to the speech output or vary the speaking rate. These modifications are applied to utterances as a whole, and thus they cannot be represented on the phoneme or word level, but on a higher phrase or even sentence level.
Speech synthesisers can now be found in many devices and also toys.

5.3.2 Telecommunications and broadcast

The most significant change in the telecommunications and broadcast industries is the transition from analogue to digital transmission. This transition is fundamental: Audio and video are now treated simply as data. As a consequence, telecommunication networks and broadcast networks can now be based on the protocols of data networks (see 5.5.2).
Digital data can be compressed. Lossless compression retains the original information, whereas in lossy compression some information is lost. For audio and video data, some loss of information may be acceptable – humans can adapt

very well to noisy speech or low quality video.
The role of SLP in this arena is the development of encoders and decoders or *codecs* for speech data. Codecs are implemented either in hardware, e.g. in mobile phones, or in software, e.g. as plug-ins for WWW browsers.
Codecs that allow speech encoding at low data rates create opportunities for new data transmission channels: mobile phone or satellite telephony operate on very low bandwidth, and Internet telephony uses codecs that can adapt to the available bandwidth (within user defined limits).

5.3.3 New services

In the context of SLP, a speech driven service is a service that formerly was provided by human operators. In a speech driven service, speech is not necessarily processed locally (as in speech interfaces), but may be processed remotely. Finally, in speech driven services the amount of speech processing required is much higher than in speech interfaces.
Many of the services traditionally provided by humans were automated with SLP technology. Dictation systems are now commonplace in offices, and simple telephone based information systems that once required human operators or touch tone operation now understand speech. It must be noted that simply installing SLP technology does not necessarily improve a service. However, it may lead to redesigning the service and in the course of this process create new services that did not exist without SLP.
Two examples of such new services are dictation and translation servers, and automated call centers and answering machine servers. Dictation and translation servers take audio files created by speaking into a PC or a digital dictation device as input, and return a transcript in some specified format, e.g. a business letter formatted for a particular word processor. All exchange of data is handled via e-mail, so that it can be accessed from all over the world, around the clock.
In automated call centers, incoming calls are analysed and either processed locally, or forwarded to the intended recipient. Local processing ranges from call completion, where the destination of the call is determined by asking the caller to speak the recipient's name, to voice-dialling where the caller simply picks up the phone and speaks the name instead of dialling. Answering machine servers are provided by telecommunication companies as a service to customers; they store incoming calls and faxes and can execute actions based on the content of the messages.
New speech processing services will continue to appear. The distinction between interfaces and services is bound to become meaningless because on the one hand processing power within devices increases, and on the other hand advanced networking will make remote access as simple and common as accessing local resources.

5.3.4 SLP as a research tool

SLP is a research area in its own right, but SLP technology has yet to become a tool for research work. The general principle is to use SLP tools in a bootstrapping manner: collect speech resources to build SLP technology, and then apply this technology to create further resources.

For example, the current performance of speech recognisers was reached by training them on large speech databases. These speech databases consist of the speech proper and annotation data. They were created by recording large numbers of speakers, either directly, or via the telephone. These speech recognisers can now be used to support the annotation of future data collections.
The effort that goes into the annotation outweighs by far the original recording effort – an orthographic annotation of read speech takes ten times the duration of the speech, for a narrow phonetic annotation one to five hundred times is common. SLP tools could help to reduce the time spent on annotation, e.g. by determining the speech parts of a signal that contains noise, by providing a first version of an orthographic transcription, or by a semi-automatic segmentation and labelling procedure where user input by the annotator, e.g. setting a boundary, automatically starts relabelling the remaining signal.

5.3.4.1 Online speech recognition

There are only very few online speech recognisers accessible in the WWW – the main problem being the transfer of speech data to the recogniser. One such system is provided by the University of Erlangen, Germany, at "`http://www5.informatik.uni-erlangen.de/HTML/German/Thesis/rrgruhnsa/Spracherkenner.html`"

5.3.4.2 Online speech synthesis

Some of the speech synthesisers that can be tested online via the WWW are listed below.

Lucent Bell Labs: American English, German, Mandarin, Chinese, Spanish, French, Italian
"`http://www.bell-labs.com/project/tts/index.html`"
Gerhard-Mercator-University Duisburg: German, English, Japanese
"`http://www.fb9-ti.uni-duisburg.de/demos/speech.html`"
ICP Grenoble: French
"`http://www.icp.inpg.fr/cgi-bin/synthese`"
Bonn University: German
"`http://asl1.ikp.uni-bonn.de/ tpo/Hadiq.en.html`"
ETH Zurich: German
"`http://www.tik.ee.ethz.ch/cgi-bin/w3svox`"
AT&T Labs: American English
"`http://www.research.att.com/projects/tts/`"
CSTR Edinburgh: British English
"`http://www.cstr.ed.ac.uk/projects/festival/userin.html`"
British Telecom: British English
"`http://innovate.bt.com/showcase/laureate/index.htm`"
University of Delaware: American English
"`http://www.asel.udel.edu/speech/Dsynterf.html`"
University of York: British English
"`http://www-users.york.ac.uk/ lang4/Yorktalk.html`"
Mons Polytechnicum: Arabic, Brazilian, Breton, Croatian, German, Estonian, Spanish, French, Greek, Italian, Dutch, Romanian, Swedish, British English, American English

"http://tcts.fpms.ac.be/synthesis/mbrola.html"
Microsoft Corp.: American English
"http://research.microsoft.com/stg/ssproject.htm"
Apple Computer Inc.: American English, Mexican Spanish
"http://www.apple.com/macos/speech/"
SoftVoice Inc.: American English
"http://www.text2speech.com/"
Eloquent Technology: American English
"http://www.eloq.com/"
AIST Nara University: Japanese
"http://www.aist-nara.ac.jp/IS/Shikano-lab/database
/lecture/SS/voice_of_computer/e-voice_of_computer.html"

5.3.4.3 Operational SLP devices and products

The original list of SLP devices and Products was originally compiled by Russ Wilcox ("rwilcox@tiac.net") at "http://www.tiac.net/users/rwilcox/speech.html". His WWW pages provide an up-to-date overview of SLP related resources.

Lernout & Hauspie: Foreign languages and speech recognition, speech synthesis and speech compression products
http://www.lhs.com/
Dragon Systems: Dictation software, speech recognition
http://www.dragonsys.com/
Verbex Voice Systems: Speech recognition
http://www.verbex.com/
Microsoft Research Speech Technology Group: Whisper speech synthesis, SAPI (Speech Application, Programming Interface) SDK
http://www.research.microsoft.com/research/srg/
Command Corporation: Speech recognition
http://www.commandcorp.com/incube_welcome.html
Northern Telecom: Speech recognition
http://www.nortel.com/
STAR Lab: Speech recognition
http://www-speech.sri.com/
SpeechWorks: Telephone speech recognition
http://www.speechworks.com/index_ns.html
IBM VoiceType Dictation: Dictation software, speech recognition
http://www.software.ibm.com/workgroup/voicetyp/
AT&T Advanced Speech Products Group: SAPI-compatible speech recognition and speech synthesis, speaker verification
http://www.att.com/aspg/
ART: Handwriting and speech recognition
http://www.artcomp.com/
Voice Control Systems: Telephone speech recognition
http://www.voicecontrol.com
Nuance: Speech recognition
http://www.nuancecom.com/
Fonix: Speech synthesis, dictation system
http://www.fonix.com/

Locus Speech Corporation: Telephone speech recognition
http://www.locus.ca/
A&G Graphics Interface: Speech recognition SDKs
http://www.customvoice.com
Vocalis: Speech-computer telephony products
http://www.vocalis.com/
Sensory Inc.: Speech recognition and synthesis hardware
http://www.sensoryinc.com
Philips Speech Processing: Speech recognition and speech-enabled telephony and consumer products
http://www.speech.be.philips.com/
Defense Group Inc. (DGI): Robust speech recognition
http://www.ca.defgrp.com/noise.html
Oki Semiconductor: Speech DSPs
http://www.okisemi.com/public/fm/Home_c.html
Speech Solutions: Speech-enabled Active-X controls
http://www.speechsolutions.com/
BaBel Technologies SA: Speech recognition and synthesis
http://www.babeltech.com/

5.4 SLP procedures, tools, and formats

In SLP, just like in other engineering fields, the creation of resources consists of four main phases: design, production, validation and distribution. For all phases, the use of suitable tools and the implementation of procedures is recognised as 'good practice'. However, the more innovative a project, the fewer tools and procedures are available.
Often the phases are not strictly sequential but there are iterations: a first design is tested in a small-scale production; the validation of this production leads to a modification of the design, etc.
In the design phase, the specifications are laid out and written down in a specification report. These specifications cover both the format and the contents of the resource to be created. The format specification describes the technical setup, data organisation, storage, and implementation issues, as well as the annotation format and procedure. The contents specification describes the actual contents of the resource – type, quantity, and quality of the speech material, demographic and administrative data, etc.
The production phase consists of a data collection and an annotation task. For the data collection, speakers have to be recruited, prompting material has to be produced, and the speech material has to be recorded. In order to detect deviations from the specifications as early as possible, the data collection must be monitored closely. These checks can be performed automatically, e.g. by logging the data that is critical to the success of the data collection. Annotation can begin as soon as some data has been collected. Again, close monitoring is strongly recommended. In general, only formal aspects of the annotation can be monitored automatically. A data production report describes the data collection and annotation.
Validation can be internal, i.e. performed by the resource producer, or external, i.e. by an independent agent. It is necessary that the validator discusses

any problems with a resource with its producer prior to the publication of the validation report because there may be good reasons for particular decisions.
Finally, SLP resources are distributed. A complete distribution consists of the resource data, documentation, program sources to access the resource data, and the validation report. Either an SLP agency, such as ELRA, LDC (see Section 5.2.1) is charged with this task, or the producer distributes the resource himself. Agencies have the advantage that they take care of all contractual issues, and that they have duplication facilities and experience with distribution channels.

5.4.1 Annotation

Annotation is the process of obtaining a symbolic representation from signal data. At the very minimum, annotation data for speech corpora consists of an orthographic transcription of the recorded speech and a pronunciation lexicon. Annotation can be performed on different levels of representation. These levels can be arranged hierarchically by the proximity to the speech signal: Closest to the signal is the phonetic segmentation, where signal fragments are labelled with a phonetic symbol (usually in IPA notation) representing a speech sound. This phonetic segmentation is time-aligned, i.e. every label also contains a signal address. Phonemic transcriptions usually are not time-aligned. A phonemic transcription is derived (automatically) from the orthographic transcription by looking up the word items in a pronunciation dictionary. The orthographic transcription usually contains markers for non-speech items, e.g. noise. In a prosodic annotation, markers for the rise and fall of intonation are inserted into the orthographic transcription. Other annotation levels are syntax structures, part-of-speech tagging, discourse representation, or dialogue structure annotation.
For the exchange of SLP resources, annotations must meet two requirements: first, their representation must allow a mapping to other levels, e.g. for each phonetic segment it should be possible to retrieve the orthographic word it belongs to. Second, any annotation format must be defined formally and must be accompanied by tools to access and edit annotation data. A recent proposal for a formal framework for linguistic annotations that can model many of the existing annotations within a single framework has been proposed by Bird and Liberman (1999).

5.4.1.1 Partitur-Format

The Partitur-Format is a multi-tier annotation format which allows the alignment of time-aligned signal data and symbolic representations by symbolic markers (Schiel et al. 1997). The KAN (for "kanonisch") tier is mandatory, it contains the canonical or citation form phonemic representation of the words of the current utterance, plus a numerical index.
All other tiers are optional. If a tier is present, it must include a reference to the KAN tier. The alignment between a phonetical tier and the KAN tier is achieved via an extended phonetic label which includes an explicit link to a phoneme, e.g. ['a:-'E:] if the phoneme /'a:/ from the citation form is produced and labelled as the phone ['E:].

Source, Availability

"http://www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html#Partitur"

5.4.2 Validation, evaluation

A speech resource needs to be validated to check whether it complies with its specifications. Formal specifications can be checked automatically, whereas the contents of a resource can in most cases be validated only by human experts. A validation procedure must meet several goals – some of which may be in conflict: it must be efficient, but deliver good results; it should be standardised, but also adaptable to different requirements. Finally, a validation must be reliable to become an accepted standard.
Ideally, a validation is performed by an agent who is independent of the resource producer. Validation centers, i.e. SLP labs that have built up expertise in the area of validation, are such independent agents. It is strongly recommended to make use of validation centers because this guarantees a minimum standard of quality and gives credibility to the producer's quality claims.
The availability of human experts is crucial to a validation. Validation centers in general provide such experts either directly or by contractual assignment. However, for rare languages or non-standard or highly innovative resources the only experts available may be the producers themselves – in such a case, the resource creator must provide a complete log of all resource related activities.
Especially for large resources, not all material can be validated. The usual procedure is to select a subset of the material, either randomly or guided by knowledge, e.g. experience from similar validations. The selection of material is not necessarily proportional to the size of the resource because there exist lower limits for the significance of tests, and upper limits for the feasibility of a validation. All activities related to the validation are logged and summarised in a validation report.
In an evaluation, the suitability of an SLP resource for a particular task is measured. Evaluations are carried out by end users who need to determine whether a given resource meets their requirements, or by SLP agencies who are interested in comparing SLP resources against each other. The best known evaluation in the SLP community are the DARPA competitions, where there is one common resource, a task to be performed, and many different competing approaches to solve the problem.
To make an evaluation fair, the SLP resource is usually divided into a training and a test set – these sets can be disjoint, but need not be. The training set may be used by all participants to train their system, and the test set is used for the actual system performance evaluation. The definition of both the training and test set is usually public in order to ensure a maximum transparency of the evaluation.
Evaluations can be only one measure amongst many others – for example, the word error rate of a speech recogniser does not say anything about the overall performance of a dialogue system, e.g. at an automated information kiosk. Nevertheless the evaluation results are often cited to characterise (and sell) an SLP resource or an SLP system.

5.4.2.1 Validation manual

Via ELRA, SPEX (Speech Expertise Centre) in the Netherlands has published a manual for the validation of speech resources. This document is based on the experience gained in the validation of the SpeechDat telephone speech database collection during which more than 65,000 speakers in 16 countries were recorded.

Source, Availability

"`http://www.icp.grenet.fr/ELRA/home.html`"

5.4.3 Tools and standards

SLP tools are basically editors or processors which are particularly suited for SLP applications. Such tools exist for all representation levels of SLP data.
Signal editors present the signal in such a way that human experts can manipulate a speech or a derived signal by signal analysis, signal modification, signal transformation, etc. Signal processors operate in the background and in general do not have a visible interface, e.g. speech codecs that compress speech signals to minimise bandwidth requirements, or signal filters. The outcome of both signal editors and processors is again a signal.
A phonetic editor at least consists of an oscillogram and a sonagram display, and an editing field for the phonetic transcription. It supports segmenting and labelling of a speech signal, and features a phonetic alphabet and formal consistency checks for the transcription. Sophisticated phonetic editors offer a variety of signal display types, and may contain a large number of embedded signal processors such as filters, or speech recognition or synthesis modules.
Phonemic editors must contain an editing field and a pronunciation lexicon. Such a lexicon ideally contains not only the canonic pronunciation of an item, but also the most frequent pronunciation variants, and additional information, e.g. morphological, syntactical, semantical and other information. These different types of information must be stored in such a way that they can be accessed independently of each other.
Similarly, editors and processors exist for the higher speech related representations, e.g. prosody, syntax, etc.

5.4.3.1 CHILDES

The child language data exchange system (CHILDES) is a computerised exchange system for language data. It was originally developed within the field of child language to foster the sharing of transcribed language data of children's spontaneous speech. CHILDES consists of an archive of recordings and annotations, and of an annotation and processing software to create and access CHILDES data.

Source, Availability

"`http://ipra-www.uia.ac.be/ipra/childes.html`"

5.4.3.2 EMU

EMU is a collection of software for the creation, manipulation and analysis of speech databases. At the core of EMU is a database search engine which allows the researcher to find various speech segments based on the sequential and hierarchical structure of the utterances in which they occur. EMU includes an interactive labeller which can display spectrograms and other speech waveforms, and which allows the creation of hierarchical, as well as sequential, labels for a speech utterance.

Source, Availability

"`http://www.shlrc.mq.edu.au/emu/index.html`"

5.4.3.3 ESPS Waves

ESPS/waves+ is a suite of programs used for the analysis and display of speech signal data. It includes a collection of programs to assist in computing spectra, analysing speech, converting data, and applying time-referenced labels. By means of a flexible, open interface to the Entropic Signal Processing System (ESPS) it can easily be customised.

Source, Availability

"`http://www.entropic.com`"

5.4.3.4 HTK

HTK (Hidden Markov Toolkit) is a toolkit for building Hidden Markov Models (HMM). HMMs can be used to model any time series; they have been particularly successful in speech recognition.
HTK runs under Unix and Linux and is a commercial software.

Source, Availability

"`http://www.entropic.com`"

5.4.3.5 SFS (Speech Filing System)

SFS provides a computing environment for SLP research. It comprises software tools, file and data formats, subroutine libraries, graphics, standards and special programming languages. It performs standard operations such as acquisition, replay, display and labelling, spectrographic and formant analysis and fundamental frequency estimation.
SFS is copyrighted University College London, and is currently supplied free of charge to research establishments for non-profit use.

Source, Availability

"`ftp://pitch.phon.ucl.ac.uk/pub/sfs`"

5.4.3.6 TRANSCRIBER

TRANSCRIBER is a tool for segmenting, labelling, and transcribing speech. It is written in Tcl/tk script language and is freely available as free software. TRANSCRIBER allows segmenting, labelling, and transcribing long duration signals. The output is in a standard SGML format. Multiple languages are supported. The tool can be ported to various platforms and is very flexible so that new functions can be easily added.

Source, Availability

"`http://www.etca.fr/English/Projects/Transcriber`"

5.4.3.7 Signalyze

Signalyze is a powerful data analysis, display, segmentation and labelling software for speech signal processing on the Macintosh.

Source, Availability

"`http://www.epfl.ch/`"

5.4.3.8 WWWTranscribe

WWWTranscribe is a transcription system based on the WWW. It is platform independent and allows network access to speech databases. It consists of a number of template HTML files and cgi-scripts written in perl that instantiate the template files with current variable values. Its modular structure makes it flexible, and it connects easily to existing signal processing applications or database management systems.

Source, Availability

"`http://www.speechdat.org/Tools/WWWTranscribe`"

5.4.3.9 MAUS

MAUS is an automatic segmentation and labelling tool for speech verification. Its primary feature is a generator of pronunciation variants for a given utterance; these variants are stored as a hypothesis graph. A standard Viterbi alignment then finds the best path through the graph.

Source, Availability

"`http://www.phonetik.uni-muenchen.de/`"

5.4.3.10 Praat

Praat is a powerful signal analysis, annotation tool, and speech synthesis developed at the Phonetics department of Amsterdam University. It runs under Windows, UNIX, and Macintosh.

Source, Availability

"`http://fonsg3.hum.uva.nl/praat/praat.html`"

5.4.3.11 CSLU Toolkit

The CSLU Speech Toolkit is a comprehensive software environment for research, development, and education of spoken language systems. It integrates a set of core technologies including speech recognition, speech synthesis, facial animation and speaker recognition. It also features authoring and analysis tools enabling quick and easy development of desktop and telephone-based speech applications.
The software is available free of charge for research and education at non-profit institutions. A restricted evaluation copy is also available for personal and commercial use. All toolkit use is covered by a license agreement.

Source, Availability

"`http://www.cse.ogi.edu/CSLU/toolkit/toolkit.html`"

5.4.3.12 SOX

SOX ("Sound Exchange") is a versatile tool for converting between various audio formats. It can read and write various types of audio files, and optionally applies some special effects (e.g. echo, channel averaging, or rate conversion).

Source, Availability

"`http://www.spies.com/Sox`"

5.4.3.13 UNIX tools

The operating system UNIX has introduced a number of very powerful concepts: devices and files are treated alike, and process in- and output can be piped from one process to another one.
The UNIX tools *grep*, *sed*, and *awk* (Aho et al. 1987) are powerful filter programs that analyse and modify a data stream by applying regular expressions. A regular expression is a text pattern that consists of normal characters and meta characters with a special meaning. A regular expression matches a fragment of the data stream to which it is applied if the characters in the data stream can be made to fit the pattern in the regular expression. *lex* (Levine et al. 1995) is a UNIX tool to build lexical analysers (or *tokenisers*).
Regular expressions have the expressive powers of finite automata, i.e. they can be used as tokenisers for lexical items, but they cannot represent arbitrary bracketed structures. Despite this limitation, the three UNIX commands are very useful for low-level text manipulation, e.g. formatting.
Other UNIX commands useful to SLP work are *sort*, *comm*, and *diff*. *sort* sorts a data stream, *comm* extracts the lines that two data streams have in common, and *diff* computes the difference between two data streams.
These UNIX commands come with UNIX installations; most of them are available for all other platforms as well.

5.4.3.14 Grammars

Context-free grammars have the expressive power to count brackets: $a^n b^n$. They are thus suitable for the creation and parsing of nested marker formalisms. Context-sensitive grammars have an even greater expressive power: $a^n b^n c^n$. However, they are less common in SLP applications.
A number of grammar implementations is available. The best known is probably *yacc* (Levine et al. 1995), which is used for building parsers. *yacc* is available for free for many platforms. Another formalism are Definite Clause Grammars (DCGs) (Pereira and Shieber 1987). Most Prolog implementations can use DCGs directly.
Grammars are given as grammar rules, with a left hand side (or *head*) and a right hand side (*body*). A *terminal* is a symbol that stands for itself, a *non-terminal* is substituted by other non-terminal terminal symbols.
In a top–down parser, the head is substituted by its body, and the body elements then become the new elements to be substituted. This process continues until either all symbols are terminals that correspond to the input string – parsing was successful – or no more rules can be applied – parsing failed.
In a bottom-up parser, a right hand side is selected that matches part of the input string. The matching terminal symbols are replaced by the head of the rule, and the selection process continues. Parsing is successful if the top-most rule of the grammar (the *axiom*) is reached.

5.4.4 Text

An SLP annotation basically is a text, and as such it uses an alphabet to build lexical items which are organised according to syntactic constraints.
Any text can be described either by its structure, or by its layout. SLP texts, e.g. annotations, lexica, etc. have an explicit structure – in fact this structure distinguishes them from texts in general. An explicit structure consists of markers and the marked text. Markers and marked text must always be distinguishable unambiguously by formal procedures. Markers may be nested to allow complex structures.
Using explicit markers leads to two distinct types of document: a template document, i.e. a kind of marker dictionary that defines the allowed markers and their nesting relationships, and a document instance, i.e. a document that contains an actual marker structure and marked text. Clearly, every representation level used in SLP annotations must be defined by a template document, and actual annotations then are applications of this template.
The Standard Generalized Markup Language (SGML) is an ISO standard for describing text through its structure.
Hierarchically organised symbolic annotation levels used in SLP can be described with a single template document. A typical example is the hierarchical relationship between phonemes and words, words and phrases, or phrases and sentences. If such a hierarchical relationship between annotation levels does not exist – which is often the case in SLP – then distinct template documents are needed for every annotation level. This is the case e.g. for phonetic segments and phonemes where several distinct phonemes may have been realised by a single phone because of coarticulation, or for truly independent phenomena,

such as background noise occurring during an utterance.

5.4.4.1 SGML

SGML (Standard Generalized Markup Language) is a specification for describing the structure of a text. In SGML, a DTD (Document Type Definition) defines the markers and their syntax for each document type, e.g. letter, technical report, etc. A document instance then is an application of SGML.
For the creation and manipulation of SGML formatted documents a number of freeware and commercial software tools is available.
SGML has been standardised by the ISO as ISO 8879, and it is a standard that is difficult to implement fully.

Source, Availability

"`http://www.iso.ch`"

5.4.4.2 XML

The Extended Markup Language (XML) is a subset of SGML. Its goal is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML has been designed for ease of implementation and for interoperability with both SGML and HTML.

Source, Availability

"`http://www.w3.org/TR/REC-xml`"

5.4.4.3 HTML

HTML, the HyperText Markup Language, is the publishing language of the World Wide Web. HTML supports text, multimedia, and hyperlink features, plus scripting languages, and style sheets. HTML is based on ISO 10639 to allow international code tables.
HTML is an application of SGML, and its current version is HTML 4.0.

Source, Availability

"`http://www.w3.org/TR/REC-html40/`"

5.5 Technology

This section contains references to material relevant not only to SLP, but to a broad range of applications:

- alphabets,
- network technology,
- file formats,
- programming languages and
- storage technology.

The material presented here is either an official ISO standard, or a de facto standard. Often the original standard has been defined and proposed by a research laboratory, a company, or an agency, and once the standard has become widely accepted, it has been adopted by the ISO.

5.5.1 Alphabets

Alphabets have been covered extensively in Appendix A of the EAGLES *Handbook of Standards and Resources for Spoken Language Systems* (Gibbon et al. 1997). Only material that has changed since then is included here.

5.5.1.1 Phonetic fonts

Phonetic fonts are required to render phonetic transcriptions on the screen or on paper. Unfortunately, font management for different platforms is not trivial, and porting a document from one word processor to another or from one platform to another may result in malformatted or illegible texts.
There are many sources for phonetic fonts: one of the most widespread phonetic fonts is that of the SIL, which is available for both Windows and Macintosh. Other font sources are the ftp archives, or dedicated font sites in the WWW. Commercial sources for fonts are Adobe Corporation, Linguist's Software, and others.

5.5.2 Networks

Networks connect computers to allow the exchange of data. The Internet is a heterogeneous network that consists of subnetworks connected to each other via gateways.
A network has a physical topology and possibly multiple logical topologies. The physical topology is determined by the cables or other communication media connecting the computers; local area networks today are based on twisted-pair Ethernet with a raw transfer rate of up to 100 Mb/s, wide area networks use fibre optic cables with transfer rates of several Gb/s. In remote areas of the world, a network can also be established via telephone lines, e.g. satellites.
The ISO has defined a 7 layer model for networks: the lower layers describe media access, e.g. how bits are fed into the medium, the middle layers describe the transfer of logical units, e.g. TCP/IP data packets, and the highest layers make up the application layer that deals with inter-application communication. The logical topology of a network is determined by the protocol that is used to transmit data via the network. There exist many different protocols, and they may share the same medium. The most widespread protocol is TCP/IP (transmission control protocol/internet protocol).
A second important network is the public telephone network. It spans the entire globe, and in fact many subnetworks of the Internet are connected by telephone lines. The telephone network is being converted from analogue to digital technology, and ISDN (Integrated Services Digital Network) has been deployed in many industrial countries. Mobile phone networks are an alternative to fixed telephone networks – they either rely on earth-bound radio transmitters or satellites. Telephone networks are important to SLP because many speech operated services will make use of the telephone, and the ubiquity of telephones

makes data collection feasible even for smaller languages or remote locations.

5.5.2.1 IETF

The Internet Engineering Task Force is a "loosely self-organised group of people who make technical and other contributions to the engineering and evolution of the Internet and its technologies".
IETF holds regular meetings, and maintains mailing lists. Proposals to the IETF are commonly published as numbered RFCs (Request for Comments); they are then discussed and approved. There are two types of RFC: FYIs (For Your Information) are introductory texts, whereas STDs (Standards) are real Internet standards.

Source, Availability

IETF: "`http://www.ietf.org`"
RFC editor: "`http://www.rfc-editor.org`"

5.5.2.2 TCP/IP

TCP/IP (transmission control protocol/internet protocol) is the most widespread protocol for computer networks. It was designed in the late 1960s and has shown to be sufficiently robust, simple and scalable. In TCP/IP, data is split into small packages labelled with the recipient's Internet address. An Internet address is a four-tuple of numbers from 0 to 255, i.e. a 32 bit address (allowing for roughly 2 billion different addresses). Routers read the address label of a package and pass it on to the next known router; at the receiving end, all packages are collected and reordered to that the original message can be restored.
TCP/IP is now being revised by the IETF, the Internet Engineering Task Force. The main goal is to provide more IP addresses by up to 128 address bits, to support Multicast, and to allow quality of service guarantee for connections. This is especially important for high bandwidth transmission, e.g. speech or video, where a minimum throughput must be ensured. This new protocol is called IPv6 (for version 6) or IPng (for next generation), and it is specified in RFC 1883 (see page 311).

Source, Availability

"`http://www.ietf.org`", follow the RFC links

5.5.2.3 MIME

MIME (Multi-purpose Internet Mail Extension) is a standard document descriptor. It consists of a document type description, e.g. *text*, *audio*, *application*, and a format description, e.g. *ISO-Latin-1*, *wav*, *javascript* separated by a slash "/" and followed by two new-line characters. The MIME type is transferred together with a document, and the receiving application interprets the MIME information to see whether it can handle the document itself or needs to call an external helper application.

Originally, MIME was specified in RFC 1341 by N. Borenstein and N. Freed. Since then, many RFCs have extended MIME.

Source, Availability

"`http://www.ietf.org/`", follow the RFC links

5.5.2.4 WWW

The World Wide Web (WWW) is a client–server system. A client (WWW browser) requests a document via a URL (Uniform Resource Locator) from a WWW server. A URL has the form

```
protocol://address:port/path/file#anchor?value_list
```

with

- `protocol`: an Internet protocol such as `http`, `ftp`, `news`, etc.
- `address`: either an IP-number or IP-address
- `port`: an operating system communications port number
- `path`: a path name relative to the web server's root directory
- `file`: a file name
- `anchor`: a named position within the file
- `value_list`: a list of attribute–value pairs written as *attribute=value*; attribute–value pairs are separated by &.

URLs can be partial only – missing parts are substituted with default values by the server. The server interprets the URL and returns the requested document to the client.
`http` is the protocol of the WWW. It defines the communication between client and server.

Source, Availability

"`http://www.w3.org/TR/http1.1`"

5.5.2.5 WWW browsers

WWW browsers, e.g. Netscape Navigator, Internet Explorer, Lynx, or Opera are basically viewers for HTML documents transferred using the `http` protocol. Most browsers can be obtained free of charge; Netscape has released the source code of its browser to allow developers world wide to optimise and extend the browser.
There are differences in each browser's implementation of the HTML document format, the JavaScript scripting language, or the support of style sheets. For truly platform independent code, only the common subset can be used.
Lynx is a text-only browser that can be run in command shells and terminal emulators.
The capabilities of browsers can be enhanced by plug-ins: multi-media I/O, PDF viewing, etc. The original plug-in API was proposed by Netscape and it has been adopted by the other browsers.

Source, Availability

Table 5.3: WWW browsers

Browser	Platform
Lynx "`http://lynx.browser.org/`"	Unix, VMS
Microsoft Internet Explorer "`http://www.microsoft.com/ie`"	Mac, Windows
Netscape Navigator "`http://www.netscape.com`"	Mac, UNIX, Windows
Opera "`http://www.opera.com`"	Windows
StarOffice "`http://www.sun.com/dot-com/staroffice.html`"	Windows, UNIX

5.5.2.6 WWW servers

WWW servers are available for all platforms, both as commercial and as shareware or freeware software. All servers naturally provide the basic `http` capabilities; most also have non-standardised system administration features that allow remote administration. Server-side applets, so-called *servlets*, usually implemented in Java, can provide services otherwise not offered by the server, e.g. data transfer between an applet and the file system of the server.
Often servers are part of a larger application, e.g. the Oracle Web server as part of the Oracle Relational DBMS, or WebCompanion in the FileMaker Pro DBMS, etc.
In the Apache project, the source for the powerful Apache web server is freely available.

Source, Availability

See Table 5.4

Table 5.4: Web server

Server	Platform	URL
Apache	Unix	"`http://www.apache.org`"

5.5.2.7 ISO 3166 country codes

The ISO has given two-letter mnemonic codes to all countries of the world. These two letter codes are used, amongst other purposes, for the top-level domain names for the individual countries.

Afghanistan	af
Albania	al
Algeria	dz
American Samoa	as

Andorra	ad
Angola	ao
Anguilla	ai
Antarctica	aq
Antigua and Barbuda	ag
Argentina	ar
Armenia	am
Aruba	aw
Australia	au
Austria	at
Azerbaijan	az
Bahamas	bs
Bahrain	bh
Bangladesh	bd
Barbados	bb
Belarus	by
Belgium	be
Belize	bz
Benin	bj
Bermuda	bm
Bhutan	bt
Bolivia	bo
Bosnia and Herzegowina	ba
Botswana	bw
Bouvet Island	bv
Brazil	br
British Indian Ocean Territory	io
Brunei Darussalam	bn
Burkina Faso	bf
Cambodia	kh
Canada	ca
Cayman Islands	ky
Central African Republic	cf
Chad	td
Chile	cl
China	cn
Christmas Island	cx
Cocos (Keeling) Islands	cc
Colombia	co
Comoros	km
Congo	cg
Cook Islands	ck
Costa Rica	cr
Cote D'ivoire	ci
Croatia	hr
Cuba	cu
Cyprus	cy
Czech Republic	cz
Denmark	dk
Djibouti	dj
Dominica	dm
Dominican Republic	do

East Timor	tp
Ecuador	ec
Egypt	eg
El Salvador	sv
Equatorial Guinea	gq
Eritrea	er
Estonia	ee
Ethiopia	et
Falkland Islands (Malvinas)	fk
Faroe Islands	fo
Fiji	fj
Finland	fi
France	fr
France, Metropolitan	fx
French Guiana	gf
French Polynesia	pf
French Southern Territories	tf
Gabon	ga
Gambia	gm
Georgia	ge
Germany	de
Ghana	gh
Gibraltar	gi
Greece	gr
Greenland	gl
Grenada	gd
Guadeloupe	gp
Guam	gu
Guatemala	gt
Guinea	gn
Guinea-bissau	gw
Guyana	gy
Haiti	ht
Heard and Mc Donald Islands	hm
Honduras	hn
Hong Kong	hk
Hungary	hu
Iceland	is
India	in
Indonesia	id
Iran (Islamic Republic Of)	ir
Iraq	iq
Ireland	ie
Israel	il
Italy	it
Jamaica	jm
Japan	jp
Jordan	jo
Kazakhstan	kz
Kenya	ke
Kiribati	ki
Korea, Democratic People's Republic Of	kp

Korea, Republic Of	kr
Kuwait	kw
Kyrgyzstan	kg
Lao People's Democratic Republic	la
Latvia	lv
Lebanon	lb
Lesotho	ls
Liberia	lr
Libyan Arab Jamahiriya	ly
Liechtenstein	li
Lithuania	lt
Luxembourg	lu
Macau	mo
Macedonia, The Former Yugoslav Republic Of	mk
Madagascar	mg
Malawi	mw
Malaysia	my
Maldives	mv
Mali	ml
Malta	mt
Marshall Islands	mh
Martinique	mq
Mauritania	mr
Mauritius	mu
Mayotte	yt
Mexico	mx
Micronesia (Federated States Of)	fm
Moldova, Republic Of	md
Monaco	mc
Mongolia	mn
Montserrat	ms
Morocco	ma
Mozambique	mz
Myanmar	mm
Namibia	na
Nauru	nr
Nepal	np
Netherlands	nl
Netherlands Antilles	an
New Caledonia	nc
New Zealand	nz
Nicaragua	ni
Niger	ne
Nigeria	ng
Niue	nu
Norfolk Island	nf
Northern Mariana Islands	mp
Norway	no
Oman	om
Pakistan	pk
Palau	pw
Panama	pa

Papua New Guinea	pg
Paraguay	py
Peru	pe
Philippines	ph
Pitcairn	pn
Poland	pl
Portugal	pt
Puerto Rico	pr
Qatar	qa
Reunion	re
Romania	ro
Russian Federation	ru
Rwanda	rw
St. Helena	sh
Saint Kitts and Nevis	kn
Saint Lucia	lc
St. Pierre and Miquelon	pm
Saint Vincent and The Grenadines	vc
Samoa	ws
San Marino	sm
Sao Tome and Principe	st
Saudi Arabia	sa
Senegal	sn
Seychelles	sc
Sierra Leone	sl
Singapore	sg
Slovakia	sk
Slovenia	si
Solomon Islands	sb
Somalia	so
South Africa	za
South Georgia and The South Sandwich Islands	gs
Spain	es
Sri Lanka	lk
Sudan	sd
Suriname	sr
Svalbard and Jan Mayen Islands	sj
Swaziland	sz
Sweden	se
Switzerland	ch
Syrian Arab Republic	sy
Taiwan, Province of China	tw
Tajikistan	tj
Tanzania, United Republic Of	tz
Thailand	th
Togo	tg
Tokelau	tk
Tonga	to
Trinidad and Tobago	tt
Tunisia	tn
Turkey	tr
Turkmenistan	tm

Turks and Caicos Islands	tc
Tuvalu	tv
Uganda	ug
Ukraine	ua
United Arab Emirates	ae
United Kingdom	uk
United States	us
United States Minor Outlying Islands	um
Uruguay	uy
Uzbekistan	uz
Vanuatu	vu
Vatican City State (Holy See)	va
Venezuela	ve
Vietnam	tn
Virgin Islands (British)	vg
Virgin Islands (U.S.A.)	vi
Wallis and FutunaIslands	wf
Western Sahara	eh
Yemen	ye
Yugoslavia	yu
Zaire	zr
Zambia	zm
Zimbabwe	zw

5.5.2.8 ISO 639 language codes

The ISO has specified two-letter mnemonic codes for the languages of the world.

Abkhazian	ab
Afar	aa
Afrikaans	af
Albanian	sq
Amharic	am
Arabic	ar
Armenian	hy
Assamese	as
Aymara	ay
Azerbaijani	az
Bashkir	ba
Basque	eu
Bengali	bn
Bhutani	dz
Bihari	bh
Bislama	bi
Breton	br
Bulgarian	bg
Burmese	my
Byelorussian	be
Cambodian	km
Catalan	ca
Chinese	zh
Corsican	co
Croatian	hr

Czech	cs
Danish	da
Dutch	nl
English	en
Esperanto	eo
Estonian	et
Faeroese	fo
Farsi	fa
Fiji	fj
Finnish	fi
French	fr
Frisian	fy
Galician	gl
Georgian	ka
German	de
Greek	el
Greenlandic	kl
Guarani	gn
Gujarati	gu
Hausa	ha
Hebrew	iw
Hindi	hi
Hungarian	hu
Icelandic	is
Indonesian	in
Interlingua	ia
Interlingue	i.e.
Inupiak	ik
Irish	ga
Italian	it
Japanese	ja
Javanese	jw
Kannada	kn
Kashmiri	ks
Kazakh	kk
Kinyarwanda	rw
Kirghiz	ky
Kirundi	rn
Korean	ko
Kurdish	ku
Laothian	lo
Latin	la
Latvian	lv
Lingala	ln
Lithuanian	lt
Macedonian	mk
Malagasy	mg
Malay	ms
Malayalam	ml
Maltese	mt
Maori	mi
Marathi	mr

Moldavian	mo
Mongolian	mn
Nauru	na
Nepali	ne
Norwegian	no
Occitan	oc
Oriya	or
Oromo	om
Pashto	ps
Polish	pl
Portuguese	pt
Punjabi	pa
Quechua	qu
Rhaeto-Romance	rm
Romanian	ro
Russian	ru
Samoan	sm
Sangro	sg
Sanskrit	sa
Scots-Gaelic	gd
Serbian	sr
Serbo-Croatian	sh
Sesotho	st
Setswana	tn
Shona	sn
Sindhi	sd
Singhalese	si
Siswati	ss
Slovak	sk
Slovenian	sl
Somali	so
Spanish	es
Sudanese	su
Swahili	sw
Swedish	sv
Tagalog	tl
Tajik	tg
Tamil	ta
Tatar	tt
Tegulu	te
Thai	th
Tibetan	bo
Tigrinya	ti
Tonga	to
Tsonga	ts
Turkish	tr
Turkmen	tk
Twi	tw
Ukranian	uk
Urdu	ur
Uzbek	uz
Vietnamese	vi

Volapuk	vo
Welsh	cy
Wolof	wo
Xhosa	xh
Yiddish	ji
Yoruba	yo
Zulu	zu

5.5.2.9 Telephone

ISDN (Integrated Services Digital Network) is a world-wide standard for digital telephony. A BRI (base rate interface) consists of two 64 Kbit data channels and a 16 Kbit command channel, a PRI (primary rate interface) has 30 data and 2 command channels.
ISDN can be used for voice telephony and data communication. In telephony, speech is converted to digital format in the ISDN handset and is then transmitted digitally in a-law (Euro-ISDN) or μ-law format (US-ISDN). These logarithmic compression schemes allow a dynamic range of 12 bit quantisation in 8 bits with little loss of signal quality.

Source, Availability

Your local telecom.

5.5.2.10 CAPI

Common-ISDN-API (CAPI) is an application programming interface standard used to access ISDN equipment connected to basic rate interfaces (BRI) and primary rate interfaces (PRI).

Source, Availability

"`http://www.capi.org`"

5.5.2.11 GSM

GSM (Global System for Mobile communications) is a standard for digital transmission for mobile telephony. It was proposed by ETSI, the European Telecommunications Standards Institute, and is now installed in large parts of the world.
The frequency bands available to GSM telephony are 900 and 1800 MHz. GSM requires speech codecs that can transfer speech data through a limited bandwidth of about 13 Kb/s. Signalling data, e.g. dial tones, are transmitted in an extra signalling channel so that they do not get distorted by the codecs.

Source, Availability

"`http://www.etsi.fr`"

5.5.2.12 DECT

DECT means Digital Enhanced Cordless Telecommunications. It is a European standard for local mobile communications. DECT allows a high density of users, high quality of speech because it features a bandwidth of 32 Kb/s and discontinuous transmission to save battery power.

Source, Availability

"`http://www.etsi.fr`"

5.5.3 File formats

Audio and video data is stored in many different file formats. On early computers, sound output capabilities were limited to 8 bit quantisation and a sample rate between 5.5 and 8 KHz (resulting in a data rate of 5.5 and 8 KB/s). With more computing power available, higher data rates could be supported – nowadays, CD-quality audio (stereo audio channels with 16 bit quantisation and 44.1 or 48 KHz sample rate, i.e. 192 KB/s data rate) can be handled even by low end PCs.
Because of the multi-media aspects of SLP processing file formats capable of storing multi-media data are becoming increasingly important. These file formats must support different types of data, e.g. audio, video, text and signal data. Ideally, they should be suitable for data transmission, data storage, and easy processing.
File formats can be divided into raw data formats and meta formats. Raw data formats store data of one type, and may not require any header at all, or only a minimal header describing the data. However, the burden of accessing the data is placed on the user – every new file format requires writing new access procedures.
Meta formats combine a variety of raw signal formats with a set of standardised and platform independent access interfaces for programming languages or applications. Meta formats are flexible in that they allow the incorporation of new data types, e.g. compressed data, and provide means of access to this data at the same time.

5.5.3.1 Audio formats

Audio formats are described in detail in the Audio file formats FAQ (see Section 5.2.2.1).

5.5.3.2 QuickTime

QuickTime is a meta file format for multi-media data and a toolbox for accessing this data. QuickTime was developed by Apple Computer, and is available for both Windows and Macintosh operating systems (for other operating systems, a subset of the QuickTime functionality is accessible).
The basic metaphor underlying QuickTime is that of a multi-track recording, where each track may contain text, graphics, audio, or video data in a large variety of formats, including streaming audio and video, and MPEG data. The

tracks are synchronised, and may be switched on or off for playback, e.g. to play movies in different languages.
The current version is QuickTime 3.0, and simple players and plug-ins for web browsers can be downloaded free of charge. QuickTime is supported by most multi-media editing tools, and a system development kit may be licensed from Apple.

Source, Availability

"`http://quicktime.apple.com/`"

5.5.3.3 MPEG

A committee called the Motion Picture Experts Group (MPEG) has proposed a family of standards for multi-media file formats. MPEG is now an ISO standard. MPEG-3 is defined specifically for audio data. It is a lossy compression scheme that results in very low data rates ($\approx 10\%$ of the data rate of audio CD-ROMs) at little or no perceivable loss of quality.

Name	Media	Description	Data rate
MPEG-1	audio, video	video recorder or standard TV quality data	< 4 Mb/s
MPEG-2	audio, video	high definition TV (HDTV) quality data	2–15 Mb/s
MPEG-3	audio	low data rate, high quality audio	8–320 Kb/s
MPEG-4	audio, video	low quality, very low data rate for videoconferencing via telephone or ISDN lines	8–64 Kb/s

Source, Availability

MPEG Web site: "`http://www.mpeg.org`"
MPEG-3 Web site: "`http://www.iis.fhg.de/amm`"

5.5.3.4 PostScript

PostScript is a language for describing the page layout of documents. It is platform independent and has become the de facto standard language for laser printers. Word processors, graphics applications, etc. create PostScript files which are then transferred to a printer. PostScript features a font inclusion mechanism so that a document can be printed on any suitable printer.
PostScript was developed by Adobe Corporation. The current version is PostScript level 3. PostScript files can be viewed with the popular freeware software Ghostview, but in general they cannot be edited once they have been created.

Source, Availability

5.5.3.5 Portable Document Format

The Portable Document Format (PDF) is a language for describing the page layout of documents combined with the ability to perform text searches in the document, dynamic linking of documents, multi-media content, and input via forms, e.g. for interactive documents. PDF files generally are much smaller than PostScript, and they may be edited. PDF has become the most widespread format for online manuals and document collections on CD-ROM, e.g. conference proceedings.
PDF has been developed by Adobe Corporation. A PDF viewer software Acrobat is freely available for almost every platform. For the creation of PDF formatted documents a commercial software is needed.

Source, Availability

"`http://www.adobe.com`"

5.5.4 Programming

Any SLP work requires a substantial amount of programming. High-level programming languages facilitate the development of software by

- object-oriented or modular design,
- strict type checking at compile time, and
- built-in consistency mechanisms such as automatic memory allocation and garbage collection

Programming languages can be classified as either embedded languages, script languages, or programming languages proper. Embedded languages run inside an application, e.g. a PC database system, word processor, or a web browser. Script languages may run on their own and call other applications or even access functions inside these applications. Programming languages proper are the classical programming languages which are used to implement applications in the first place.
Database Management Systems are software systems designed to safely store large amounts of data and to provide guarded access to this data. DBMSs range from small, single-user and PC-based to large, distributed and multi-user on workstations or mainframes. Relational databases were developed in the late 1970s and are now commonplace; modern object-oriented DBMSs are beginning to penetrate the market. SQL is the de facto standard language for relational databases and SQL-3 is currently being standardised by the ISO; important new features are the computation of the transitive closure, and object-oriented concepts. An initiative by researchers and developers of object-oriented DBMSs has resulted in the ODMG, which has published a draft standard for an object-oriented high-level database query language called OSQL.

5.5.4.1 Database Management Systems

Databases are covered in detail in the EAGLES *Handbook of Standards and Resources for Spoken Language Systems* (Gibbon et al. 1997).

A source of information about object-oriented DBMSs is the Object Database Management Group, a consortium of researchers and providers of DBMSs that has defined a common object-oriented DB query language OQL and has specified the minimum requirements for object-oriented DBMSs.

Source, Availability

"`http://www.odmg.org/`"

5.5.4.2 Java

Java is an object-oriented programming language developed by Javasoft of SUN Microsystems. It has become the de facto standard programming language for applets, i.e. programs which are distributed over the WWW to run inside a WWW browser.
Java features a large class library including classes for graphical display, and audio data access. The Java Speech API specification supports voice command recognition, dictation, and text-to-speech synthesis.
The current version of Java is 2.0 (version 1.2 was renamed to 2.0).

Source, Availability

"`http://www.javasoft.com/`"
More information about the Java Speech API may be found at
"`http://java.sun.com/products/java-media/speech/`"

5.5.4.3 C++

C++ is the standard object-oriented programming language for standalone applications both for SLP and other purposes. It was specified by B. Stroustroup (Stroustroup 1991).
C++ is being standardised by the ISO.

Source, Availability

Commercial and freeware C++ compilers are available for every platform.

5.5.4.4 perl

perl is an interpreted programming language designed for rapid programming of scripts; its main features are powerful text manipulation operations such as regular expressions, associative arrays, and ease of system access, e.g. for file and directory access and manipulation.
perl is freely available for almost every platform; the current version number is 5. Because of its powerful text operators and system access it has become the most commonly used language for programming cgi-applications for the WWW.

Source, availability

"`http://www.perl.org`"

5.5.4.5 python

python is a modern object-oriented programming language designed to overcome the limited data modelling capabilities of perl. One of its distinguishing features is the built-in interface to many windowing environments.
Python is freely available for most platforms.

Source, Availability

"`http://www.python.org`"

5.5.4.6 JavaScript, ECMAScript

JavaScript is a script language that runs inside WWW browsers. Client-side computations are implemented in JavaScript, e.g. consistency checkers for form input.
JavaScript by Netscape and J-Script by Microsoft are not completely compatible.
ECMAScript is the script language proposed by the ECMA consortium (European Computer Manufacturer's Association) to establish a common script language for all browsers.

Source, Availability

The specification of JavaScript can be found at
"`http://developer.netscape.com/docs`",
J-Script at "`http://msdn.microsoft.com/scripting/`", and
ECMAScript at "`ftp://ftp.ecma.ch/ecma-st/e262-pdf.pdf`"

5.5.4.7 tcl/tk

tcl/tk is a graphical toolbox to the tcl scripting language. tcl/tk allows "glueing" together applications and provides a graphical interface to these applications, e.g. buttons, menus, and windows.
tcl/tk is available for all platforms.

Source, Availability

tcl was originally proposed by John Osterhout (Osterhout 1994). tcl/tk is available at many ftp archives.

5.5.5 Storage

SLP data processing requires very much storage capacity. Storage devices are classified by their capacity, access speed, and whether the storage medium is removable or not. On the one hand the general cost of storage space is decreasing, but on the other hand there is an increasing demand of storage space from new applications – including SLP.

5.5.5.1 RAID

RAID means Redundant Array of Inexpensive Disks. In a RAID array, several hard disks are combined in such a way that failure or removal of a disk does not interrupt the operation of the array as a whole. This is possible by distributing data over the individual hard disks, and by data duplication.
Several RAID levels have been specified. They differ in the degree of redundancy and safety.

5.5.5.2 DVD

DVD (digital versatile disk) is an advanced optical medium. DVDs have the same size as CD-ROMs (5 1/4"), but have up to two data layers on both sides, and a higher storage density. They can store up to roughly 18 GB on one disk. DVD was originally devised for entertainment purposes (full size video films) and thus has the same structural problems as CDs (helical track, constant angular velocity, i.e. variable disk rotation speed). For entertainment media content, DVDs can be marked with a country code that allows this medium to be played only in a region with the correct code.
DVD is backward compatible so that DVD drives can read DVD, DVD-ROM, and traditional CD-ROMs.
DVD-RAM is a phase change medium with a large capacity, but it is incompatible with DVD or DVD-ROM.

Bibliographical references

References

C. D. I. 14915 (1998). Multimedia user interface design – Software ergonomic requirements.

C. Abry and M. Lallouache (1995). Le MEM: Un modèle d'anticipation paramétrable par locuteur. Données sur l'arrondissement en français. *Bulletin de la Communication Parlée* 3: 85–99.

A. Adjoudani (1996). *Reconnaissance audiovisuelle de la parole.* Ph.D. thesis, Institut National Polytechnique, Grenoble, France.

A. Adjoudani and C. Benoît (1996). On the integration of auditory and visual parameters in an HMM-based ASR. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines, Models, Systems, and Applications*, volume 150 of *Computer and Systems Sciences*, Berlin. NATO ASI Series, Springler-Verlag.

A. Adjoudani, T. Guiard-Marigny, B. LeGoff, L. Reveret and C. Benoît (1997). A multimedia platform for audio-visual speech processing. In: *Eurospeech'97.*

A. Aho, B. Kernighan and P. Weinberger (1987). *The AWK Programming Language.* Addison Wesley, Reading.

J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, E. Maier, N. Reithinger, B. Schmitz and M. Siegel (1997). Dialogue acts in VERBMOBIL-2. VM-Report 204, DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken.

J. Allen, W. Bradford, E. Ringger and T. Sikorshi (1996). A robust system for natural spoken dialogue. In: *Proceedings of the Annual Meeting*, pp. 62–70. Association for Computational Linguistics.

N. Allison, A. Ellis, B. Flude and A. Luckman (1992). A connectionist model of familiar face recognition. *IEEE Colloquium on Machine Storage and Recognition of Faces* 5: 1–10.

B. Altenberg (1990). Spoken English and the dictionary. In: J. Svartvik, ed., *The London-Lund Corpus of Spoken English: Description and research*, Lund Studies in English 82, pp. 275–286. Lund University Press, Lund.

A. Anderson, M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson and R. Weinert (1991). The HCRC Map Task Corpus. *Language and Speech* 34(4): 351–366.

E. André, W. Finkler, W. Graf, T. Rist, A. Schauder and W. Wahlster (1993). WIP: The automatic synthesis of multimodal presentations. In: M. Maybury, ed., *Intelligent Multimedia Interfaces*, pp. 75–93. AAAI Press.

E. André and T. Rist (1993). The design of illustrated documents as a planning task. In: M. Maybury, ed., *Intelligent Multimedia Interfaces*, pp. 94–116. AAAI Press. Also DFKI Research Report RR-92-45.

N. Arends (1993). *The visual speech apparatus: An aid for speech training.* Ph.D. thesis, Nijmegen University/NICI, Sint Michielsgestel (NL): Instituut voor Doven.

Y. Arens, E. Hovy and M. Vossers (1993). On the knowledge underlying multimedia presentations. In: M. Maybury, ed., *Intelligent Multimedia Interfaces*, pp. 280–306. AAAI Press.

M. Argyle and M. Cook (1976). *Gaze and Mutual gaze.* Cambridge University Press.

R. Arntz and H. Picht (1989). *Einführung in die Terminologiearbeit.* Studien zu Sprache und Techni. Olms, Hildesheim, New York, Zürich.

A. Arvaniti (1994). Acoustic features of Greek rhythmic structure. *Journal of Phonetics* 22: 239–268.

J. Austin (1962). *How to do things with words.* Clarendon Press, Oxford.

C. Avesani (1990). A contribution to the synthesis of Italian intonation. In: *Proc ICSLP 90*, volume 1, pp. 833–836, Kobe, Japan.

P. Badin and C. Abry (1996). Articulatory synthesis from X-rays and inversion for an adaptive speech robot. In: *Proceedings of ICSLP'96: The Fourth International Conference on Spoken Language Processing.*

N. Badler, R. Bindiganavale, Bourne, Palmer, Shi and B. Webber (1998). A parametrized action representation for virtual human agents. In: *WECC'98, The First Workshop on Embodied Conversational Characters.*

B. Bahan (1996). *Non-Manual Realization of Agreement in American Sign Language.* Ph.D. thesis, Boston University, Boston, MA.

B. Bailey, J. Konstan, R. Cooley and M. Dejong (1998). Toolkit for building interactive multimedia presentations. In: *ACM Multimedia 98 – Electronic Proceedings*, "`http://www.acm.org/sigmm/MM98/electronic_proceedings/bailey/index.html`".

G. Bailly (1996). Building sensori-motor prototypes from audiovisual exemplars. In: *Proceedings of ICSLP'96: The Fourth International Conference on Spoken Language Processing.*

S. Balbo, J. Coutaz and D.Salber (1993). Towards automatic evaluation of multimodal user interfaces. In: *Conference on Intelligent User Interfaces*, pp. 201–208.

G. Ball and Breese (1998). Emotion and personality in a conversational character. In: *WECC'98, The First Workshop on Embodied Conversational Characters.*

J. Ball and D. Ling (1994). Spoken language processing in the Persona conversational assistant. In: *Lifelike Computer Characters'94.*

J. Barnett, P. Bamberg, M. Held, J. Huerta, L. Manganaro and A. Weiss (1995). Comparative performance in large vocabulary isolated-word recognition in five European languages. In: *Proceedings Eurospeech 1995, Madrid, Spain*, pp. 189–192.

S. Basu and A. Pentland (1997). Recovering 3D lip structure from 2D observations using a model trained from video. In: *AVSP'97 workshop*, Rhodos, Greece.

J. Bates (1994). Realism and believable agents. In: *Lifelike Computer Characters'94.*

B. Batliner, H. Block, A. Kießling, R. Kompe, H. Niemann, E. Nöth, T. Ruland and S. Schacht (1997). Improving parsing of spontaneous speech with the help of prosodic boundaries. VM-Report 210, F.-A.-Universität Erlangen–Nürnberg/Siemens AG, München.

R. Battison (1975). Phonological deletion in American Sign Language. *Sign Language Studies* 5: 1–19.

S. Bayer, R. Kozierok and J. Kurtz (1995). Multimodal interfaces on the web. In: *Third Python Workshop.*

M. Bearne, S. Jones and J. Sapsford-Francis (1994). Towards usability guidelines for multimedia systems. In: *Proceedings of the second ACM International Conference on Multimedia (MULTIMEDIA'94)*, pp. 105–110, San Francisco.

G. Beattie (1981). Sequential temporal patterns of speech and gaze in dialogue. In: T. Sebeok and J. Umiker-Sebeok, eds., *Nonverbal Communication, Interaction, and Gesture*, pp. 297–320. The Hague, New-York.

M. Beckman and G. Ayers Elam (1997). Guidelines for ToBI labelling. March 1997, Ohio State University.

Y. Bellik (1996). MEDITOR: A multimodal text editor for blind users. In: *ACM UIST96, Ninth Annual Symposium on User Interface Software*, Seattle, USA.

Y. Bellik and D. Teil (1992). Les types de multimodalité. In: *Actes des Quatrièmes journées sur l'ingénierie des interfaces homme-machine (IHM'92)*, pp. 22–28. Telecom Paris: 92S004, Paris.

A. Benguerel and H. Cowan (1974). Coarticulation of upper lip protrusion in french. *Phonetica* 30: 40–51.

C. Benoît (1990). Why synthesize talking faces? In: *Proceedings of the ESCA*

Workshop on Speech Synthesis, pp. 253–256, Autrans. ESCA.

C. Benoît, T. Guiard-Marigny, B. LeGoff and A. Adjoudani (1996). Which components of the face do humans and machines best speechread? In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO ASI Series. Series F: Computer and Systems Sciences*, pp. 315–325. Springler-Verlag, Berlin.

C. Benoît, T. Lallouache, T. Mohamedi, A. Tseva and C. Abry (1990). Nineteen (+- two) french *visemes* for visual speech synthesis. In: *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans. ESCA.

C. Benoît and L. Pols (1992). On the assessment of synthetic speech. In: *Talking machines: Theories, Models, and Designs*, pp. 435–442. North Holland, Amsterdam.

R. Benzmüller and M. Grice (1997). Trainingsmaterialien zur Etikettierung deutscher Intonation mit GToBI. Phonus 3, Institute of Phonetics, University of the Saarland. pp. 9–34.

N. Bernsen (1996). Towards a tool for predicting speech functionality. "`http://cse.ogi.edu/CSLU/fsj/issues/issue1/`".

N. Bernsen (1997). A toolbox of output modalities: Representing output information in multimodal interfaces. The Maersk Mc-Kinney Moller Institute for Production Technology, Odense University, Denmark.

L. Bernstein (1991). Lipreading Corpus V-VI: Disc 3 and Corpus VII-VIII: Disc 4, Washington, DC.

L. Bernstein, E. Auer and P. Seitz (1996a). Lipreading Corpus XI-XII: Disc 6. Vowel Stimuli - Two Talkers. Los Angeles, CA: House Ear Institute.

L. Bernstein, E. Auer and P. Seitz (1996b). Lipreading Corpus XIII-XIV: Disc 7. Medial Consonant Stimuli - Two Talkers. Los Angeles, CA: House Ear Institute.

L. Bernstein, E. Auer and P. Seitz (1996c). Lipreading Corpus XV-XVI: Disc 8. Final Consonant Stimuli - Two Talkers. Los Angeles, CA: House Ear Institute.

L. Bernstein and S. Eberhardt (1986). Johns Hopkins Lipreading: Corpus I-II: Disc 1 and Corpus III-IV: Disc 2, Baltimore, MD.

L. Bernstein, P. Seitz and E. Auer (1995). Lipreading Corpus IX-X: Disc 5. Initial Consonant Stimuli - Two Talkers. Washington, DC.

M. Berthod and J. Maroy (1979). Learning in syntactic recognition of symbols drawn on a graphic tablet. *Computer Graphics Image Processing* 9: 166–182.

J. Beskow (1995). Rule-based visual speech synthesis. In: *ESCA - EUROSPEECH '95. 4th European Conference on Speech Communication and Technology*, Madrid.

J. Beskow (1997a). Animation of talking agents. In: *AVSP'97 workshop*, Rhodos, Greece.

J. Beskow (1997b). Olga - A conversational agent with gestures. In: *Proc. of IJCAI'97 - Workshop on Animated Interface Agents - Making them intelligent*, Nagoya, Japan.

J. Beskow, M. Dahlquist, B. Granström, M. Lundeberg, C. Spens and T. Öhman (1997). The teleface project - disability, feasability and intelligibility. In: *PHONUM 4/1997; Proceedings of Fonetik 97, Swedish Phonetics Conference.*

J. Beskow, K. Elenius and S. McGlashan (1996). Olga - A dialogue system witn an animated talking agent. In: *International Conference on Spoken Language Processing*, volume 3, pp. 1651–1654, Philadelphia (PA). IEEE Computer Society.

D. Beymer (1995). *Pose-Invariant face recognition using real and virtual views.* Ph.D. thesis, M.I.T.

D. Beymer and T. Poggio (1995). Face recognition from one example view. Technical Report A.I. Memo No. 1536, M.I.T., A.I. Lab., Cambridge.

D. Biber, S. Johansson, G. Leech, S. Conrad and E. Finegan, eds. (1999, forthcoming). *The Longman grammar of spoken and written English*. Longman, London.

G. Bieger and M. Glock (1986). Comprehending spatial and contextual information in picture-text instructions. *The Journal of Experimental Education* 54(4): 181–188.

K. Binsted (1998). Designing portable characters. In: *WECC'98, The First Workshop on Embodied Conversational Characters*.

S. Bird and M. Liberman (1999). A formal framework for linguistic annotation. Technical report ms-cis-99-01, Department of Computer and Information Science, University of Pennsylvania, USA.

R. Birdwhistell (1952). *Introduction to kinesics, an annotation system for analysis of body motion and gesture*. University of Louisville.

A. Blake and M. Isard (1994). 3D position, attitude and shape input using video tracking of hands and lips. *Computer Graphics Proceedings, Annual Conference Series* pp. 185–192.

G. Blakowski and R. Steinmetz (1996). A media synchronization survey: Reference model, specification, and case studies. *IEEE JSAC* .

M. Blattner and R. Dannenberg (1990). CHI '90 workshop on multimedia and multimodal interface design. *SIGCHI Bulletin* 22(2): 54–57.

M. Blattner and E. Glinert (1996). Multimodal integration. *Multimedia* 3(4): 14–24.

T. Blum, D. Keislar, J. Wheaton and E. Wold (1997). Audio databases with content-based retrieval. In: M. Maybury, ed., *Intelligent Multimedia Information Retrieval*, pp. 113–135. AAAI Press.

K. Bohm, W. Hubner and K. Vaananen (1992). GIVEN: Gesture Driven Interactions in Virtual Environments - A toolkit approach to 3D interactions. In: *Interantional Conference on Interface to Real and Virtual Worlds (Informatique '92)*, Montpellier (France).

D. Bolinger (1989). *Intonation and its Uses*. Stanford University Press.

R. Bolt (1980). Put-that there: Voice and gesture at the graphics interface. *Computer Graphics Journal of the Association of Computing and Machinery* 14(3): 262–270.

R. Bolt (1987). The integrated multi-modal interface. *The Transactions of the Institute of Electronics, Information and Communication Engineers* J79-D(11): 2017–2025. Japan.

M. Bordegoni, G. Faconti, M. Maybury, T. Rist, S. Ruggieri, P. Tahanias and M. Wilson (1997). A standard reference model for intelligent presentation systems. In: *IJCAI'97 workshop on Intelligent Multimodal Systems*, pp. 85–99, Nagoya, Japan.

E. Bos, C. Huls and W. Claassen (1994). EDWARD: Full integration of language and action in a multimodal user interface. *Int. J. Human-Computer Studies* 40: 473–495.

D. Boston (1973). Synthetic facial animation. *British Journal of Audiology* 7: 373–378.

H. Bothe (1996). Relations of audio and visual speech signals in a physical feature space: Implications for the hearing-impaired. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO ASI Series. Series F: Computer and Systems Sciences*, pp. 445–460. Springler-Verlag, Berlin.

C. Bregler, M. Covell and M. Stanley (1997). Video rewrite: Driving visual speech with audio. *Computer Graphics Annual Conference Series* .

C. Bregler and Y. Konig (1994). Eigenlips for robust speech recognition. In: *Proceedings of the Int. Conf. on Acoustics Speech and Signal Processing (IEEE-ICASSP)*, Adelaide, Australia.

C. Bregler, S. Manke, H. Hild and A. Waibel (1993). Improving connected letter recognition by lipreading. In: *International Conference on Acoustic, Speech, and Signal Processing*, Minneapolis (MN).

S. Brennan and J. Ohaeri (1994). Effect of message style on users' attributions toward agents. In: *CHI' 94 Conference Companion Human Factors in Computing Systems*, pp. 281–282, Boston. ACM Press.

R. Briggs, B. Beck, A. Dennis, E. Carmel, J. Nunamaker and R. Pfarrer (1992). Is the pen mightier than the keyboard? In: *25th Hawai International Conference on Systems Sciences*, volume 3, pp. 201–210, Kauai (HI). IEEE.

N. Brooke (1996). Using the visual component in automatic speech recognition. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1656–1659, Philadelphia, PA.

N. Brooke and S. Scott (1994). Animated computer graphics of talking faces based on stochastic models. In: *Proceedings of the International Symposium on Speech, Image-processing and Neural Networks*, pp. 73–76, Hong Kong. IEEE.

N. M. Brooke and E. D. Petajan (1986). Seeing speech: Investigations into the synthesis and recognition of visible speech movements using automatic image processing and computer graphics. In: *Proceedings of the International Conference on Speech Input/Output: Techniques and Applications*, pp. 104–109, London, UK. Institution of Electrical Engineers.

R. Brunelli and T. Poggio (1993a). Caricatural effects in automated face perception. *Biological Cybernetics* 69: 235–241.

R. Brunelli and T. Poggio (1993b). Face recognition: Features versus templates. *IEEE Trans. Pattern Analysis and Machine Intelligence* 15(10): 1042–1052.

U. Bub, M. Hunke and A. Waibel (1995). Knowing how to listen to in speech recognition: Visually guided beamforming. In: *International Conference on Acoustics, Speech and Signal Processing*, Detroit (MI).

G. Budin (1996). *Wissensorganisation und Terminologie: die Komplexität und Dynamik wissenschaftlicher Informations- und Kommunikationsprozesse*. Günter Narr Verlag, Tübingen.

M. Bunt (1989). Speech is more than just an audible version of text. In: F. N. M.M. Taylor and D. Bouwhuis, eds., *The structure of multimodal dialogue*, pp. 287–299. Elsevier Science Publishers B.V. (North Holland).

G. Burdea (1996). Multimodal virtual reality: Input-output devices, system integration, and human factors. *HCI* 8(1): 5–24.

J. Burger and R. Marshall (1993). The application of natural language models to intelligent multimedia. In: M. Maybury, ed., *Intelligent Multimedia Interfaces*, pp. 174–196. AAAI Press.

S. Burger and H. G. Tillmann (1997). Comparison of commercial dictation systems for personal computers. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)* 35: 107–114.

L. Burnard, ed. (1995). *Users' reference guide for the British National Corpus version 1.0*. Oxford University Computing Services, Oxford.

H. Bussmann (1990). *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart.

B. Butterworth and U. Hadar (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review* 96(1): 168–174.

J. Carletta (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2): 249–254.

J. Carletta, A. Isard, S. Isard, J. Kowtko, A. Newlands, G. Doherty-Sneddon and

A. Anderson (1995). *HCRC Dialogue Structure Coding Manual.* Human Communication Research Centre, 2 Buccleugh Place, Edinburgh EH8 8LW, Scotland.

J. Carletta and J. Taylor (1996). The SGML representation of the HCRC Map Task Corpus. Technical Report, Human Communication Research Centre, 2 Buccleugh Place, Edinburgh EH8 8LW, Scotland.

J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost and M. Stone (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Computer Graphics Annual Conference Series* pp. 413–420.

L. Cerrato, F. Leoni and A. Paoloni (1997). A methodology to quantify the contribution of visual and prosodic information to the process of speech comprehension. In: *AVSP'97 workshop*, Rhodos, Greece.

D. Chandramohan and P. Silsbee (1996). A multiple deformable template approach for visual speech recognition. In: *Proceedings of ICSLP'96: The Fourth International Conference on Spoken Language Processing.*

C. Chaudron (1988). *Second Language Classrooms: Research on teaching and learning.* Cambridge University Press, Cambridge.

R. Chellappa, C. Wilson and S. Sirohey (1995). Human and machine recognition of faces: A survey. *Proceedings of the IEEE* 83(5): 705–740.

H. Chen, T. Chen, B. Haskell, A. Kaplan, S. Keshav and E.Petajan (1994). Audio-assisted video coding/processing. In: *ISO/MPEG meeting*, Paris, France.

A. Cheyer (1997). MVIEWS: Multimodal tools for the video analyst. In: *International Conference on Intelligent User Interfaces*, pp. 55–62, San Francisco (CA).

A. Cheyer and L. Julia (1995). Multimodal maps: An agent-based approach. In: *International Conference on Cooperative Multimodal Communication*, Eindhoven (NL).

N. Chovil (1989). *Communicative Functions of Facial Displays in Conversation.* Ph.D. thesis, University of Victoria.

N. Chovil (1991). Social determinants of facial displays. *Journal of Nonverbal Behavior* 15(3): 141–154.

E. Churchill, S. Prevost, T. Bickmore, P. Hodgson, T. Sullivan and Cook (1998). Design issues for situated conversational characters. In: *WECC'98, The First Workshop on Embodied Conversational Characters.*

R. Cipolla, Y. Okamoto and Y. Kuno (1993). Robust structure from motion using motion parallax. In: *International Conference on Computer Vision*, pp. 374–382. IEEE.

J. Clark and C. Yallop (1995). *An introduction to phonetics and phonology.* Blackwell, Oxford UK; Cambridge USA, 2nd edition.

Cockpit (1996). Commissie ontwikkeling defensie materieel (codema) project "toepassing automatische spraakherkenning in de cockpit".

E. Codd (1970). A relational model of data for large shared data banks. *CACM* 13(6).

J. Cohen (1988). *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

M. Cohen and D. Massaro (1990). Synthesis of visible speech. *Behavioral Research Methods and Instrumentation* 22(2): 260–263.

M. Cohen and D. Massaro (1993). Modeling coarticulation in synthetic visual speech. In: M. Magnenat-Thalmann and D. Thalmann, eds., *Models and Techniques in Computer Animation*, Tokyo. Springer-Verlag.

M. Cohen, R. Walker and D. Massaro (1996). Perception of synthetic visual speech. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines,*

Models, Systems, and Applications, volume 150 of *Computer and Systems Sciences*, pp. 153–168, Berlin. NATO ASI Series, Springler-Verlag.

P. Cohen (1992). The role of natural language in a multimodal interface. In: *Proceedings of the fith annual ACM symposium on user interface software and technology (UIST'92)*, pp. 143–149, Monterey, CA. New-York: ACM Press.

P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen and J. Clow (1997). QuickSet: Multimodal interaction for distributed applications. In: *Fifth ACM International Multimedia Conference MULTIMEDIA'97*, pp. 31–40, New York/Reading. ACM Press/Addison-Wesley.

R. Cole, T. Carmell, P. Connors, M. Macon, J. Wouters, J. de Villiers, A. Tarachow, D. Massaro, M. Cohen, J. Beskow, J. Yang, U. Meier, A. Waibel, P. Stone, G. Fortier, A. Davis and C. Soland (1998). Intelligent animated agents for interactive language training. Unpublished manuscript.

R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen and V. Zue, eds. (1995). *Survey of the State of the Art in Human Language Technology.* Center for Spoken Language Understanding CSLU.

M. Coleman (1969). Text editing on a graphic display device using hand-drawn proofreader's symbols. In: M. Fairman and J. Nievergelt, eds., *Second University of Illinois Conference on Computer Graphics*, Pertinent Concepts in Computer Graphics, pp. 283–290, Urbana Champaign (IL). University of Illinois Press.

S. Condon and C. Cech (1995). *Manual for coding decision-making interactions.* Université des Acadiens.

W. Condon (1988). An analysis of behavioral organization. *Sign Language Studies* 58: 55–88.

S. Coquillard (1990). Extented free form deformation: A sculpturing tool for 3D geometric modeling. *Proceedings of Siggraph '90, Computer Graphics* 22(2): 260–263.

S. Coquillard and P. Jancene (1991). Animated free form deformation: An interactive animation technique. *Proceedings of Siggraph '91, Computer Graphics* 25(4): 23–26.

R. Cornett, R. Beadles and B. Wilson (1977). Automatic cued speech. In: *Proc. Res. Conf. on Speech-Processing Aids for the Deaf*, Gallaudet College.

P. Cosi and E. Magno-Caldognetto (1996). Lips and jaws movements for vowels and consonants: Spatio-temporal characteristics and bimodal recognition applications. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO ASI Series. Series F: Computer and Systems Sciences.* Springler-Verlag, Berlin.

P. Cosi, E. Magno-Caldognetto, F. Ferrero, M. Dugatto and K. Vagges (1996). Speaker independent bimodal phonetic recognition experiments. In: *Proceedings of ICSLP'96: The Fourth International Conference on Spoken Language Processing.*

J. Coutaz (1992). Multimedia and multimodal user interfaces: A software engineering perspective. In: *StPetersburg International Workshop on Human Computer Interaction.*

I. Craw (1992). Recognising face features and faces. *IEEE Colloquium on Machine Storage and Recognition of Faces* 7: 1–4.

D. Cruse (1986). *Lexical semantics.* CUP, Cambridge.

D. Crystal (1988). *A dictionary of linguistics and phonetics.* Basil Blackwell in association with André Deutsch, Oxford, 2nd edition.

N. Dahlbäck, A. Jëonsson and L. Ahrenberg (1992). Wizard of Oz-studies - why and how. In: W. D. Gray, W. E. Hefley and D. Murray, eds., *Proceedings of the International Workshop on Intelligent User Interfaces*, pp. 193–200, New York, NY,

USA. ACM Press.

B. Dalton, R. Kaucic and A. Blake (1996). Automatic speechreading using dynamic contours. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO ASI Series. Series F: Computer and Systems Sciences*, pp. 373–382. Springler-Verlag, Berlin.

J. Davis and J. Hirschberg (1988). Assigning intonational features in synthesized spoken discourse. In: *ACL88*, pp. 187–193, Buffalo.

J. Davis and M. Shah (1993). Gesture recognition. Technical Report CS-TR-93-11, University of Central Florida.

G. de Haan, G. van der Veer and J. van Vliet (1991). Formal modelling techniques in human-computer interaction. *Acta Psychologica* 78: 27–67.

D. DeCarlo (1998). Personal communication.

D. DeCarlo and D. Metaxas (1996). The integration of optical flow and deformable models with applications to human face shape and motion estimation. In: *Proceedings CVPR'96*.

B. deGraf (1990). "Performance" facial animation. In: *Vol 26: State of the Art in Facial Animation*, pp. 10–14. ACM Siggraph'90 Course Notes.

E. den Os and G. Bloothoofd (1998). Evaluating various spoken dialogue systems with a single questionnaire: Analysis of the elsnet olympics. In: *Linguistic Evaluation and Resources Conference*.

A. Denda, T. Itoh and S. Nakagawa (1997). Evaluation of spoken dialogue system for a sightseeing guidance with multi-modal interface. In: *IJCAI'97 workshop on Intelligent Multimodal Systems*, Nagoya, Japan.

X. Deng (1988). *A finite element analysis of surgery of the human facial tissue*. Ph.D. thesis, Columbia University, New-York.

M. D'Imperio (1997). Narrow focus and focal accent in the Neapolitan variety of Italian. In: *Proc. ESCA Workshop: Intonation: Theory, Models and Applications*, pp. 87–90, Athens, Greece.

A. Dix, J. Finlay, G. Abowd and R. Beale (1998). *Human-computer interaction*. Prentice-Hall.

P. Doenges, F. Lavagetto, J. Ostermann, I. Pandzic and E. Petajan (1997). MPEG-4: Audio/video and synthetic graphics/audio for mixed media. *Image Communications Journal* 5(4).

J. Dowell, Y. Shmueli and I. Salter (1995). Applying a cognitive model of the user to the design of a multimodal speech interface. In: *Pre-Proceedings of the First International Workshop on Intelligence and Multimodality in Multimedia Interfaces (IMMI-1)*, Edinburgh, Scotland.

C. Downton and H. Drouet (1991). Image analysis for model-based sign language coding. In: *6th International Conference on Image Analysis and Processing*, pp. 637–644.

P. Duchnowski, U. Meier and A. Waibel (1994). See me, hear me: Integrating automatic speech recognition and lipreading. In: *International Conference on Spoken Language Systems and Processing*, Yokohama (Japan). IEEE.

C. Dugast (1998). Personal communication.

D. Duke and I. Herman (1998). A standard for multimedia middleware. In: *ACM Multimedia 98 - Electronic Proceedings*, "`http://www.acm.org/sigmm/MM98/electronic_proceedings/duke/index.html`".

S. Duncan (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society* 3: 161–180.

K. D. Dutz, ed. (1985). *Studien zur Klassifikation, Terminologie und Systematik:*

Theorie und Praxis. Akten der 6. Arbeitstagung des Münsterander Arbeitskreises für Semiotik. MAKS-Publikationen, Münster, Institut für allgemeine Sprachwissenschaft. Studium der Sprachwissenschaft Beiheft 5, Arbeiten zur Klassifikation; 5.

J. Edwards and M. Lampert, eds. (1993). *Talking data: Transcription and coding in discourse research*. Erlbaum, Hillsdale, New Jersey.

K. Ehlich and J. Rehbein (1975). Zur Konstitution pragmatischer Einheiten in einer Institution: Das Speiserestaurant. In: D. Wunderlich, ed., *Linguistische Pragmatik*, pp. 209–254. Athenäum, Frankfurt am Main.

P. Ekman (1979). About brows: Emotional and conversational signals. In: M. von Cranach, K. Foppa, W. Lepenies and D. Ploog, eds., *Human ethology: Claims and limits of a new discipline: contributions to the Colloquium*, pp. 169–248. Cambridge University Press, Cambridge, England; New-York.

P. Ekman (1989). The argument and evidence about universals in facial expressions of emotion. In: H. Wagner and A. Manstead, eds., *Handbook of Social Psychophysiology*, pp. 143–164. Wiley, Chichester; New-York.

P. Ekman (1992). Facial expressions of emotion: New findings, new questions. *American Psychological Society* 3(1): 34–38.

P. Ekman and W. Friesen (1978). *Facial Action Coding System*. Consulting Psychologists Press, Inc.

A. Emmett (1985). Digital portfolio: Tony de Peltrie. *Computer Graphics World* 8(10): 72–77.

N. Erber and C. deFilippo (1978). Voice/mouse synthesis and tactual/visual perception of /pa, ba, ma/. *Journal of the Acoustical Society of America* 64: 1015–1019.

I. Essa, S. Basu, T. Darrell and A. Pentland (1996). Modeling, tracking and interactive animation of faces and heads using input from video. In: *Computer Animation'96*, pp. 68–79, Geneva, Switzerland. IEEE Computer Society Press.

I. Essa and A. Pentland (1994). A vision system for observing and extracting facial action parameters. *Proceedings of Computer Vision and Pattern Recognition (CVPR 94)* pp. 76–83.

I. A. Essa (1995). *Analysis, Interpretation, and Synthesis of Facial Expressions*. Ph.D. thesis, MIT, Media Laboratory, Cambridge, MA.

I. A. Essa, T. Darrell and A. Pentland (1994). Tracking facial motion. In: *Proceedings of IEEE Workshop on Nonrigid and Articulated Motion*.

J. P. et al. (1994). *Human-computer interaction*. Addison Wesley.

EURODICAUTOM (1998). "`http://www2.echo.lu/eurod/support/-presentationfr.html, http://www2.echo.lu/edic/`".

E. Eyes (1996). *The BNC Treebank: Syntactic annotation of a corpus of modern British English*. M.A. dissertation, Department of Linguistics and Modern English Language,Lancaster University.

T. Ezzat and T. Poggio (1997). Videorealistic talking faces: A morphing approach. In: *AVSP'97 workshop*, Rhodos, Greece.

G. Fang (1992). Vocal tract area functions of swedish vowels and a new three-parameter model. In: *Proceedings of the International Conference on Spoken Language Processing*, volume 1, pp. 807–810, Banff, Canada.

P. Faraday and A. Sutcliffe (1996). An empirical study of attending and comprehending mm presentations. In: *Proceedings of the ACM conference on Multimedia*, Boston, USA.

P. Faraday and A. Sutcliffe (1997). Multimedia: Design for the "moment". In: *Electronic proceedings of the ACM conference on Multimedia*, Seattle, USA.

R. Farag (1979). Word level recognition of cursive script. *IEEE Transactions on Computers* 28: 172–175.

C. Faure and L. Julia (1993). Interaction homme-machine par la parole et le geste pour l'édition de documents: TAPAGE. In: *International Conference on Interfaces to Real and Virtual Worlds*, pp. 171–180.

H. Felber and G. Budin (1989). *Terminologie in Theorie und Praxis.* Forum für Fachsprachenforschung. Günter Narr Verlag, Tübingen.

K. Fischer (1996). Distributed representation formalisms for discourse particles. In: D. Gibbon, ed., *Natural language processing and speech technology.* Mouton de Gruyter, Berlin.

K. Fischer (1998). *A cognitive lexical pragmatic approach to the polysemy of discourse particles.* Ph.D. thesis, University of Bielefeld.

K. Fischer and H. Brandt-Pook (1998). Automatic disambiguation of discourse particles. In: *Proceedings of the Workshop on Discourse Relations and Discourse Markers*, Montreal. COLING-ACL.

G. Flammia and V. Zue (1995). Empirical evaluation of human performance and agreement in parsing discourse constituents in spoken dialogue. In: *Eurospeech 95, 4th European conference on speech communication and technology*, volume 3, pp. 1965–1968, Madrid, Spain.

J. Flanagan (1997). Synergetic modalities for human/machine communication. In: S. Furui, B. Jang and W. Chou, eds., *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 1–8, Santa Barbara (CA). IEEE Signal Processing Society.

J. Foley, A. van Dam, S. Feiner and J. Hughes (1990). *Computer Graphics: Principles and Practice.* Addison-Wesley, Reading, MA. Second Edition.

D. Forsey and R. Bartels (1990). Hierarchical Bspline refinement. *Computer Graphics* 22(4): 205–212.

C. Frankish, R. Hull and P. Morgan (1995). Recognition accuracy and user acceptance of pen interfaces. In: *International Conference on Computer-Human Interaction*, pp. 503–510, Denver (CO). ACM.

A. Fridlund (1994). *Human facial expression: An evolutionary view.* Academic Press, New York.

V. Fromkin (1964). Lip positions in American-English vowels. *Language and Speech* 7(3): 215–225.

S. Frota (1995). On the prosody of intonation of focus in european portuguese. University of Lisbon.

K. Fu (1981). *Syntactic Pattern Recognition and Applications.* Prentice Hall.

Y. Fung (1993). *Biomechanics: Mechanical Properties of Living Tissues.* Springer-Verlag, 2nd edition.

J. Galliers and K. Sparck Jones (1996). *Evaluating natural language processing systems: An analysis and review.* Springer, Berlin.

R. Garside, G. Leech and T. McEnery, eds. (1997). *Corpus annotation: Linguistic information from computer text corpora.* Longman, London.

J. Gauvain, S. Bennacef, L. Devillers, L. Lamel and S. Rosset (1997). Spoken language component of the mask kiosk. In: K. Varghese and S. Pfleger, eds., *Human Comfort and Security of Information Systems*, pp. 93–103. Springer.

D. Gibbon (1999, forthcoming). Computational lexicography. In: F. Van Eynde and D. Gibbon, eds., *Lexicon Development for Speech and Language Processing.* Kluwer, Dordrecht.

D. Gibbon, R. Moore and R. Winski, eds. (1997). *Handbook of standards and resources*

for spoken language systems. Mouton de Gruyter, Berlin.

J. Gibson (1966). *The Senses Considered as Perceptual Systems.* Houghton Mifflin Co., Boston.

J. Gibson (1979). *The Ecological Approach to Visual Perception.* Houghton Mifflin Co., Boston.

A. Goldschen, O. Garcia and E. Petajan (1996). Rationale for phoneme-viseme mapping and feature selection in visual speech recognition. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO ASI Series. Series F: Computer and Systems Sciences*, pp. 505–515. Springler-Verlag, Berlin.

A. J. Goldschen (1993). *Continuous Automatic Speech Recognition by Lipreading.* Ph.D. thesis, George Washington University.

C. Goodwin (1986). Gestures as a resource for the organization of mutual orientation. *Semiotica* 62(1/2): 29–49.

S. Goose, A. Lewis and H. Davis (1997). OHRA: Towards an open hypermedia reference architecture and a migration path for existing systems. "`http://www.mmrg.ecs.soton.ac.uk/publications/papers/jodi97/-ohra.htm#Osterbye96`".

P.-F. Gou (1970). Strain energy function for biological tissues. *Journal of Biomechanics* 3: 547–550.

V. Govindaraju, K. Gyeonghwan and S. Srihari (1997). Paradigms in handwriting recognition. In: *IEEE International Conference on Systems, Man , and Cybernetics*, volume 2, pp. 1498–1503, Orlando (FL). IEEE.

H. Graf, E. Cosatto and M. Potamianos (1997). Robust recognition of faces and facial features with a multimodal system. In: C. Malmborg, ed., *IEEE International Conference on Systems, Man, and Cybernetics*, volume 4, pp. 2034–2039, Orlando (FL). IEEE Computer Society.

E. Grant (1968). An ethological description of non-verbal behaviour during interviews. *British Journal of Medecine and Psychology* 41.

E. Grant (1969). Human facial expression. *Man* 4: 525–536.

H. Gray (1973). *Anatomy of the Human Body.* Lea Febiger, Philadelphia.

S. Greenbaum, ed. (1996). *English worldwide: The International Corpus of English.* Clarendon Press, Oxford.

S. Greenbaum and Y. Ni (1996). About the ICE tagset. In: S. Greenbaum, ed., *English worldwide: The International Corpus of English*, pp. 92–109. Clarendon Press, Oxford.

H. Grice (1969). Utterer's meaning and intentions. *Philosophical Review* 68(2): 147–177.

M. Grice (1995). *The intonation of interrogation of Palermo Italian: Implications for intonation theory.* Niemeyer, Tübingen.

M. Grice, M. Reyelt, R. Benzmüller, J. Mayer and A. Batliner (1996). Consistency in transcription and labelling of German intonation with GToBI. In: *Conference on Spoken Language Processing, Philadelphia*, pp. 1716–1719.

K. Grœnbæk and U. Wiil (1997). Towards a reference architecture for open hypermedia. In: *Proceedings of the 3rd Workshop on Open Hypermedia Systems. Hypertext '97*, pp. 6–11, Southampton, England.

B. Guenter, C. Grimm, H. Malvar and D. Wood (1998). Making faces. *Computer Graphics Annual Conference Series* .

T. Guiard-Marigny (1996). *Modélisation tridimensionnelle des articulateurs de la parole pour l'animation faciale: Implémentation temps réel et mesure d'intelligibilité*

bimodale. Ph.D. thesis, Institut National Polytechnique, Grenoble, France.

T. Guiard-Marigny, A. Adjoudani and C. Benoît (1994). A 3-D model of the lips for visual speech synthesis. In: *Proc. of the 2nd ESCA/IEEE workshop on Speech Synthesis*, pp. 49–52, New Paltz, NY.

T. Guiard-Marigny, N. Tsingos, A. Adjoudani, C. Benoît and M.-P. Gascuel (1996). 3D models of the lips for realistic speech animation. In: *Computer Animation'96*, pp. 80–89, Geneva, Switzerland. IEEE Computer Society Press.

C. Gussenhoven (1984). *On the grammar and semantics of sentence accents*. Foris, Dordrecht.

C. Gussenhoven (1993). The Dutch foot and the chanted call. *Journal of Linguistics* 21: 37–63.

C. Gussenhoven and T. Rietveld (1991). An experimental evaluation of two nuclear-tone taxonomies. *Linguistics* 29: 423–449.

I. Guyon, D. Henderson, P. Albrecht, Y. LeCun and J. Denker (1992). Writer independent and writer adaptive neural network for on-line character recognition. In: S. Impedovo and J. Simon, eds., *From Pixels to Feature III: Frontiers in Handwriting Recognition*. Elsevier Science Publishers.

U. Hadar, T. Steiner, E. Grant and F. C. Rose (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science* 2: 35–46.

F. G. Halasz and M. Schwartz (1994). The dexter hypertext reference model. *Communications of the ACM* 37(2): 30–39.

D. Hand (1982). *Kernel Discriminant Analysis*. Research Studies Press (A Division of John Wiley and Sons), New York (NY).

K. Hapeshi and D. Jones (1992). Interactive multimedia for instruction: A cognitive analysis of the role of audition and vision. *International Journal of Human-Computer Interaction* 4(1): 79–99.

M. Hare, A. Doubleday, M. Ryan and I. Bennet (1995). Intelligent presentation of information retrieved from heterogeneous multimedia databases. In: *Pre-Proceedings of the First International Workshop on Intelligence and Multimodality in Multimedia Interfaces (IMMI-1)*, Edinburgh, Scotland.

L. Harmon, M. Khan, R. Lasch and P. Ramig (1981). Machine identification of human faces. *Pattern Recognition* 13(2): 97–110.

R. Harper, A. Wiens and J. Matarazzo (1978). *Nonverbal Communication: The State of the Art*. J. Wiley and Sons, New York.

H. Hart (1978). *Hart's rules for compositors and readers at the University Press Oxford*. Oxford University Press, Oxford, 38th revised edition.

H. Hartson and P. Gray (1992). Temporal aspects of tasks in the user action notation. *Human-Computer Interaction* 7: 1–45.

K. Hartung, S. Münch, L. Schomaker, T. Guiard-Marigny, B. Le Goff, R. MacLaverty, J. Nijtmans, A. Camurri, I. Defée and C. Benoît (1996). Development of a system architecture for the acquisition, integration, and representation of multimodal information. Esprit project miama report, European Project.

A. Hauptmann (1989). Speech and gestures for graphic image manipulation. In: *International Conference on Computer-Human Interaction*, volume 1, pp. 241–245, Austin (TX). ACM.

A. Hauptmann and M. Witbrock (1997). Informedia: News-on-demand multimedia information acquisition and retrieval. In: M. Maybury, ed., *Intelligent Multimedia Information Retrieval*, pp. 215–239. AAAI Press.

B. Hayes and A. Lahiri (1991). Bengali intonational phonology. *Natural Language and Linguistic Theory* 9: 47–96.

W. Hayes (1993). *Statistics for the Social Sciences.* Holt, Rinehard, and Winston, New-York.

R. S. Heller (1990). The role of hypermedia in education: A look at the research issues. *Journal of Research on Computing in Education* pp. 431–441.

G. W. Helmut Felber, Friedrich Lang, ed. (1979). *Terminologie als angewandte Sprachwissenschaft. Gedenkschrift für Univ. Prof. Dr. Eugen Wüster*, München, London, Paris, New York. Saur.

C. Henton and P. Litwinowicz (1994). Saying and seeing it with feeling: Techniques for synthesizing visible, emotional speech. In: *Proc. of the 2nd ESCA/IEEE workshop on Speech Synthesis*, pp. 73–76, New Paltz, NY.

T. Hildebrandt and W. Liu (1993). Optical recognition of handwritten chinese characters: Advances since 1980. *Pattern Recognition* 26(2): 205–225.

D. Hill, A. Pearce and B. Wyvill (1988). Animating speech: An automated approach using speech synthesised by rules. *The Visual Computer* 3: 277–289.

W. Hill et al. (1992). *Architectural Qualities and Principles for Multimodal and Multimedia Interfaces.* ACM Press.

J. Hirschberg (1990). Accent and discourse context: Assigning pitch accent in synthetic speech. In: *AAAI90*, pp. 952–957.

D. Hirst (1991). Intonation models: Towards a third generation. In: *Actes du XIIème Congrès International des Sciences Phonétiques. 19-24 août 1991, Aix-en-Provence, France*, pp. 305–310, Aix-en-Provence. Université de Provence, Service des Publications.

D. Hirst and R. Espesser (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix* 15: 71–85.

D. Hirst, N. Ide and J. Véronis (1994). Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project. In: *Proceedings of the ESCA/IEEE Workshop on Speech Synthesis, New York, September 1994.*

D. Hirst, P. Nicolas and R. Espesser (1991). Coding the F0 of a continuous text in French: An experimental approach. In: *Actes du XIIème Congrès International des Sciences Phonétiques. 19–24 août 1991, Aix-en-Provence, France*, volume 5, pp. 234–237. Aix-en-Provence: Université de Provence, Service des Publications.

E. Holden and G. Roy (1992). The graphical translation of English text into signed English in the hand sign translator system. In: A. Kilgour and L. Kjelldahl, eds., *Eurographics '92.* Blackwell Publishers.

J. Hollan, E. Rich, W. Hill, D. Wroblewski, W. Wilner, K. Wittenberg and J. Grudin (1988). An introduction to HITS: Human Interface Tool Suite. Technical Report ACA-HI-406-88, Microelectronics and Computer Technology Corporation.

Howard Hughes medical institute (1995). Seeing, hearing and smelling the world. "`http://www.hhmi.org/senses/`". report.

C. Huls and E. Bos (1995). Studies into full integration of language and action. In: *Proceedings of the International Conference on Cooperative Multimodal Communication(CMC/95)*, pp. 161–174, Eindhoven.

C. Huls, E. Bos and A. Dijkstra (1994). Talking pictures. In: P. McKevitt, ed., *Twelth National Conference on Artificial Intelligence (AAAI94),*, pp. 83–90, Seattle, Washington, USA. Working notes of the workshop "Integration of Natural Language and Vision Processing".

K. Hunt (1965). Grammatical structures written at three grade levels. Technical Report, N.C.T.E, Champaign, Ill.

M. Hunt (n.d.). Private Communication.

D. Hymes (1972/1986). Models of the interaction of language and social life. In:

J. Gumperz and D. Hymes, eds., *Directions in sociolinguistics: The ethnography of communication*, pp. 35–71. Blackwell, Oxford. (Originally published by Holt, Rinehart and Winston, 1972).

IBM (n.d.). "`http://www.software.ibm.com/is/voicetype/dev_vvsdk.html`".

A. Ichikawa, Y. Okada, A. Imiya and Y. Horiuchi (1997). Analytical method for linguistic information of facial gestures in natural dialogue languages. In: *AVSP'97 workshop*, Rhodos, Greece.

N. Ide, G. Priest-Dorman and J. Véronis (1996). EAGLES recommendations on corpus encoding. EAGLES Document EAG–TCWG–CES/R–F. Version 1.4, October, 1996.

T. Ishii, T. Yasuda and J. Toriwaki (1993). A generation model of the human skin texture. In: N. Magnenat-Thalmann and D. Thalmann, eds., *Computer Graphics International '93: Communicating with real world*, pp. 139–150, Tokyo. Springer-Verlag.

R. Jacob (1993). Eye-movement based human-computer interaction techniques: Toward non-command interfaces. In: H. Hartson and D. Hix, eds., *Advances in Human-Computer Interaction*, volume 4, pp. 151–189. Ablex Publishing.

Javasoft (n.d.). "`http://www.javasoft.com/products/java-media/speech/-index.html`".

S. Jekat, H. Tappe, H. Gerlach and T. Schöllhammer (1997). Dialogue interpreting: Data and analysis. VM-Report 189, University of Hamburg.

S. Johansson (1995). The approach of the Text Encoding Initiative to the encoding of spoken discourse. In: G. Leech, G. Myers and J. Thomas, eds., *Spoken English on computer: Transcription, mark-up and application*, pp. 82–98. Longman, London and New York.

S. Johansson et al. (1991). Text Encoding Initiative, Spoken Text Work Group: Working paper on spoken texts. October 1991, Manuscript.

B. E. John and D. E. Kieras (1994). The GOMS family of analysis techniques: Tools for design and evaluation. Technical Report CMU-CS-94-181, Carnegie Mellon University School of Computer Science.

M. Johnston, P. Cohen, D. McGee, S. Oviatt, J. Pittman and I. Smith (1997). Unification-based multimodal integration. In: *35th Annual Meeting of the Association for Computational Linguistics*, pp. 281–288, Madrid (Spain).

G. Jones, J. Foote, K. Jones and S. Young (1997). The video mail project: experiences in retrieving spoken documents. In: M. Maybury, ed., *Intelligent Multimedia Information Retrieval*, pp. 191–214. AAAI Press.

P. Kalra (1993). *An Interactive Multimodal Facial Animation System.* Ph.D. thesis, Swiss Federal Institute of Technology, Lausanne, Switzerland.

P. Kalra, A. Mangili, N. Magnenat-Thalmann and D. Thalmann (1991). SMILE: A multilayered facial animation system. In: T. Kunii, ed., *Modeling in Computer Graphics.* Springer-Verlag.

P. Kalra, A. Mangili, N. Magnenat-Thalmann and D. Thalmann (1992). Simulation of muscle actions using rational free form deformations. *Proc Eurographics '92, Computer Graphics Forum* 2(3): 59–69.

H. Kamio, M. Koorita, H. Matsuura, M. Tamura and T. Nitta (1994). A UI design support tool for multimodal spoken dialogue system. In: *International Conference on Spoken Language Processing*, volume 3, pp. 1283–1286, Philadelphia (PA). IEEE Computer Society.

T. Kanade (1973). *Picture processing by computer complex and recognition of human face.* Ph.D. thesis, Kyoto University, Dept. of Information Science.

F. Karlsson, A. Voutilainen, J. Heikkilä and A. Anttila, eds. (1995). *Constraint Grammar, a language-independent system for parsing unconstrained text.* Mouton de Gruyter, Berlin and New York.

R. Kausic, B. Dalton and A. Blake (1996). Real-time lip tracking for audio-visual speech recognition applications. In: *Proc. European Conf. Computer Vision*, pp. 376–387, Cambridge, UK.

D. Keltner (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology* 68(3): 441–454.

A. Kendon (1974). Movement coordination in social interaction: Some examples described. In: S. Weitz, ed., *Nonverbal Communication.* Oxford University Press.

A. Kendon (1990). *Conducting interaction: Pattern of behavior in focused encounter.* Cambridge University Press.

R. Kent and F. Minifie (1977). Coarticulation in recent speech production models. *Journal of Phonetics* 5: 115–135.

M. Kirby and L. Sirovich (1990). Application of Karhunen-Loeve procedure for the characterisation of human faces. *PAMI* 12(10): 103–108.

J. Kleiser (1989). A fast, efficient, accurate way to represent the human face. In: *Vol 22: State of the Art in Facial Animation*, pp. 20–33. ACM Siggraph'89 Course Notes.

Kleiser-Walczak (1988). Sextone for president. *ACM SIGGRAPH '88 Film and Video Show* issue 38/39. Kleiser Walczak Construction Comp.

Kleiser-Walczak (1989). Dozo. *ACM SIGGRAPH '89 Film and Video Show* Kleiser Walczak Construction Comp.

G. Knowles (1987). *Patterns of spoken English.* Longman, London.

G. Knowles (1991). Prosodic labelling: The problem of tone group boundaries. In: S. Johansson and A.-B. Stenström, eds., *English computer corpora: Selected papers and research guide*, pp. 149–163. Mouton de Gruyter.

G. Knowles, A. Wichmann and P. Alderson (1996). *Working with speech: Perspectives on research into the Lancaster/IBM Spoken English Corpus.* Longman, London and New York.

T. Koda and P. Maes (1996). Agents with faces: The effects of personification of agents. In: *HCI'96.*

K. Kohler (1987). Categorical pitch perception. In: *Proc. IX ICphS*, volume 5, pp. 331–333, Tallin.

K. Kohler, ed. (1991). *Studies in German intonation.* Arbeitsberichte 25, Universität Kiel.

K. Kohler (1995). PROLAB – the Kiel system of prosodic labelling. In: *Proc. ICPhS 95*, pp. 162–165, Stockholm.

K. Kohler (1996). Parametric control of prosodic variables by symbolic input in TTS synthesis. In: van Santen et al., ed., *Progress in Speech Synthesis*, pp. 459–475. Springer, New York.

D. Koons, C. Sparrell and K. Thorisson (1993). Integrating simultaneous input from speech, gaze, and hand gestures. In: M. Maybury, ed., *Intelligent Multimedia Interfaces*, pp. 257–275. Morgan Kaufmann.

G. Kramer (1994). An introduction to auditory display. In: G. Kramer, ed., *Auditory Display*, pp. 1–77. Addison-Wesley.

J. Kuch and T. Huang (1995). Vision based hand modeling and tracking. In: *International Conference on Computer Vision*, Cambridge (MA).

D. Ladd (1996). *Intonational Phonology.* CUP, Cambridge.

D. Ladd, K. Scherer and K. Silverman (1985). An integrated approach to studying intonation and attitude. In: C. Johns-Lewis, ed., *Intonation in discourse*. C. Johns-Lewis.

M. Lallouache (1991). *Un poste visage-parole couleur. Acquisition et traitement automatique des lèvres*. Ph.D. thesis, Institut National Polytechnique, Grenoble, France.

J. Landay and B. Myers (1993). Extending an existing user interface toolkit to support gesture recognition. In: *INTERCHI '93*, Amsterdam (Netherlands). ACM Press.

W. Larrabee (1986). A finite element model of skin deformation. I. Biomechanics of skin and soft tissue: A review. *Laryngoscope* 96: 399–405.

J. Lassiter (1987). Principles of traditional animation applied to 3D computer animation. *SIGGRAPH'87, Computer Graphics* 21(4): 35–44.

F. Lavagetto and P. Lavagetto (1996). Time delay neural networks for articulatory estimation from speech: Suitable subjective evaluation protocols. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO ASI Series. Series F: Computer and Systems Sciences*. Springler-Verlag, Berlin.

J. Lee and T. Kunii (1995). Model-based analysis of hand posture. *IEEE Computer Graphics and Applications* pp. 77–86.

Y. Lee, D. Terzopoulos and K. Waters (1995). Realistic modeling for facial animation. *Computer Graphics Annual Conference Series* pp. 55–62.

G. Leech, R. Barnett and P. Kahrel (1996). Guidelines for the standardization of syntactic annotation of corpora. EAGLES Document EAG-TCWG-SASG/1.8.

G. Leech and R. Garside (1991). Running a grammar factory: The production of syntactically analysed corpora or 'treebanks'. In: S. Johansson and A.-B. Stenström, eds., *English computer corpora: Selected readings and research guide*, pp. 15–32. Mouton de Gruyter, Berlin and New York.

G. Leech, G. Myers and J. Thomas, eds. (1995). *Spoken English on computer: Transcription, mark-up and application*. Longman, London and New York.

G. Leech and A. Wilson (1994). EAGLES morphosyntactic annotation. EAGLES Report EAGCSG/IR-T3. 1. Pisa, Istituto di Linguistica Computazionale. Reissued (Version of Mar. 1996) as: *Recommendations for the morphosyntactic annotation of corpora*. EAGLES Document EAG-TCWG-MAC/R.

B. LeGoff (1997). *Synthèse à partir du texte de visage 3D parlant français*. Ph.D. thesis, Institut National Polytechnique, Grenoble, France.

B. LeGoff and C. Benoît (1997). A French speaking synthetic head. In: *AVSP'97 workshop*, Rhodos, Greece.

B. LeGoff, T. Guiard-Marigny and C. Benoît (1996). Analysis-synthesis and intelligibility of a talking face. In: J. van Santen, R. Sproat, J. Olive and J. Hirschberg, eds., *Progress in Speech Synthesis*. Springler-Verlag.

C. Lehmann (1996). Linguistische Terminologie als relationales Netz. In: C. Knobloch and B. Schaeder, eds., *Nomination – fachsprachlich und gemeinsprachlich*, pp. 215–267. Westdeutscher Verlag, Opladen. AVG, Allgemein-Vergleichende Grammatik, Arbeitspapier Nr. 8, DFG-Projekt, Universität Bielefeld.

J. Leopold and A. Ambler (1997). Keyboardless visual programming using voice, handwriting, and gesture. In: *IEEE Symposium on Visual Languages*, pp. 28–35, Capri (Italy). IEEE Computer Society.

E. Levin and R. Pieraccini (1995). CHRONUS: The next generation. In: *ARPA Workshop on Spoken Language Technology*, Austin (TX). Morgan Kaufman.

J. Levine, T. Mason and D. Brown (1995). *lex & yacc*. O'Reilly & Associates, Se-

bastopol.

S. Levinson (1979). Activity types and language. *Linguistics* 17.5/6: 356–399.

A. Levy-Schoen (1969). *L'etude des mouvements oculaires.* Dunod, Paris.

J. Lewis and F. Parke (1987). Automated lipsynch and speech synthesis for character animation. In: J. Caroll and P. Tanner, eds., *Proceedings Human Factors in Computing Systems and Graphics Interface '87*, pp. 143–147.

S. Liddell (1980). *American Sign Syntax Language.* The Hague.

P. Limantour (1994). *FACIES: Facial Animation controlled by an interactive and efficient system.* Ph.D. thesis, Université de Paris XI-Orsay, Orsay, France.

J. Lipscomb (1991). A trainable gesture recognizer. *Pattern Recognition* .

P. Litwinowicz (1994). Animating images with drawings. *Computer Graghics Annual Conferences Series* pp. 413–420.

J. Llisterri (1996). EAGLES preliminary recommendations on spoken texts. EAGLES document EAG–TCWG–SPT/P.

A. Lofqvist (1990). Speech as audible gestures. *Speech Production and Speech Modeling* pp. 289–322.

Longman (1992). *Dictionary of English Language and Culture.* Longman.

J. Loomis, H. Poizner, U. Bellugi, A. Blakemore and J. Hollerbach (1983). Computer graphic modeling of American Sign Language. *Computer Graphics* 17(3): 105–114.

J. Luettin, N. Thacker and S. Beet (1996). Active shape models for visual speech feature extraction. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO ASI Series. Series F: Computer and Systems Sciences.* Springler-Verlag, Berlin.

J. Lyons (1977). *Semantics*, volume I and II. Cambridge University Press, Cambridge.

M2VTS (1996). www.uk.infowin.org/acts/rus/projects/ac102.htm.

B. MacWhinney (1995). *The CHILDES project: Tools for analyzing talk.* Lawrence Erlbaum, Hillsdale, NJ.

S. Madhvanath (1996). *The Holistic Paradigm in Handwritten Word Recognition and its Application to Large and Dynamic Lexicon Scenarios.* Ph.d., State University of New York.

P. Maes, T. Darrell, B. Blumberg and A. Pentland (1995). The ALIVE system: Full-body interaction with autonomous agents. In: *Computer Animation'95*, Geneva, Switzerland. IEEE Computer Society Press.

Maggioni (1985). A novel gestural input device for virtual reality. In: *IEEE Annual Virtual Reality International Symposium*, pp. 118–124.

N. Magnenat-Thalmann and D. Thalmann (1987). The direction of synthetic actors in the film: Rendez-vous à Montrèal. *IEEE Computer Graphics and Applications* pp. 9–19.

E. Magno-Caldognetto and I. Poggi (1997). Micro- and macro-bimodality. In: C.Benoît and R.Campbell, eds., *Proceedings of the Workshop on Audio Visual Speech Perception*, Rhodes.

U. Malaske (1998). Sprechen statt schreiben. *c't Magazin für Computer Technik* 5: 110–119.

S. Manke (1998). *On-line Erkennung kursiver Handschrift bei groβen Vokabularien (On-line Recognition of Cursive Handwriting with Large Vocabularies).* Ph.d., Fredericiana.

J. Manschot and A. Brakee (1986). The measurement and modelling of the mechanical properties of human skin in vivo - II. The model. *Journal of Biomechanics* 19(7): 517–521.

F. Marcos-Marín, A. Ballester and C. Santamaría (1993). Transcription conventions

used for the Corpus of Spoken Contemporary Spanish. *Literary and Linguistic Computing* 8(4): 283–292.

H. Marmolin (1991). Multimedia from the perspectives of psychology. In: L. Kjelldahl, ed., *Multimedia Systems, Interaction and Applications. 1st Eurographics Workshop*, Berlin. Springer-Verlag.

J. Martin (1995). *Cooperations between modalities and binding through synchrony in multimodal interfaces.* Ph.D. thesis, ENST 95 E 015, Orsay, France. In French.

J. Martin (1997). Towards "intelligent" cooperation between modalities. In: *IJCAI Workshop on Intelligent Multimodal Systems*, Nagoya (Japan).

J. Martin, R. Veldman and D. Beéroule (1995). Towards adequate representation technologies for multimodal interfaces. In: *International Conference on Cooperative Multimodal Communication*, volume 1, pp. 207–223.

K. Mase (1991). Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan* 22(6): 67–75.

J. Mason, F. Deravi, C. Chibelushi and S. Gandon (1996). Project: DAVID (Digital Audio Visual Integrated Database). Technical Report, Univ. of Swansea, Dept. of Electrical and Electronic Engineering.

D. Massaro and M. Cohen (1990). Perception of synthesized audible and visible speech. *Psychological Science* 1(1): 55–63.

M. Maybury (1993). *Intelligent Multimedia Interfaces.* AAAI Press.

M. Maybury (1997). *Intelligent Multimedia Information Retrieval.* AAAI / The MIT Press.

C. Mayo, M. Aylett and D. Ladd (1997). Prosodic transcription of Glasgow English: An evaluation study of GlaToBI. In: *Proc. ESCA Workshop on Intonation: Theory, Models and Applications. Athens, Greece, September 18–20.*

P. Mc Kevitt, D. Partridge and Y. Wilks (1992). Approaches to natural language discourse processing. Technical Report, Computing Research Laboratory Dept. 3CRL, Box 30001, New Mexico State University, Las Cruces.

D. McAllister, R. Rodman, D. Bitzer and A. Freeman (1997). Lip synchronization of speech. In: *AVSP'97 workshop*, Rhodos, Greece.

S. McGlashan (1996). Towards multimodal dialogue management. In: *Proceedings of Twente Workshop on Language Technology*, volume 11, Enschede, The Netherlands.

H. McGurk and J. MacDonald (1976). Hearing lips and seeing voices. *Nature* 264: 746–748.

D. McNeill (1992). *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago.

U. Meier, R. Stiefelhagen and J. Yang (1997). Preprocessing of visual speech under real world conditions. In: *AVSP'97 workshop*, Rhodos, Greece.

A. K. Melby and S. E. Wright (1998). The CLS framework. Technical Report, TTT.org. Draft, June 1998.

B. Mellor and C. Baber (1997). Modelling of speech-based user interfaces. In: *Proceedings of Eurospeech'97.*

I. Mennen and E. den Os (1993). Intonation of modern Greek sentences. In: *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, volume 17, pp. 111–128.

MHEG ISO (1998). Mheg iso, coding of multimedia and hypermedia information. "`http://www.demon.co.uk/tcasey/ MHEG, http://www.chips.ibm.com/.sc29/29w42912.htm`".

M. Monachini (1995). ELM-IT: An Italian incarnation of the EAGLES-TS. Definition of lexicon specification and guidelines. Technical Report, Istituto di Linguistica

Computazionale, Pisa.

M. Monachini and N. Calzolari (1996). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: A common proposal and applications to european languages. Eagles document eag-clwg-morphsyn/r, Instituto di Linguistica Computazionale, Pisa. Revised edition.

A. Monaghan (1991). *Intonation in a Text-to-Speech Conversion System*. Ph.D. thesis, University of Edinburgh.

D. Moran, A. Cheyer, L. Julia and D. Martin (1997). Multimodal user interfaces in the Open Agent Architecture. In: *Intelligent User Interfaces*, pp. 61–68, Orlando (FL).

T. Morimoto, T. Takezawa, F. Yato, S. Sagayama, T. Tashiro, M. Nagata and A. Kurematsu (1993). ATR's speech translation system: ASURA. In: *European Conference on Speech Communication and Technology*, volume 2, pp. 1295–1298, Berlin (Germany).

S. Morishima (1996). Face-feature extraction from spatial frequency for dynamic expression. In: *Siggraph'96 Tutorial Course No. 25: Life-like, Believable Communication Agents*, pp. 47–55.

S. Morishima and H. Harashima (1991). Speech-to-image media conversion based on VQ and neural network. In: *Proceedings of ICASSP91, M10.11*, pp. 2865–2868.

J. Mostow, S. Roth, A. Hauptmann and M. Kane (1994). A prototype reading coach that listens. In: *Proceedings of Twelfth National Conference on Artificial Intelligence AAAI*, pp. 785–792.

MPEG ISO (1998). Mpeg iso, coded representation of moving pictures and associated audio. "`http://www.chips.ibm.com/.sc29/29w42911.htm`".

B. Myers, R. McDaniel, R. Miller, A. Ferrency, A. Faulring, B. Kyle, A. Mickish, A. Klimovitski and P. Doane (1997). The Amulet environment: New models for effective user interface software development. *IEEE Transactions on Software Engineering* 23(6): 347–365.

M. Nahas, H. Huitric and M. Saintourens (1988). Animation of a B-spline figure. *The Visual Computer* 3(5): 272–276.

C. Nakatani, B. Grosz, D. Ahn and J. Hirschberg (1995). Instructions for annotating discourses. Technical Report, Center for Research in Computing Technology, Harvard University, Cambridge, MA.

C. Nakatani and J. Hirschberg (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America* 95(3): 1603–1616.

C. Nakatani and D. Traum (1998). *Draft: Discourse Structure Coding Manual*. "`http://www.cs.umd.edu/users/traum/DSD/ntman.ps`".

G. Nelson (1996). Markup systems. In: *English worldwide: The International Corpus of English*, pp. 36–53. Clarendon Press, Oxford.

J. Nespoulous and A. Lecours (1986). Gestures: Nature and function. In: J. Nespoulous, Perron and A. Lecours, eds., *The Biological Foundations of Gestures*. Lawrence Erlbaum Associates, Hillsdale (NJ).

W. Newman and R. Sproull (1979). *Principles of Interactive Computer Graphics*. McGraw-Hill.

H. Niemann, E. Nöth, S. Harbeck and V. Warnke (1997a). Topic spotting using subword units. Verbmobil-Report 205, F.-A.-Universität Erlangen-Nürnberg.

H. Niemann, E. Nöth, A. Kießling, R. Kompe and A. Batliner (1997b). Prosodic processing and its use in Verbmobil. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pp. 75–78, München.

L. Nigay and J. Coutaz (1993). A design space for multimodal systems: Concurrent

processing and data fusion. In: *International Conference on Computer-Human Interaction*, pp. 172–178. ACM Press.

L. Nigay and J. Coutaz (1995). A generic platform for addressing the multimodal challenge. In: *International Conference on Computer-Human Interaction*, pp. 98–105, Denver (CO). ACM.

K. Nitta, O. Hasegawa, T. Akiba, T. Kamishima, T. Kurita, S. Hayamizu and K. Itoh (1997). An experimental multimodal disputation system. In: *Proc. of IJCAI'97*, Nagoya, Japan.

F. Nolan and E. Grabe (1997). Can 'ToBI' transcribe intonational variation in British English? In: Botinis, Kouroupetroglou and Carayiannis, eds., *Intonation: Theory, Models and Applications*. Proceedings of the ESCA Workshop, Athens, Greece.

F. Nouboud and R. Plamondon (1990). On-line recognition of handprinted characters: Survey and beta tests. *Pattern Recognition* 23(9): 1031–1044.

R. Ochsman and A. Chapanis (1974). The effects of 10 communication modes on the behavior of teams during cooperative problem solving. *Intern. J. Man-Machine Studies* 6: 579–619.

K. Østerbye and U. Wiil (1996). The flag taxonomy of open hypermedia systems. In: *Proceedings of the ACM Hypertext '96 Conference*, pp. 129–139, Washington, DC.

J. Osterhout (1994). *Tcl and the Tk Toolkit*. Addison Wesley Professional Computing.

S. Oviatt, A. DeAngeli and K. Kuhn (1997). Integration and synchronization of input modes during multimodal human-computer interaction. In: *International Conference on Computer-Human Interaction*, volume 1, pp. 415–422, Atlanta (GA). ACM.

S. Oviatt and R. VanGent (1996). Error resolution during multimodal human-computer interaction. In: *International Conference on Spoken Language Processing*, volume 2, pp. 204–207, Philadelphia (PA).

D. Pallett, J. Fiscus, W. Fisher, J. Garofolo, B. Lund, A. Martin and M. Przybocki (1994). 1994 benchmark tests for the ARPA spoken language program. In: *ARPA Workshop on Spoken Language Technology*, pp. 5–36, Princeton (NJ). Margan Kaufmann Publishers, Inc.

F. Parke (1972). *Computer Generated Animation of Faces*. Master's thesis, University of Utah, Salt Lake City, UT. UTEC-CSc-72-120.

F. Parke (1991). Control parametrization for facial animation. In: N. Magnenat-Thalmann and D. Thalmann, eds., *Computer Animation '91*, pp. 3–14. Springer-Verlag.

F. Parke and K. Waters (1996). *Computer Facial Animation*. A.K. Peters, Wellesley, MA.

M. Patel and P. Willis (1991). FACES–The Facial Animation, Construction and Editing System. In: *Proceedings of Eurographics'91 Conference*, pp. 33–45, Austria.

E. Patterson, P. Litwinowicz and N. Greene (1991). Facial animation by spatial mapping. In: N. Magnenat-Thalmann and D. Thalmann, eds., *Computer Animation '91*, pp. 45–58. Springer-Verlag.

A. Pearce, B. Wyvill and D. Hill (1986). Speech and expression: A computer solution to face animation. *Graphics and Vision Interface '86* pp. 136–140.

C. Pelachaud, N. Badler and M. Steedman (1996). Generating facial expressions for speech. *Cognitive Science* 20(1): 1–46.

C. Pelachaud and I. Poggi (1998). Multimodal communication between synthetic agents. In: *Advanced Visual Interface*, Aquila, Italy.

C. Pelachaud and S. Prevost (1995). Coordinating vocal and visual parameters for 3D virtual agents. In: *2nd Eurographics Workshop on Virtual Environments*, Monte

Carlo.

C. Pelachaud, M. Viaud and H. Yahia (1993). Rule-structured facial animation system. In: *IJCAI 93*, volume 2, pp. 1610–1615.

A. Pentland and K. Mase (1989). Lipreading: Automatic visual recognition of spoken words. In: *Proc. Image Understanding and Machine Vision*. Optical Society of America.

F. Pereira and S. Shieber (1987). Prolog and natural language analysis. CLSI Lecture Notes No. 10.

E. Petajan (1984). Automatic lipreading to enhance speech recognition. In: *Proceedings of the IEEE Communication Society Global Telecommunications Conference*.

E. Petajan (1997). Facial animation coding, unofficial derivative of MPEG-4 standardization, work-in-progress. Technical Report, Human Animation Working Group, VRML Consortium.

E. Petajan and H. P. Graf (1996). Robust face feature analysis for automatic speachreading and character animation. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO ASI Series. Series F: Computer and Systems Sciences*, pp. 425–436. Springler-Verlag, Berlin.

S. Pieper (1991). *CAPS: Computer-Aided Plastic Surgery*. Ph.D. thesis, Massachusetts Institute of Technology, Media Arts and Sciences.

S. Pieper and D. Zeltzer (1989). A biologically inspired model of human facial tissue for computer animation. In: *Vol 22: State of the Art in Facial Animation*, pp. 71–124. ACM Siggraph'89 Course Notes.

J. Pierrehumbert (1980). *The phonology and phonetics of English intonation*. Ph.D. thesis, MIT. Published 1988 by Indiana Univerity Linguistics Club.

J. Pierrehumbert and J. Hirschberg (1990). The meaning of intonational contours in the interpretation of discourse. In: P. Cohen, J. Morgan and M. Pollack, eds., *Intentions in Communication*, pp. 271–312. MIT Press, Cambridge, MA.

S. Platt (1985). *A Structural Model of the Human Face*. Ph.D. thesis, University of Pennsylvania, Dept. of Computer and Information Science, Philadelphia, PA.

S. Platt and N. Badler (1981). Animating facial expressions. *Computer Graphics* 15(3): 245–252.

I. Poggi and E. M. Caldognetto (1996). A score for the analysis of gestures in multimodal communication. In: L.Messing, ed., *Proceedings of the Workshop on the Integration of Gesture and Language in Speech, Applied Science and Engineering Laboratories*, pp. 235–244, Newark and Wilmington, Del.

B. Post (1993). *A phonological analysis of French intonation*. Master's thesis, University of Nijmegen.

G. Potamianos, E. Cosatto, H. Graf and D. Roe (1997). Speaker independent audio-visual database for bimodal ASR. In: *AVSP'97 workshop*, Rhodos, Greece.

PREMO (1998). "`http://dbs.cwi.nl/cwwwi/owa/cwwwi.print_projects?ID=44`".

S. Prevost (1996). Modeling contrast in the generation and synthesis of spoken language. In: *Proceedings of ICSLP'96: The Fourth International Conference on Spoken Language Processing*.

S. Prevost and C. Pelachaud (to appear). *Talking Heads*. MIT Press.

J. A. Provine and L. T. Bruton (1996). 3-D model based coding - An very low bit rate coding scheme for video-conferencing. In: *Proc. IEEE Intl. Symp. on Circuits and Sys.*, pp. 798–801.

M. Rahim, C. Goodyear, W. Kleijn and J. Schroeter (1993). On the use of neural networks in articulatory speech synthesis. *Journal of the Acoustical Society of*

America 93(2): 1109–1121.

A. Rahman and G. Sampson (1998). Extending grammar annotations to spontaneous speech. In: B. Krenn, T. Brants, W. Skut and H. Uszkoreit, eds., *Recent advances in corpus annotation*. ESSLII 98, DFKI, Saarbrücken.

P. Rauss, J. Phillips, M. Hamilton and A. DePersia (1996). FERET (face-recognition technology) recognition algorithms. In: *Proc. of the Fifth Automatic Target Recognizer System and Technology Symposium*.

C. Reed, W. Rabinowitz, N. Durlach, , L. Delhome, L. Braida, J. Pemberton, B. Mulcahey and D. Washington (1992). Analytic study of the Tadoma method: Improving performance with supplementary tactual displays. *Journal of Speech and Hearing Research* 35: 455–465.

C. Reed, W. Rabinowitz, N. Durlach, L. Braida, S. Conway-Fithian and M. Schultz (1985). Research on the Tadoma method of speech communication. *Journal of the Acoustical Society of America* 77(1): 247–257.

B. Reeves (1990). Simple and complex facial animation: Case studies. In: *Vol 26: State of the Art in Facial Animation*, pp. 90–106. ACM Siggraph'90 Course Notes.

J. Rehg and T. Kanade (1993). Digiteyes: Vision-based human hand tracking. Technical Report CMU-CS-93-220, Carnegie Mellon University.

E. Reiter (1997). Choosing a media for presenting information. *Electronic Transactions on Artificial Intelligence* 1(1). Discussion paper.

L. Revéret, F. Garcia, C. Benoît and E. Vatikiotis-Bateson (1997). An hybrid approach to orientation-free liptracking. In: *AVSP'97 workshop*, Rhodos, Greece.

M. Reyelt, M. Grice, R. Benzmüller, J. Mayer and A. Batliner (1986). Prosodische Etikettierung des Deutschen mit ToBI. In: D. Gibbon, ed., *Natural Language and Speech Technology: Results of the third KONVENS conference, Bielefeld*, pp. 144–155. Mouton de Gruyter, Berlin.

J. Rhyne (1987). Dialogue management for gestural interfaces. *Computer Graphics* 21(2): 137–142.

J. Rhyne and C. Wolf (1993). Recognition-based user interfaces. In: H. Hartson and D. Hix, eds., *Advances in Human-Computer Interaction*, volume 4, pp. 191–212. Ablex Publishing, Norwood (NJ).

S. Rickel and Johnson (1998). Task-oriented dialogs with animated agents in virtual reality. In: *WECC'98, The First Workshop on Embodied Conversational Characters*.

E. L. Riegelsberger (1997). *The Acoustic-to-Articulatory Mapping of Voiced and Fricated Speech*. Ph.D. thesis, The Ohio State University.

M. Riley (1989). Some applications of tree-based modelling to speech and language. In: *Proceedings of the Speech and Natural Language Workshop*, Cape Cod MA. DARPA. Morgan Kaufmann.

A. Risberg and J. Lubker (1978). Prosody and speechreading. Technical Report Quaterly Progress and Status Report 4, Speech Transmission Laboratory, KTH, Stockholm, Sweden.

T. Rist, E. André and J. Müller (1997). Adding animated presentation agents to the interface. In: *Intelligent User Interface*, pp. 79–86.

P. Roach (1994). Conversion between prosodic transcription systems: 'Standard British' and ToBI. *Speech Communication* 15: 91–99.

B. Robertson (1988). Mike the talking head. *Computer graphics world* 11(7).

D. Roe, F. Pereira, R. Sproat and M. Riley (1992). Efficient grammar processing for a spoken language translation system. In: *International Conference on Acoustics, Speech and Signal Processing*, volume 1, pp. 213–216.

T. Roks (1997). Autovisie-tno onderzoek naar voice dialling in de auto. *Autovisie* pp. 60–61. December 1997.

R. Rosenthal and R. Rosnow (1991). *Essentials of Behavioral Research: Methods and Data Analysis*. McGraw-Hill series in Psychology. McGraw-Hill.

D. Rubine (1991a). *The automatic recognition of gestures*. Ph.d. thesis, Carnegie Mellon University.

D. Rubine (1991b). Specifying gestures by example. *ACM Journal on Computer Graphics* 25(4): 329–337.

A. Rudnicky and A. Hauptmann (1991). Models for evaluating interaction protocols in speech recognition. In: *Proceedings of the CHI conference*, pp. 285–291.

A. Rudnicky, S. Reed and E. Thayer (1996). SpeechWear: A mobile speech system. In: *Proceedings of ICSLP'96: The Fourth International Conference on Spoken Language Processing*, pp. 538–541.

H. Sacks (1967–1972). Unpublished lecture notes, 1967–72. University of California.

H. Sacks, E. Schegloff and G. Jefferson (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50: 696–735.

J. Sager (1990). *A practical course in terminology processing*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

J. Sager and M.-C. L'Homme (1994). A model for the definition of concepts: Rules for analytical definitions in terminological databases. *Terminology* 1(2): 351–374.

J. C. Sager and B. Nkwenti-Azeh (1989). *Terminological problems involved in the process of exchange of new technoloy between developing and developed countries*. UNESCO, Paris. Terminological problems involved in the process of exchange of new technology between developing and developed countries, booklet, J.C. Sager and Blaise Nkwenti-Azeh at Centre for Computational Linguistics UMIST, prepared for the Division of Economic and Social Sciences.

H. Said and T. Tan (1996). A brief review on integrated audio-visual processing for personal identification. In: *IEEE Colloquium on Integrated Audio-Visual Processing for Recognition*, London (UK).

M. Salisbury, J. Hendrickson, T. Lammers and C. Fu (1990). Talk and draw: Bundling speech and graphics. *Computer* 23(8): 59–65.

E. Salztman and K. Munhall (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology* 1(4): 333–382.

S. Samal and P. Iyengar (1992). Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition* 25(1): 65–77.

G. Sampson (1987). Probabilistic models of analysis. In: R. Garside, G. Leech and G. Sampson, eds., *The computational analysis of English*, pp. 16–29. Longman, London.

G. Sampson (1995). *English for the computer*. Clarendon Press, Oxford.

R. Sarukkai and C. Hunter (1997). Integration of eye fixation information with speech recognition systems. In: *European Conference on Speech Communication and Technology*, pp. 97–100, Rhodes (Greece).

E. Schegloff and H. Sacks (1973). Opening up closings. *Semiotica* 8.

K. Scherer (1979). Social markers in speech. In: K. Scherer, ed., *Personality markers in speech*, pp. 147–209. Cambridge University Press.

K. Scherer (1980). The functions of nonverbal signs in conversation. In: H. G. R. St. Clair, ed., *The Social and Physiological Contexts of Language*, pp. 225–243. Lawrence Erlbaum Associates.

K. Scherer, D. Ladd and K. Silverman (1984). Vocal cues to speaker affect: Testing two models. *Journal of Acoustical Society of America* 76: 1346–1356.

F. Schiel, S. Burger, A. Geumann and K. Weilhammer (1997). The partitur format at bas. Forschungsberichte des instituts für phonetik und sprachliche kommunikation, fipkm 35, Universität München.

K.-D. Schmitz (1997). Über wichtige Aspekte bei der Einrichtung einer rechnergestützten Terminologieverwaltung. In: L. Lundquist, H. Picht and J. Qvistgaard, eds., *LSP Identity and Interface, Research, Knowledge and Society, Proceedings of the 11th European LSP Symposium on Languages for Special Purposes, Copenhagen, August 1997*, pp. 391–398, Copenhagen. Copenhagen Business School.

K.-D. Schmitz (1998). Terminographie und Terminologienormung. In: M. Snell-Hornby, H. Hönig, P. Kußmaul and P. Schmitt, eds., *Handbuch Translation*, pp. 83–91. Stauffenburg-Verlag, Tübingen.

K.-D. Schmitz, G. Budin and C. Galinski (1994). *Empfehlung für Planung und Aufbau von Terminologiedatenbanken*. Gesellschaft für Terminologie und Wissenstransfer e.V.

L. Schomaker (1998). From handwriting analysis to pen-computer applications. *IEE Electronics Communication Engineering Journal* 10(3): 93–102.

L. Schomaker, J. Nijtmans, A. Camurri, F. Lavagetto, P. Morasso, C. Benoît, T. Guiard-Marigny, B. LeGoff, J. Robert-Ribes, A. Adjoudani, I. Defée, S. Münch, K. Hartung and J. Blauert (1995a). A taxonomy of multimodal interaction in the human information processing system. Technical Report, Esprit Project 8579 MIAMI.

L. Schomaker, J. Nijtmans, A. Camurri, F. Lavagetto, P. Morasso, C. Benoît, T. Guiard-Marigny, B. LeGoff, J. Robert-Ribes, A. Adjoudani, I. Defée, S. Münch, K. Hartung and J. Blauert (1995b). Common concepts and software tools. Technical Report, Esprit Project 8579 MIAMI, WP 1.

C. Schwippert and C. Benoît (1997). Audiovisual intellegibility of an androgynous speaker. In: *AVSP'97 workshop*, Rhodos, Greece.

J. Searle (1969). *Speech acts: An essay in the philosophy of language*. CUP, Cambridge.

J. Searle (1980). *Expression and meaning*. CUP, Cambridge.

R. Shankar and D. Krishnaswamy (1993). Classification of pen gestures using learning vector quantization. In: *Neural and Stoachstic Methods in Image and Signal*, volume 1, pp. 138–143, San Diego (CA).

R. Sharma, T. Huang and V. Pavlovic (1995). A multimodal framework for interacting with virtual environments. In: C. Ntuen and E. Park, eds., *Human Interaction with Complex Systems: Conceptual Principles and Design Practice*. Kluwer Academic Publishers.

A. Shaw (1970). Parsing of graph-representable pictures. *Journal of the ACM* 17(3): 453.

B. Shneiderman (1997). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison Wesley, Menlo Park.

S. Siegel and J. Castellan Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

P. Silsbee (1994). Motion in deformable templates. In: *First IEEE Intl. Conference on Image Processing*, volume 1, pp. 323–327. IEEE.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg (1992). ToBI: a standard for labeling English prosody. In: *Proceedings of the Second International Conference on Spoken Language Processing*, volume 2, pp. 867–870, Banff, Canada.

J. M. Sinclair and R. Coulthard (1975). *Towards an analysis of discourse*. OUP,

Oxford.

SMIL (1998). "`http://www.w3.org/AudioVideo/`".

J. Sosa (1991). *Fonética y fonologia de la entonación del Espanõl Hispanoamericano.* Ph.D. thesis, University of Massachusetts, Amherst.

C. Sperberg-McQueen and L. Burnard (1994). Guidelines for text encoding and interchange (TEI P3). ACH-ACL-ALLC Text Encoding Initiative, TEI, Chicago and Oxford.

R. Srihari and C. Baltus (1993). Incorporating syntactic constraints in recognizing handwritten sentences. In: *International Joint Conference on Artificial Intelligence*, p. 1262, Chambery (France). AAAI.

T. Starner, J. Makhoul, R. Schwartz and G. Chou (1994). On-line cursive handwriting recognition using speech recognition methods. In: *International Conference on Acoustics, Speech and Signal Processing*, Adelaide. IEEE.

B. Stein and M. Meredith (1993). *The merging of the senses.* Cognitive Neuroscience series. MIT Press.

A.-B. Stenström (1990). Lexical items peculiar to spoken discourse. In: J. Svartvik, ed., *The London-Lund Corpus: Description and Research*, Lund Studies in English 82, pp. 137–176. Lund University Press.

A.-B. Stenström (1994). *An introduction to spoken interaction.* Longman, London.

R. Stetson (1928). Motor phonetics. *Archives Néerlandaises de Phonétique Expérimentale* 3: 1–216. 2nd ed., 1951, Amsterdam; re-ed. 1988, by J.A.S. Kelso & K.G. Munhall, Boston.

R. Stiefelhagen, J. Yang and U. Meier (1997a). Real time lip tracking for lipreading. In: *Eurospeech'97.*

R. Stiefelhagen, J. Yang and A. Waibel (1997b). A model-based gaze tracking system. *Int. Journal of Artificial Intelligence Tools* 6(2): 193–209.

W. Stokoe, D. Casterline and C. Croneberg (1965). *A dictionary of American Sign Language on linguistic principles.* Gallaudet College Press, Washington, D.C.

R. Stone (1991). Virtual reality and telepresence - a UK initiative. In: *Virtual Reality 91 - Impacts and Applications. Proc. of the 1st Annual Conf. on Virtual Reality.* Meckler Ltd.

D. Stork, G. Wolff and E. Levine (1992). Neural network lipreading system for improved speech recognition. *IJCNN* .

B. Stroustroup (1991). *The C++ Programming Language.* Addison Wesley, Reading, 2nd edition.

M. Stubbs (1983). *Discourse analysis: The sociolinguistic analysis of natural language.* Blackwell, Oxford.

D. Sturman (1998). Computer puppetry. *IEEE Computer Graphics and Applications* pp. 38–45.

D. Sturman and D. Zeltzer (1994). A survey of glove-based input. *IEEE Computer Graphics and Applications* 14(1): 30–39.

Q. Su and P. Silsbee (1996). Robust audiovisual integration using semicontinuous Hidden Markov Models. In: *Proceedings of ICSLP'96: The Fourth International Conference on Spoken Language Processing.*

B. Suhm (1997). Empirical evaluation of interactive multimodal error correction. In: S. Furui, B. Jang and W. Chou, eds., *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 583–590, Santa Barbara (CA). IEEE Signal Processing Society.

B. Suhm (1998). *Multimodal Interactive Error Recovery for Non-Conversational Speech User Interfaces.* Ph.D. thesis, Karlsruhe University, Germany.

B. Suhm, B. Myers and A. Waibel (1996). Interactive recovery from speech recognition errors in speech user interfaces. In: *International Conference on Spoken Language Processing*, volume 2, pp. 861–864, Philadelphia (PA).

Q. Summerfield (1992). Lipreading and audio-visual speech perception. *Philosophical transactions of the royal society of London* 335: 71–78.

H. Suonuuti (1997). *Guide to Terminology*. TSK, Nordterm 8, Helsinki.

J. Svartvik and M. Eeg-Olofsson (1982). Tagging the London-Lund Corpus of Spoken English. In: S. Johansson, ed., *Computer corpora in English language research*, pp. 85–109. Norwegian Computer Centre for the Humanities, Bergen.

M. Sweeney, M. Maguire and B. Shackel (1993). Evaluating user-computer interaction: A framework. *International Journal of Man-Machine Studies* 38: 689–711.

A. Takeuchi and S. Franks (1992). A rapid face construction lab. Technical Report SCSL-TR-92-010, Sony Computer Science Laboratory Inc.

A. Takeuchi and K. Nagao (1993). Communicative facial displays as a new conversational modality. In: *ACM/IFIP INTERCHI'93*, Amsterdam.

A. Takeuchi and T. Naito (1995). Situated facial displays: Towards social interaction. In: *International Conference on Computer Human Interaction CHI*, pp. 450–455, Denver (CO). ACM.

H. Tan, W. Rabinowitz and N. Durlach (1989). Analysis of a synthetic Tadoma system as a multidimensional tactile display. *Journal of the Acoustical Society of America* 86(3): 981–988.

R. Taylor (1989). Integrating voice, visual and manual transactions: Some practical issues from aircrew station design. In: F. N. M.M. Taylor and D. Bouwhuis, eds., *The structure of multimodal dialogue*, pp. 259–265. Elsevier Science Publishers B.V. (North Holland).

E. N. Technology (n.d.). "`http://www.foodexplorer.com/product/TECHTUT/-FF10754.HTM`".

TERMITE (1999). "`http://www.itu.int/search/wais/Termite/index.html-#Introduction`".

TERMIUM (1999). "`http://www.translationbureau.gc.ca/termium1.htm`".

D. Terzopoulos and K. Waters (1990). Physically-based facial modelling, analysis, and animation. *Journal of Visualization and Computer Animation* 1(2): 73–90.

D. Terzopoulos and K. Waters (1991). Techniques for realistic facial modelling and animation. In: N. Magnenat-Thalmann and D. Thalmann, eds., *Computer Animation '91*, pp. 45–58. Springer-Verlag.

D. Terzopoulos and K. Waters (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(6): 569–579.

S. Teufel (1996). EAGLES specifications for English morphosyntax. Draft Version. [ELM-EN] University of Stuttgart. "`ftp://ftp.ims.uni-stuttgart.de/pub/eagles/`".

S. Teufel and C. Stöckert (1996). EAGLES specifications for German morphosyntax. [ELM-DE] University of Stuttgart. "`ftp://ftp.ims.uni-stuttgart.de/pub/eagles/`".

C. Thomas (1987). Designing electronic paper to fit user requirements. In: D. Diaper and R. Winder, eds., *People and Computers III*, British Computer Society Workshop Series, pp. 247–257. Cambridge University Press, University of Exeter, proceedings of the third conference of the british computer society human-computer interaction specialist group edition.

H. Thompson (1997). Towards a base architecture for spoken language tran-

script{s,tion}. COCOSDA meeting, Rhodes.

H. Thompson, A. Anderson and M. Bader (1995). Publishing a spoken and written corpus on CD-ROM: the HCRC Map Task experience. In: G. Leech, G. Myers and J. Thomas, eds., *Spoken English on Computer: Transcription, mark-up and application*, pp. 168–180. Longman, London and New York.

K. Thórisson (1997). Layered modular action control for communicative humanoids. In: *Computer Animation'97*, Geneva, Switzerland. IEEE Computer Society Press.

H. Tillmann (1997). Eight main differences between collections of written and spoken language data. Forschungsberichte des instituts für phonetik und sprachliche kommunikation, fipkm 35, Universität München.

TM (1998). www.cse.ogi.edu/cslu/tm.

M. Turk and A. Pentland (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3(1): 71–86.

R. Uschanski, L. Delhorne, A. Dix, L. Braida, C. Reed and N. Durlach (1992). Automatic speech recognition to aid the hearing impaired. Prospects for the automatic generation of cued speech. *Journal of Rehabilitation Research and Development* .

D. Valentin, H. Abdi, A. O'Toole and G. Cottrell (1994). Connectionist models of face processing: A survey. *Pattern Recognition* 27(9): 1209–1230.

F. Van Eynde and D. Gibbon, eds. (1999, forthcoming). *Lexicon Development for Speech and Language Processing*. Kluwer, Dordrecht.

D. A. van Leeuwen and H. J. M. Steeneken (1997). Within-speaker variability of the word error rate for a continuous speech recocognition system. In: *Proceedings of Eurospeech, Rhodes, Greece*, pp. 1915–1918.

E. Vatikiotis-Bateson, K. Munhall, M. Hirayama, Y. Lee and D. Terzopoulos (1996). The dynamics of audiovisual behavior of speech. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO ASI Series. Series F: Computer and Systems Sciences*, pp. 221–232. Springler-Verlag, Berlin.

M. Vazirgiannis, Y. Theodoridis and T. Sellis (1998). Spatio-temporal composition and indexing for large multimedia applications. *Multimedia Systems* 6: 284–298.

J. Venditti (1995). Japanese ToBI Labelling Guidelines. In: K. Ainsworth-Darnell and M. D'Imperio, eds., *Ohio State Working Papers in Linguistics*, volume 50, pp. 127–162.

D. Veronda and R. Westmann (1970). Mechanical characterization of skin-finite deformations. *Journal of Biomechanics* 3: 111–124.

M. Viaud and H. Yahia (1992). Facial animation with wrinkles. In: *3rd Workshop on animation and simulation, Eurographic's 92*, Cambridge, England.

S. Vinoski (1997). CORBA: Integrating diverse applications within distributed heterogeneous environments. *IEEE Communications* 14(2).

M. Vo (1998). *A Framework and Toolkit for the Construction of Multimodal Learning Interfaces*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, USA.

M. Vo and A. Waibel (1997). Modeling and interpreting multimodal inputs: A semantic integration approach. Technical Report CMU-CS-97-192, Carnegie Mellon University.

M. Vo and C. Wood (1996). Building an application framework for speech and pen input integration in multimodal learning interfaces. In: *International Conference on Acoustics, Speech and Signal Processing*, Atlanta (GA). IEEE.

M. Vogt (1996). Fast matching of a dynamic lip model to color video sequences under regular illumination conditions. In: D. Stork and M. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems, and Applications*, volume 150 of *NATO*

ASI Series. Series F: Computer and Systems Sciences. Springler-Verlag, Berlin.

M. Vogt (1997). Interpreted multi-state lip models for audio-visual speech recognition. In: *AVSP'97 workshop*, Rhodos, Greece.

P. Vossen, M. Maguire, R. Graham and J. Heim (1998). Design guide for multimedia. Technical Report 2nd Edition, Version 4, INUSE, Telematics Applications Project IE 2016.

A. Waibel (1996). Interactive translation of conversational speech. *Computer* 29(7).

A. Waibel, B. Suhm, M. Vo and J. Yang (1997). Multimodal interfaces for multimedia information agents. In: *International Conference on Acoustics, Speech and Signal Processing*, Munich (Germany). IEEE Signal Processing Society.

C. Waite (1989). *The Facial Action Control Editor, FACE: A Parametric Facial Expression Editor for Computer Generated Animation*. Master's thesis, Massachusetts Institute of Technology, Media Arts and Sciences, Cambridge.

G. Walker and P. Sheppard (1997). Telepresence - The future of telephony. *British Telecommunications Technology Journal, Special Issue on Telepresence* October 1997.

J. Walker, L. Sproull and R. Subramani (1994). Using a human face in an interface. In: *Human Factors in Computing Systems*, pp. 85–91.

M. Walker, D. Litman, C. Kamm and A. Abella (1997). A general framework for evaluating spoken dialogue agents. In: *35th Annual Meeting of the Association of Computational Linguistics*, Madrid.

M. Walker and J. Moore (1997). Empirical studies in discourse. *Computational Linguistics* 23(1): 1–12.

C. Wang (1993). *Langwidere: A Hierarchical Spline Based Facial Animation System with Simulated Muscles*. Master's thesis, University of Calgary, Department of Computer Science, Calgary, AB.

E. Wang, H. Shahnvaz, L. Hedman, K. Papadopoulos and N. Watkinson (1993). *Human-Computer Interaction: Software and Hardware Inter-faces*, chapter A usability evaluation of text and speech redundant help messages on a reader interface, pp. 724–729. G. Salvendy and M. Smith.

J. Wang (1995). Integration of eye-gaze, voice and manual response in multimodal user interface. In: *IEEE Conference on Systems, Man, and Cybernetics*, volume 5, pp. 3938–3942, Vancouver (BC). IEEE.

M. Wang and J. Hirschberg (1992). Automatic classification of intonational phrase boundaries. *Computer Speech and Language* 6: 175–196.

W. Ward (1991). Understanding spontaneous speech: The Phoenix system. In: *International Conference on Acoustics, Speech and Signal Processing*, volume 1, pp. 365–367, Toronto (Canada). IEEE Computer Society.

K. Waters (1987). A muscle model for animating three-dimensional facial expressions. *Computer Graphics* 21(4): 17–24.

K. Waters and T. Levergood (1993). DECface: An automatic lip-synchronization algorithm for synthetic faces. Technical Report Technical Report CRL 93/4, Cambridge Research Laboratory, Digital Equipment Corporation.

K. Waters, J. Rehg, M. Loughlin, S. Kang and D. Terzopoulos (1996). Visual sensing of humans for active public interfaces. Technical Report Technical Report CRL 96/5, Cambridge Research Laboratory, Digital Equipment Corporation.

A. Wexelblat (1995). An approach to natural gesture in virtual environments. *ACM Transactions on Computer-Human Interaction* 2.

C. Wolf and P. Morrel-Samuels (1987). The use of hand-drawn gestures for text editing. *International Journal of Man-Machine Studies* 27: 91–102.

P. Woodward, T. Mohamadi, C. Benoît and G. Bailly (1992). Synthèse à partir du texte d'un visage parlant français. In: *19ième Journée d'Etudes sur la Parole*, Bruxelles.

S. E. Wright and G. Budin (1997). *Handbook of Terminology Management. Volume 1: Basic Aspects of Terminology Management.* Benjamins, Amsterdam, Philadelphia.

E. Wüster (1991). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie.* Springer, Wien, 3rd edition.

Y. Yacoob and L. Davis (1994). *Computer Vision and Pattern Recognition Conference*, chapter Computing spatio-temporal representations of human faces, pp. 70–75. IEEE Computer Society.

E. Yamamoto, S. Nakamura and K. Shikano (1997). Speech to lip movement synthesis by HMM. In: *AVSP'97 workshop*, Rhodos, Greece.

J. Yang and A. Waibel (1997). A real-time face tracker. *Int. Journal of Artificial Intelligence Tools* 6: 193–209.

B. Yuhas, M. Holdstein and T. Sejnowski (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine* pp. 65–71.

A. Yuille (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience* 3(1): 59–70.

A. Yuille, D. Cohen and P. Hallinan (1989). Feature extraction from faces using deformable templates. In: *Conference on Vision and Pattern Recognition*, pp. 104–109.

R. Zacharski, A. Monaghan, D. Ladd and J. Delin (1993). BRIDGE: Basic research on intonation in dialogue generation. Technical Report, HCRC: University of Edinburgh. Unpublished manuscript.

A SAMPA and X-SAMPA phonetic symbols

The SAMPA alphabet was developed in the late 1980s by John Wells, in consultation with a wide range of colleagues, to meet a need for a simple machine-readable encoding of phonetic transcriptions with symbols of the International Phonetic Alphabet (IPA) for file interchange purposes. At that time, standardisation of symbol codes and IPA fonts was not highly developed. The underlying principle of SAMPA was to select those IPA symbols which were conventionally used to represent phonemes in the major languages of the European Union, and to assign a 7-bit ASCII code number (below 128) to each. One of the secondary criteria was the visual similarity of the IPA symbol and the letter representing the ASCII code.

Since that time, the standardisation of IPA encoding has progressed, with the system developed by John Esling (the 'Esling codes'), and, more recently, Unicode representations. For practical purposes, however, little has changed at the time of writing, and there is still a need for a straightforward machine-readable encoding.

In the meantime, SAMPA is widely used, and extensions of SAMPA have now been developed for many other languages. In order to aid the development of such extensions, the extended code-set X-SAMPA was devised by John Wells, and encompasses the complete set of IPA conventions. For a number of symbols, human readability had to be sacrificed in favour of simple, unambiguous meachine-readability, owing to the restricted number of ASCII codes. The present collation of SAMPA and X-SAMPA is by Inge Mertins.

For further details, consult Gibbon et al. (1997) and the relevant IPA and SAMPA Internet sites, including project sites with working versions of SAMPA for specific languages.

For prosodic annotation, a number of systems are available. A number of these are discussed in Chapter 1. The most widely used in extensive corpus annotation, computational linguistics and speech technology is currently ToBI (Tones and Break Indices); the SAMPROSA system (see Gibbon et al. 1997) contains additional symbols which are suitable for more detailed dialogue transcription.

Readers should be aware that there is still considerable need for standardisation with respect to the use of IPA codes and fonts in consumer software such as word processors and Internet browsers.

VOWELS

DESCRIPTION	IPA	SAMPA/ X-SAMPA	ASCII/ ANSI
close front unrounded	i	i	105
close front rounded	y	y	121
close central unrounded	ɨ	1	49
close central rounded	ʉ	}	125
close back unrounded	ɯ	M	77
close back rounded	u	u	117
near-close front unrounded (lax i)	ɪ	I	73
near-close front rounded (lax y)	ʏ	Y	89
near-close back rounded (lax u)	ʊ	U	85
close-mid front unrounded	e	e	101
close-mid front rounded	ø	2	50
close-mid central unrounded	ɘ	@\	64, 92
close-mid central rounded	ɵ	8	56
close-mid back unrounded	ɤ	7	55
close-mid back rounded	o	o	111
mid central unrounded (schwa)	ə	@	64
open-mid front unrounded	ɛ	E	69
open-mid front rounded	œ	9	57
open-mid central unrounded	ɜ	3	51
open-mid central rounded	ɞ	3\	51, 92
open-mid back unrounded	ʌ	V	86
open-mid back rounded	ɔ	O	79
near-open front unrounded	æ	{	123
near-open central unrounded	ɐ	6	54
open front unrounded	a	a	97
open front rounded	ɶ	&	38
open back unrounded	ɑ	A	65
open back rounded	ɒ	Q	81

CONSONANTS (PULMONIC)

DESCRIPTION	IPA	SAMPA/ X-SAMPA	ASCII/ ANSI
voiceless bilabial plosive	p	p	112
voiced bilabial plosive	b	b	98
voiceless dental/alveolar plosive	t	t	116
voiced dental/alveolar plosive	d	d	100
voiceless retroflex plosive	ʈ	t`	116, 96
voiced retroflex plosive	ɖ	d`	100, 96
voiceless palatal plosive	c	c	99
voiced palatal plosive	ɟ	J\	74, 92
voiceless velar plosive	k	k	107
voiced velar plosive	g	g	103
voiceless uvular plosive	q	q	113
voiced uvular plosive	ɢ	G\	71, 92
glottal stop	ʔ	?	63
bilabial nasal	m	m	109
labiodental nasal	ɱ	F	70
dental/alveolar nasal	n	n	110
retroflex nasal	ɳ	n`	110, 96
palatal nasal	ɲ	J	74
velar nasal	ŋ	N	78
uvular nasal	ɴ	N\	78, 92
bilabial trill	ʙ	B\	66, 92
alveolar trill	r	r	114
uvular trill	ʀ	R\	82, 92
alveolar tap	ɾ	4	52
retroflex flap	ɽ	r`	114, 96
voiceless bilabial fricative	ɸ	p\	112, 92
voiced bilabial fricative	β	B	66
voiceless labiodental fricative	f	f	102
voiced labiodental fricative	v	v	118
voiceless dental fricative	θ	T	84
voiced dental fricative	ð	D	68
voiceless alveolar fricative	s	s	115
voiced alveolar fricative	z	z	122
voiceless postalveolar fricative	ʃ	S	83
voiced postalveolar fricative	ʒ	Z	90
voiceless retroflex fricative	ʂ	s`	115, 96
voiced retroflex fricative	ʐ	z`	122, 96

Consonants (pulmonic), continued

Description	IPA	SAMPA/ X-SAMPA	ASCII/ ANSI
voiceless palatal fricative	ç	C	67
voiced palatal fricative	ʝ	j\	106, 92
voiceless velar fricative	x	x	120
voiced velar fricative	ɣ	G	71
voiceless uvular fricative	χ	X	88
voiced uvular fricative	ʁ	R	82
voiceless pharyngeal fricative	ħ	X\	88, 92
voiced pharyngeal fricative	ʕ	?\	63, 92
voiceless glottal fricative	h	h	104
voiced glottal fricative	ɦ	h\	104, 92
voiceless alveolar lateral fricative	ɬ	K	75
voiced alveolar lateral fricative	ɮ	K\	75, 92
labiodental approximant	ʋ	P (*or* v\)	80 (118, 92)
alveolar approximant	ɹ	r\	114, 92
retroflex approximant	ɻ	r\‘	114, 92, 96
palatal approximant	j	j	106
velar approximant	ɰ	M\	77, 92
dental/alveolar lateral approximant	l	l	108
retroflex lateral approximant	ɭ	l‘	108, 96
palatal lateral approximant	ʎ	L	76
velar lateral approximant	ʟ	L\	76, 92

Clicks

Description	IPA	SAMPA/ X-SAMPA	ASCII/ ANSI
bilabial	ʘ	O\ (capital O)	79, 92
dental	ǀ	| \	124, 92
(post)alveolar	ǃ	! \	33, 92
palatoalveolar	ǂ	= \	61, 92
alveolar lateral	ǁ	| \ | \	124, 92, 124, 92

Ejectives, Implosives

Description	IPA	SAMPA/ X-SAMPA	ASCII/ ANSI
bilabial ejective	pʼ	p_>	112, 95, 62
dental/alveolar ejective	tʼ	t_>	116, 95, 62
velar ejective	kʼ	k_>	107, 95, 62
alveolar fricative ejective	sʼ	s_>	115, 95, 62
voiced bilabial implosive	ɓ	b_<	98, 95, 60
voiced dental/alveolar implosive	ɗ	d_<	100, 95, 60
voiced palatal implosive	ʄ	J_<	74, 92, 95, 60
voiced velar implosive	ɠ	g_<	103, 95, 60
voiced uvular implosive	ʛ	G_<	71, 92, 95, 60
The following were withdrawn from the IPA in 1993:			
voiceless bilabial implosive	ɓ̥	p_<	112, 95, 60
voiceless dental/alveolar implosive	ɗ̥	t_<	116, 95, 60
voiceless palatal implosive	ʄ̊	c_<	99, 95, 60
voiceless velar implosive	ɠ̊	k_<	107, 95, 60
voiceless uvular implosive	ʛ̥	q_<	113, 95, 60

Other Symbols

Description	IPA	SAMPA/ X-SAMPA	ASCII/ ANSI
voiceless labial-velar fricative	ʍ	W	87
voiced labial-velar approximant	w	w	119
voiced labial-palatal approximant	ɥ	H	72
voiceless epiglottal fricative	ʜ	H\	72, 92
voiced epiglottal fricative	ʢ	<\	60, 92
epiglottal plosive	ʡ	>\	62, 92
voiceless alveolo-palatal fricative	ɕ	s\	115, 92
voiced alveolo-palatal fricative	ʑ	z\	122, 92
alveolar lateral flap	ɺ	l\	108, 92
simultaneous ʃ and x	ʃ͡x	x\	120, 92
tie bar	k͡p t͡s	_	95

DIACRITICS

DESCRIPTION	IPA	SAMPA/ X-SAMPA	EXAMPLE IPA	EXAMPLE SAMPA
voiceless	◌̥	_0 (0 = figure)	n̥	n_0
voiced	◌̬	_v	s̬	s_v
aspirated	◌ʰ	_h	tʰ	t_h
more rounded	◌̹	_0 (letter O)	o̹	o_0
less rounded	◌̜	_c	o̜	o_c
advanced	◌̟	_+	u̟	u_+
retracted	◌̠	_-	e̠	e_-
centralized	◌̈	_"	ë	e_"
mid-centralized	◌̽	_x	e̽	e_x
syllabic	◌̩	= (*or* _=)	n̩	n= (*or* n_=)
non-syllabic	◌̯	_^	e̯	e_^
rhoticity	◌˞	‘	ɚ	@‘
breathy voiced	◌̤	_t	b̤	b_t
creaky voiced	◌̰	_k	ḛ	e_k
linguolabial	◌̼	_N	t̼	t_N
labialized	◌ʷ	_w	tʷ	t_w
palatalized	◌ʲ	’ (*or* _j)	tʲ	t’ (*or* t_j)
velarized	◌ˠ	_G	tˠ	t_G
pharyngealized	◌ˤ	_?\	dˤ	d_?\
velarized or pharyngealized	◌̴	_e	ɬ	l_e
velarized l, alternatively	ɫ	5		
raised	◌̝	_r	e̝	e_r
lowered	◌̞	_o	e̞	e_o
advanced tongue root	◌̘	_A	e̘	e_A
retracted tongue root	◌̙	_q	e̙	e_q
dental	◌̪	_d	t̪	t_d
apical	◌̺	_a	d̺	d_a
laminal	◌̻	_m	n̻	n_m
nasalized	◌̃	~ (*or* _~)	ẽ	e~ (*or* e_~)
nasal release	◌ⁿ	_n	dⁿ	d_n
lateral release	◌ˡ	_l	dˡ	d_l
no audible release	◌̚	_}	t̚	t_}

SUPRASEGMENTALS

DESCRIPTION	IPA	SAMPA X-SAMPA	ASCII/ ANSI
primary stress	ˈ	"	34
secondary stress	ˌ	%	37
long	ː	:	58
half-long	ˑ	:\	58, 92
extra-short	˘, eg ĕ	_X	95, 88
minor (foot) group	|		
major (intonation) group	‖	|	124
syllable break	.	$	36
linking mark	‿	-\	45, 92

TONES AND WORD ACCENTS

DESCRIPTION	IPA	SAMPA X-SAMPA	EXAMPLE IPA	EXAMPLE SAMPA
level extra high	˥ *or* ˩	_T *or* _1	e̋	e_T *or* e_1
level high	˦ *or* ˨	_H *or* _2	é	e_H *or* e_2
level mid	˧ *or* ˧	_M *or* _3	ē	e_M *or* e_3
level low	˨ *or* ˦	_L *or* _4	è	e_L *or* e_4
level extra low	˩ *or* ˥	_B *or* _5	ȅ	e_B *or* e_5
downstep	↓	!		
upstep	↑	^		
contour, rising	˨˩ *or* /	_R *or* _/ *or* _L_H	ě	e_R , e_/ , e_L_H
contour, falling	˩˨ *or* \	_F *or* _\ *or* _H_L	ê	e_F , e_\, e_H_L
contour, high rising	˧˥ *or* ˦˥	_H_T	e᷄	e_H_T
contour, low rising	˩˧ *or* ˨˧	_B_L	e᷅	e_B_L
contour, rising–falling	˩˧˥ *or* ˨˧˦	_R_F *or* _/_\ *or* _M_H_L	e᷈	e_R_F , e_/_\ e_M_H_L
global rise	↗	<R> *or* </>		
global fall	↘	<F> *or* <\>		

NB: Instead of being written as diacritics with _ , all prosodic marks can alternatively be placed in a separate tier, set off by <> , as recommended for global rise and global fall.

WIDELY USED BUT LESS STANDARDISED SYMBOLS

SAMPA	ASCII	Comment
...	46,46,46	Silent pause
	92	Phonetic case-shift (eg F might be used to signal a shift into French and would terminate the shift.
§	21	Phonological Phrase
#	35	Word Boundary
##	35,35	Absence of liaison
+	43	Morpheme boundary

B The EAGLET term database

B.1 Introduction

For a comprehensive description of the principles and standards on which the EAGLET termbank is based, see Chapter 4. The present, slightly abridged printed version was automatically created from the EAGLET termbank database. In view of rapid development of the field and the degree of specialisation of many terms, the technical literature should be consulted for detailed definitions. In doubt, rather more general definitions have been preferred.
In the printed version the following data categories are retained:

1. Orthography
   A representation of the term in standard British English orthography.
2. Pronunciation
   The phonemic transcription of the term in both IPA and SAMPA notation is given. This is unusual for a termbank, but experience in the field shows a need among non-native experts for a pronunciation guide.
3. Part of Speech
   The structure of compounds is given in attribute–value notation. For example, the term 'text-to-speech system' is analysed as '[N: [N: text][PREP: to][N: speech][N: system]]'. The tags used here are to be read as:
   - N = noun
   - V = verb
   - AJ = adjective
   - AV = adverb
   - DET = determiner
   - PREP = preposition
   - C = conjunction
   - NU = numeral

9. Inflections
   As nearly all terms in EAGLET are nouns, this category basically indicates the plural form(s) of terms. The possible values are: -s ('badger' – 'badgers'), -es ('search' – 'searches'), none ('Bayesian decision theory'). For irregular forms and the '-ies' plural in words like 'frequencies' the plural form is given in full. When 'no plural' abstract nouns and generic names denote specific instances or types, they may take a plural in some contexts, e.g. 'three LPCs' in the sense of 'three LPC analyses'.
10. Domain
    'Domain' refers to the individual chapter of Gibbon et al. (1997) or of the present volume to which the term can be assigned or in which it is defined. Subject fields such as 'physical characterisation, 'corpora', 'lexicon', 'interactive dialogue systems' are specified. For example, for 'Hidden Markov Model' the value is 'Domain: language modelling'. Many terms, however, are difficult to place because they are very general, for instance 'orthographic transcription', a term that occurs in nearly all handbook chapters and, like many others, is not restricted to the domain of spoken language technology; in such cases the Domain field has been left empty.
11. Hyperonyms
    The data category 'hyperonym' corresponds to the classical *genera proximum*

of terminological and classical definition theory. A *hyperonym* is the verbal representation of the superordinate concept of a term in a taxonomy; it is the converse of 'hyponym'. For example, *morph* is a hyperonym of *bound morph* because '*A bound morph is a type of morph*' is an acceptable sentence. Similarly, *microphone* is a hyperonym of *unidirectional microphone*, and the latter is a hyperonym of *cardioid microphone*.

12. Hyponyms
    A *hyponym* is the verbal representation of the subordinate concept of the term in question; it is the converse of 'hyperonym'. Hyponyms of very general terms are omitted. Examples: A *bound morph* is a hyponym of *morph* because *A bound morph is a kind of morph* is an acceptable sentence.
13. Synonyms
    A synonym is a term that represents the same concept as the main entry term in a term entry. In EAGLET, no distinction is made between genuine synonyms and quasi-synonyms. Quasi-synonyms are terms that represent the same concept in the same language, but for which interchangeability is limited to some contexts and inapplicable in others. An example from the *Handbook*: *wolf* is a synonym of *skilled impostor*. Strictly speaking, abbreviations are synonyms for the terms they abbreviate, and are treated as such in the termbank.
14. Cohyponyms (cf. 'Antonym' in Chapter 4
    This data category covers antonyms, i.e. terms denoting various lexical opposites, without commitment to the particular kind of antonym. It includes complementaries, i.e. terms that "divide some conceptual domain into two mutually exclusive compartments" (Cruse 1986, p. 198). For example, in the spoken language technology domain, *recognition* and *synthesis* are cohyponyms (in fact, antonyms).
15. Definitions
    As in most standard general dictionaries, EAGLET not only contains analytical definitions, i.e. definitions which give a noun phrase which formulates the meaning of the term in question (Sager and L'Homme 1994), but also definitions that contain so-called 'nonessential' characteristics and information that would be classified as 'world knowledge'. In many cases the source of the definition is given.
16. Meronymic superordinates
    Terms that are superordinates in a PARTOF hierarchy. Example: *syllable* is a meronymic superordinate of *onset* because *The/An onset is part of a syllable* is an acceptable sentence.
17. Meronymic subordinates
    Terms that are subordinates in a PARTOF hierarchy. Example: *onset* is a meronym of *syllable*, because *An onset is a part of a syllable.* is an acceptable sentence.
18. Examples
    A term and its definition is exemplified.
    Example: 'un' and 'able' in 'unbearable' are affixes.

B.2 EAGLET termbank (abridged)

2D gesture
/'tuː 'diː 'dʒestʃə/, /'tuː 'diː 'dZestS@/, [N: [AJ: 2D][N: gesture]], [plural: -s]. Domain: multimodal systems. Hyperonyms: gesture. Synonyms: graphic mark, gesture. Cohyponym: pointing, 3D gesture. Def.: 2D gestures refer to movements on a flat surface, for example marks drawn with a pen on a touch-sensitive display.

3D gesture
/'θriː 'diː 'dʒestʃə/, /'Triː 'diː 'dZestS@/, [N: [AJ: 3D][N: gesture]], [plural: -s]. Domain: multimodal systems. Hyperonyms: gesture. Cohyponym: pointing, 2D gesture. Def.: 3D gestures refer to movements of fingers, hand, or head in three dimensional space.

abstract lemma
/'æbstrækt 'leмə/, /'{bstr{kt 'lem@/, [N: [AJ: abstract][N: lemma]], [plural: abstract lemmata]. Domain: lexicon. Cohyponym: lemma, lexical lemma. Def.: An abstract lemma is a lexical database access key which may have any convenient unique name or number (or be labelled by the spelling of the canonical inflected form); all properties have equal status, so that the abstract lemma is neutral with respect to different types of lexical access, through spelling, pronunciation, semantics, etc. (Gibbon et al. 1997, p. 200)

accent identification
/'æksənt aɪdentɪfɪ'keɪʃən/, /'{ks@nt aIdentIfI'keIS@n/, [N: [N: accent][N: identification]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker classification task. Cohyponym: sex identification, age identification, mood identification, health state identification, speaker cluster identification. Def.: A task consisting in determining aspects of the social background of the speaker. (Gibbon et al. 1997, p. 409)

accent
/'æksənt/, /'{ks@nt/, [N: accent], [plural: -s]. Hyperonyms: 1. prosodic feature. Hyponyms: 1. syntactic accent, tonal accent. Synonyms: 1. pitch prominence. Def.: 1. The phonetic property which makes a particular word or syllable stand out in a stream of speech. (Crystal 1988, p. 2) 2. Regional, social or foreign pronunciation.

acceptance
/ək'septəns/, /@k'sept@ns/, [N: acceptance], [plural: -s]. Domain: speaker recognition. Hyperonyms: decision outcome (of a speaker recognition system). Hyponyms: false acceptance. Cohyponym: rejection. Def.: 1. Decision outcome which consists in responding positively to a task such as a speaker (or speaker class) verification task 2. The degree to which customers are willing to use a system or service.

accuracy
/'ækjʊrəsi/, /'{kjUr@si/, [N: accuracy], [plural: none]. Domain: speaker recognition. Hyperonyms: performance measure. Hyponyms: recognition accuracy. Synonyms: precision. Cohyponym: recall; error rate. Def.: A measure of the performance of a system such as an automatic speech recognition (ASR) system, defined as (N - S - D - I)/N, where N: number of basic units (usually words) in a test, S: number of substitution errors, D: number of deletion errors, I: number of insertion errors.

acoustic interface
/ə'kuːstɪk 'ɪntəfeɪs/, /@'kuːstIk 'Int@feIs/, [N: [AJ: acoustic][N: interface]], [plural: -s]. Domain: speech synthesis, speech recognition. Cohyponym: linguistic interface. Meronym. sup.: text-to-speech system. Def.: 1. The acoustic interface of a speech synthesiser transduces the output of the linguistic interface (lexical representation, abstract phonological code) to an audible waveform. 2. The acoustic interface of a speech recogniser converts the acoustic input signal into a set of word or sentence hypotheses.

acoustic measure
/əˈkuːstɪk ˈmeʒə/, **/@’ku:stIk ’meZ@/**, [N:[AJ: acoustic][N: measure]], [plural: -s]. Hyponyms: amplitude, intensity, fundamental frequency, F0. Def.: An acoustic measure quantifies properties of a system on the basis of properties of the speech signal it processes.

acoustic output device
/əˈkuːstɪk ˈaʊtpʊt dɪˈvaɪs/, **/@’ku:stIk ’aUtpUt dI’vaIs/**, [N: [AJ: acoustic][N: output][N: device]], [plural: -s]. Domain: multimodal systems. Hyperonyms: output device. Cohyponym: visual output device, haptic output device. Def.: Much research has been conducted into producing good quality synthetic speech, and there are several commercial products. Non-speech sounds include beep sounds, auditory icons or earcons, and auditory display (visualisation of data through sound parameters). Virtual reality systems or headphones can simulate the spatial relations of sounds.

acoustic phonetics
/əˈkuːstɪk fəˈnetɪks/, **/@’ku:stIk f@’netIks/**, [N: [AJ: acoustic][N: phonetics]], [plural: none]. Hyperonyms: phonetics. Cohyponym: articulatory phonetics, auditory phonetics. Meronym. sup.: phonetics. Def.: Acoustic phonetics is the study of the physical properties of speech sound, as transmitted between mouth and ear. (Crystal 1988)

acoustic-phonetic model
/əˈkuːstɪk fəˈnetɪk ˈmɒdəl/, **/@’ku:stIk f@’netIk ’mQd@l/**, [N: [AJ: acoustic][AJ: phonetic][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: knowledge source in a speech recognition system; model. Cohyponym: language model. Meronym. sup.: automatic speech recognition system. Def.: The acoustic-phonetic model is the conditional probability of observing the acoustic vectors when the speaker utters the words. Like the language model probabilities, these probabilities are estimated during the training phase of the recognition system. (Gibbon et al. 1997, p. 239)

acquainted impostor
/əˈkweɪntɪd ɪmˈpɒstə/, **/@’kweIntId Im’pQst@/**, [N: [AJ: acquainted][N: impostor]], [plural: -s]. Domain: speaker recognition. Hyperonyms: intentional impostor. Cohyponym: unacquainted impostor. Def.: An acquainted impostor qualifies as an intentional impostor who has some knowledge of the voice of the authorised speaker. (Gibbon et al. 1997, p. 422)

ACT
1. /ˈeɪ ˈsiː ˈtiː/ 2. /ˈækt/, **1. /’eI ’si: ’ti:/ 2. /’{kt/**, [N: ACT], [plural: -s]. Domain: consumer off-the-shelf products. Hyperonyms: tool. Synonyms: Advanced Crew Terminal. Def.: The ACT is a collection of tools that can help an astronaut in his daily work, providing electronic time schedules, procedure checking, experiment control and data acquisition. It was inplemented in a Microsoft Windows operating environment as a collection of application programs.

Action Unit
/ˈækʃən ˈjuːnɪt/, **/’{kS@n ’ju:nIt/**, [N: [N: Action][N: Unit]], [plural: -s]. Domain: multimodal systems. Hyperonyms: unit. Synonyms: AU. Def.: Basic unit used in FACS. An AU corresponds to the action of a muscle or a group of related muscles. Each AU describes the direct effect of muscle contraction as well as any secondary effects due to movement propagation, wrinkles or bulges. A facial expression is the combination of AUs. Most of the AUs combine additively. But they may also be subject to rules of dominance (an AU disappears for the benefit of another AU), substitution (an AU is eliminated when others produce the same effect), alteration (AUs cannot combine).

active contour
/'æktɪv 'kɒntʊə/, /'{ktIv 'kQntU@/, [N: [AJ: active][N: contour]], [plural: -s]. Domain: multimodal systems. Synonyms: snake. Def.: Deformable contour defined by a set of nodes connected by springs. Active contours are first located on the face. Contours are tracked by applying an image force field that is computed from the gradient of the intensity image. Muscle contraction is estimated from contour deformations. The import of visual information to recognise audio signals is around 7 percent.

active vocabulary size
/'æktɪv və'kæbjʊləri 'saɪz/, /'{ktIv v@'k{bjUl@ri 'saIz/, [N: [AJ: active][N: vocabulary][N: size]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: coverage, vocabulary size. Cohyponym: passive vocabulary size; exception vocabulary size, extension vocabulary size, user vocabulary size. Def.: A maximum number of words a speech recognition system can recognise at any given moment.

active vocabulary
/'æktɪv və'kæbjʊləri/, /'{ktIv v@'k{bjUl@ri/, [N: [AJ: active][N: vocabulary]], [plural: y/-ies]. Domain: system design. Hyperonyms: lexicon, vocabulary. Cohyponym: backup vocabulary. Def.: The vocabulary that is actively considered by a recogniser at a given instant.

activity type
/æk'tɪvɪti 'taɪp/, /{k'tIvIti 'taIp/, [N: [N: activity][N: type]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: communicative activity. Def.: A type of communicative activity with its own conventions of dialogue. E.g. negotiation, instruction, problem solving.

actual word
/'ækʃʊəl 'wɜːd/, /'{kSU@l 'w3:d/, [N: [AJ: actual][N: word]], [plural: -s]. Domain: lexicon. Hyperonyms: word, lexical item. Synonyms: lexicalised word. Cohyponym: potential word, neologism, nonce word. Def.: Word contained in a finite lexicon (Gibbon et al. 1997, p. 195)

adaptive dialogue strategy
/ə'dæptɪv 'daɪəlɒg 'strætədʒi/ , /@'d{ptIv 'daI@lQg 'str{t@dZi/ , [N: [AJ: adaptive][N: dialogue][N: strategy]], [plural: y/-ies]. Domain: interactive dialogue systems. Hyperonyms: dialogue strategy. Cohyponym: constitutive dialogue strategy, cooperative dialogue strategy, deterministic dialogue strategy. Def.: An adaptive dialogue strategy takes into account a dynamic user model by learning the user's communicative strategies and adjusting to them as each dialogue proceeds. (Gibbon et al. 1997, p. 598)

adaptive language model
/ə'dæptɪv 'læŋgwɪdʒ 'mɒdəl/, /@'d{ptIv 'l{NgwIdZ 'mQd@l/, [N: [AJ: adaptive][N: language][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: language model. Cohyponym: non-adaptive language model. Def.: Adaptive language models adapt their probabilities to the most recent history, say the last 100 to 1000 predecessor words. (Gibbon et al. 1997, p. 257)

adequacy evaluation
/'ædɪkwəsi ɪvælju'eɪʃən/, /'{dIkw@si Iv{ljU'eIS@n/, [N: [N: adequacy][N: evaluation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: evaluation. Cohyponym: diagnostic evaluation, performance evaluation, comparative assessment, benchmarking assessment. Def.: Adequacy evaluations determine the fitness of a system for a purpose: does it meet the requirements, how well, and at what cost? The requirements are mainly determined by user needs. Therefore, user needs have to be identified, which may require considerable effort in itself. Consumer reports are a typical example of adequacy evaluation.

Advanced Crew Terminal
/əd'vɑːnst 'kruː 'tɜːmɪnəl/, /@d'vA:nst 'kru: 't3:mIn@l/, [N: [AJ: Advanced][N: Crew][N: Terminal]] , [plural: -s]. Domain: consumer off-the-shelf products. Hyperonyms: tool. Synonyms: ACT. Def.: The ACT is a collection of tools that can help an astronaut in his daily work, providing electronic time schedules, procedure checking, experiment control and data acquisition. It was implemented in a Microsoft Windows operating environment as a collection of application programs.

adverb
/'ædvɜːb/, /'{dv3:b/, [N: adverb], [plural: -s]. Domain: lexicon. Hyperonyms: lexical category, part of speech, POS. Hyponyms: degree adverb, manner adverb, local adverb, temporal adverb, subjective adverb, modal adverb. Synonyms: adverbial. Cohyponym: adjective, noun, verb, preposition, conjunction, interjection. Def.: A grammatical word which modifies a verb, an adjective, or a sentence and which is relatively peripheral to the clause or sentence in which it occurs, often optional and moveable, and expressing one or a range of meanings such as time, place, manner, purpose, and reason.

adverbial
/əd'vɜːbɪəl/, /@d'v3:bI@l/, [N: adverbial], [plural: -s]. Domain: lexicon. Hyperonyms: grammatical, phrasal expression. Def.: A grammatical word or phrasal expression which modifies a verb, an adjective, or a sentence and which is relatively peripheral to the clause or sentence in which it occurs, often optional and moveable, and expressing one or a range of meanings such as time, place, manner, purpose, and reason.

affix
/'æfɪks/, /'{fIks/, [N: affix], [plural: -es]. Domain: lexicon. Hyperonyms: morph. Hyponyms: prefix, suffix, circumfix. Cohyponym: stem; infix, interfix, superfix. Meronym. sup.: word. Def.: Affixes are morphs which realise the inflectional and derivational beginnings and endings of words. (Gibbon et al. 1997, p. 215) E.g. 'un' and 'able' in 'unbearable' are affixes..

affixation
/æ'fɪkseɪʃən/, /{'fIkseIS@n/, [N: affixation], [plural: -s]. Domain: lexicon. Hyperonyms: inflection; derivation; morphological operation. Hyponyms: prefixation, suffixation; derivational affixation, inflectional affixation. Cohyponym: infixation, superfixation, interfixation. Def.: Morphological concatenation of a stem with an affix, often involving phonological modifications of the affix in the context of different stems, or of stem vowels or consonants. E.g. English 'algorithm' + 's' = 'algorithms'; English 'algebra' + 'ic' = 'algebraic'.

affricate
/'æfrɪkət/, /'{frIk@t/, [N: affricate], [plural: -s]. Domain: lexicon. Hyperonyms: consonant. Def.: A sound made when the air-pressure behind a complete closure in the vocal t ract is gradually released; the initial release produces a plosive, but the separation which follows is sufficiently slow to produce audible friction, and there is thus a fricative element in the sound also. (Crystal 1988, p. 10-11) E.g. /dZ/ in /dZVNg@l/ - jungle; /tS/ in /tS{t/ - chat.

age identification
/'eɪdʒ aɪdentɪfɪ'keɪʃən/, /'eIdZ aIdentIfI'keIS@n/, [N: [N: age][N: identification]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker classification task. Cohyponym: sex identification, health state identification, mood identification, accent identification, speaker cluster selection. Def.: When the goal is to classify a speaker within an age group, from a spoken utterance, the problem can be called age identification. (Gibbon et al. 1997, p. 409)

agent
/ˈeɪdʒənt/, **/’eIdZ@nt/**, [N: agent], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: dialogue participant. Synonyms: party. Def.: 1. In the context of interactive systems, “agent” usually refers to a dialogue participant, that is, the dialogue system or the user. 2. A human operator who takes over when a telephone-based dialogue goes wrong (’Please hold on; this call will be transferred to an agent’). 3. Anthropomorphic metaphor for intelligent search software.

agglutinative language
/əˈgluːtɪnətɪv ˈlæŋgwɪdʒ/, **/@’glu:tIn@tIv ’l{NgwIdZ/**, [N: [AJ: agglutinative][N: language]], [plural: -s]. Domain: lexicon. Hyperonyms: natural language. Cohyponym: isolating language, fusional inflectional language. Def.: Language in which large numbers of inflectional endings are concatenated (Gibbon et al. 1997, p. 197) E.g. Finnish, Turkish and Japanese display agglutination to a major extent..

agreement
/əˈgriːmənt/, **/@’gri:m@nt/**, [N: agreement], [plural: none]. Domain: lexicon. Synonyms: congruence. Cohyponym: word formation. Def.: Agreement between two or more elements of a sentence with regard to their morpho-syntactic categories (case, person, number, gender). (Bussmann, p. 404) E.g. In English ‘He sings a song’ there is congruence between subject and predicate with regard to person and number..

algorithm
/ˈælgərɪðəm/, **/’{lg@rID@m/**, [N: algorithm], [plural: -s]. Domain: language modelling. Hyperonyms: procedure. Hyponyms: search algorithm, sorting algorithm, parsing algorithm, recognition algorithm. Cohyponym: strategy, heuristic . Meronym. sub.: terminating condition, general condition . Def.: 1. A specification of a sequence of operations in a data processing procedure which terminates in a finite number of steps. 2. An algorithm has the properties of finiteness (terminates after a finite number of steps), definiteness (each step is precisely defined), generality (for all inputs it solves all problems of a particular type for which it is designed), effectiveness (all operations are mechanical and not dependent on intuition), input-output (for a given input it has a specific unique output).

alignment algorithm
/əˈlaɪnmənt ˈælgərɪðəm/, **/@’laInm@nt ’{lg@rID@m/**, [N: [N: alignment][N: algorithm]], [plural: -s]. Domain: language modelling. Hyperonyms: algorithm. Hyponyms: Viterbi alignment, Viterbi decoding, Viterbi approximation, maximum approximation. Def.: An algorithm which aligns a reference string with a hypothesis string in order to derive a measure of similarity, for example in terms of insertions, deletions and substitutions.

alignment
/əˈlaɪnmənt/, **/@’laInm@nt/**, [N: alignment], [plural: -s]. Domain: speech recognition. Hyperonyms: evaluation method. Def.: A function over two strings yielding a measure of their similarity in terms of insertions, deletions and substitutions. In determining the performance of a continuous speech recognition system, the response of the recogniser has to be compared to the transcription of the utterance presented to the system. In this process, the two word strings have to be aligned in order to compare them.

all-pass filter
/ˈɔːlpɑːs ˈfɪltə/, **/’O:lpA:s ’fIlt@/**, [N: [AJ: all][V: pass]][N: filter]], [plural: -s]. Domain: physical characterisation. Hyperonyms: filter. Synonyms: null filter. Cohyponym: low-pass filter, high-pass filter, band-pass filter, band-stop filter, notch filter. Def.: An all-pass filter passes all frequencies, i.e. in the ideal case the frequency spectrum of the input is mapped linearly into the frequency spectrum of the output.

allophone
/'æləfəʊn/, /'{l@f@Un/, [N: allophone], [plural: -s]. Domain: phonology, lexicon; structural linguistics. Hyperonyms: phone. Hyponyms: phonetic alternant, free variant, conditioned variant. Cohyponym: phoneme, contrastive phone. Def.: Allophones of a phoneme are phonetically similar variants (alternants) of that phoneme which occur in complementary environments (Gibbon et al. 1997, p. 206) 2.

alphabetic orthography
/ælfə'betɪk ɔː'θɒgrəfi/, /{lf@'betIk O:'TQgr@fi/, [N: [AJ: alphabetic][N: orthography]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: orthography. Hyponyms: Roman orthography, Cyrillic orthography, Greek orthography; standard orthography; reformed orthography. Synonyms: spelling. Cohyponym: syllabic orthography, logographic orthography. Meronym. sub.: alphabet, character. Def.: Alphabetic orthography is a concatenation of characters from a finite alphabet in a visual spatial sequence, and entails a nearly one-to-one mapping between characters and phonemes. All European languages are represented in an alphabetic orthography, with the exception of the arabic numerals, which are logographic. (Gibbon et al. 1997, p. 188)

alveolar consonant
/ælvi'əʊlə 'kɒnsənənt/, /{lvi'@Ul@ 'kQns@n@nt/, [N: [AJ: alveolar][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, labiodental consonant, dental consonant, postalveolar consonant, retroflex consonant, palatal consonant, velar consonant, uvular consonant, pharyngeal consonsant, glottal consonant. Def.: Alveolar consonants are sounds which are phonetically classified in terms of their place of articulation, made by the blade of the tongue, or the tip and the blade together, in contact with the alveolar ridge. (cf. also Crystal 1988, p. 14)

American Standard Code for Information Interchange
/ə'merɪkən 'stændəd 'kəʊd 'fɔːr ɪnfə'meɪʃən ɪntə'tʃeɪndʒ/, /@'merIk@n 'st{nd@d 'k@Ud 'fO:r Inf@'meIS@n Int@'tSeIndZ/, [N: [AJ: American][N: code][PREP: for][N: information][N: interchange]], [plural: none]. Synonyms: ASCII. Cohyponym: ANSI, EBCDIC. Def.: A standard 7-bit (decimal 0-127) numerical encoding scheme for alphabetic characters, decimal digits, punctuation marks and display control codes, widely used for plain unformatted texts, for non-proprietary information exchange via email, for standard programming languages, for many text markup and formatting languages such as SGML, LaTeX, RTF, PostScript, and for the SAMPA computer-readable phonetic alphabet.

amplitude
/'æmplɪtjuːd/, /'{mplItju:d/, [N: amplitude], [plural: -s]. Domain: physical characterisation. Hyperonyms: acoustic measure. Cohyponym: intensity, fundamental frequency, F0. Def.: The difference between a measured value of a signal and a reference line (such as zero or an averaged value over an interval of the signal). Applied to a speech signal, it yields a measure of the amount of vibration in the signal. Amplitude relates to the perceptual impression of loudness, and as the amplitude of a vibration diminishes, it becomes less audible. (cf. also Clark & Yallop, p. 207)

anacoluthon
/ænəkə'luːθɒn/, /{n@k@'lu:TQn/, [N: anacoluthon], [plural: anacolutha]. Domain: dialogue representation. Hyperonyms: dysfluency. Synonyms: syntactic blend. Def.: A type of dysfluency which takes the form of an ill-formed syntactic structure, beginning according to one structural plan, and ending according to another. E.g. The switch to a non-matching tag question in: 'And there's an accident up by the Flying Fox, is it?' (BNC, British National Corpus).

analogue representation
/ˈænəlɒg reprɪzenˈteɪʃən/, /'{n@lQg reprIzen'teIS@n/, [N: [AJ: analogue][N: representation], [plural: -s]. Domain: multimodal systems. Hyperonyms: output modality representation. Synonyms: iconic representation. Cohyponym: digital representation; linguistic representation, arbitrary representation, static-dynamic representation. Def.: 1. A representation in terms of an model in which values of variables vary along a continuous scale and correlate with the values of the continuous empirical variables they represent, as opposed to a digital or digitised representation, in which the continuous empirical variables are modelled by variables with values on a discrete scale. 2. A representation which is complementary to a symbolic linguistic representation, based on the particular physical characteristics of the object it represents. Image, sound, graphics and haptic devices may be used to give such a representation. A picture of a book may give information on the title of the book, the author, the collection, but it will not tell you who the book belongs to.

analytic approach
/ænəˈlɪtɪk əˈprəʊtʃ/, /{n@'lItIk @'pr@UtS/, [N: [AJ: analytic][N: approach]], [plural: -es]. Domain: multimodal systems. Cohyponym: global approach. Def.: A bottom-up approach to parsing (analysing) sentences by first identifying constituents, and then building higher level interpretations.

analytic testing
/ænəˈlɪtɪk ˈtestɪŋ/, /{n@'lItIk 'testIN/, [N: [AJ: analytic][N: testing]], [plural: -s]. Domain: speech synthesis. Hyperonyms: testing procedure. Cohyponym: global testing. Def.: Procedure in which the listener is instructed to evaluate specific aspects of the performance of a speech output system, e.g. suitability of tempo, quality of segments, appropriateness of word stresses, sentence accents, etc.

anechoic chamber
/ænɪˈkəʊɪk ˈtʃeɪmbə/, /{nI'k@UIk 'tSeImb@/, [N: [AJ: anechoic][N: chamber]], [plural: -s]. Domain: physical characterisation. Hyperonyms: recording room. Cohyponym: laboratory room, recording studio, soundproof booth. Def.: A room characterised by almost total lack of external noise and internal wall reflections above a critical frequency, which depends on the depth and structure of the absorptive lining of the walls. (Gibbon et al. 1997, p. 310)

annotation tier
/ænəˈteɪʃən ˈtiːə/, /{n@'teIS@n 'ti:@/, [N:[N: annotation][N: tier]], [plural: -s]. Domain: corpus representation. Hyperonyms: tier. Hyponyms: break index, tone tier, prosodic tier, segmental tier, orthographic tier. Def.: One of a set of simultanous levels of annotation of the same speech signal or text, which are generally represented as parallel streams of characters in vertical alignment corresponding intervals or points in the signal or text.

annotation tool
/ænəˈteɪʃən ˈtuːl/ , /{n@'teIS@n 'tu:l/ , [N: [N: annotation][N: tool]], [plural: -s]. Domain: corpora. Hyperonyms: tool, software. Hyponyms: text annotation tool, signal annotation tool. Synonyms: labelling tool, markup tool. Def.: A specialised tool that supports the automatic, semi-automatic or manual mark-up of corpora with annotation; e.g. a speech signal labelling tool or an automatic part-of-speech tagger.

annotation
/ænəˈteɪʃən/, /{n@'teIS@n/, [N: annotation], [plural: -s]. Domain: corpus representation. Hyperonyms: description, representation, characterisation. Hyponyms: part of speech annotation, POS annotation, segmental annotation, prosodic annotation. Synonyms: labelling, markup. Def.: 1. Symbolic description of a speech signal or text by assigning categories to intervals or points in the speech signal or to substrings or positions in the text. 2. Process of obtaining a symbolic representation of signal data. 2) The act of adding additional types of linguistic information to the transcription (representation) of a text or discourse. 3. The material added to a corpus by means of (a): e.g. part-of-speech tags.

answer system
/'ɑːnsə 'sɪstəm/, /'A:ns@ 'sIst@m/, [N:[N: answer][N: system]], [plural: -s]. Hyperonyms: Spoken language dialogue system. Def.: A spoken language dialogue system which provides an automatic telephone answering service.

anti-aliasing filter
/'æntɪ 'eɪljəsɪŋ 'fɪltə/, /'{ntI 'eIlj@sIN 'fIlt@/, [N: [AJ: anti-aliasing][N: filter]], [plural: -s]. Domain: physical characterisation. Hyperonyms: filter. Cohyponym: low-pass filter, high-pass filter, band-pass filter, band-stop filter, notch filter, all-pass filter. Def.: An anti-aliasing filter is a low-pass filter which attenuates high frequencies in an analogue signal prior to digital sampling, with the maximum cut-off threshold at half the sampling rate (Nykvist theorem, sampling theorem).

anticipatory coarticulation
/æn'tɪsɪpətəri kəʊɑːtɪkjʊ'leɪʃən/, /{n'tIsIp@t@ri k@UA:tIkjU'leIS@n/, [N: [AJ: anticipatory][N: coarticulation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: coarticulation. Synonyms: right-to-left coarticulation. Cohyponym: perseverative coarticulation, left-to-right coarticulation. Def.: Anticipatory coarticulation is the influence of a following sound on the production of an earlier sound, whereby an articulator involved in the the production of the following sound moves towards its target position for that sound during the production of the earlier sound. (cf. also Crystal 1988, p. 52) E.g. The esh-sound in 'shoe' is rounded, anticipating the lip rounding of the vowel..

antonym
/'æntənɪm/, /'{nt@nIm/, [N: antonym], [plural: -s]. Hyperonyms: co-hyponym. Hyponyms: complementary, opposite. Cohyponym: synonym. Def.: A word which stands in the lexical semantic relation of antonymy to a given word.

antonymy
/æn'tɒnəmi/, /{n'tQn@mi/, [N: antonymy], [plural: y/-ies]. Domain: lexicon. Hyperonyms: lexical semantic relation, co-hyponymy . Cohyponym: synonymy. Def.: The lexical semantic relation of antonomy is given if two words are co-hyponyms with respect to given meanings, and if they differ in meaning in respect of those details of the same meaning which are not shared by their hyperonym. E.g. Example: 'manual' and 'novel' are antonyms. Note that the term is sometimes restricted to binary oppositions, e.g. dead - alive..

API
/'eɪ 'piː 'aɪ/, /'eI 'pi: 'aI/, [N: API], [plural: -s]. Domain: system design. Hyperonyms: software interface. Hyponyms: SAPI, TAPI. Synonyms: Application Programming Interface. Cohyponym: application programme, software utility . Def.: An API is an software interface between a utility such as an input/output device (for instance a speech recogniser or speech synthesiser) and an application programme. APIs are under development for automatic speech recognition (ASR), speaker verification and text-to-speech applications. Different API definitions seem to prevail for desk top applications and telephone applications.

applet
/'æplət/, /'{pl@t/, [N: applet], [plural: -s]. Hyperonyms: computer program, software. Hyponyms: Java applet, PocketC applet. Synonyms: small scale software application. Cohyponym: large scale software application. Def.: A software application, often with restricted functionality and designed to be used as a module within a larger application. The term was originally introduced for Java applets which are downloaded on demand over the Internet and run inside a WWW browser.

applicant speaker
/'æplɪkənt 'spiːkə/, /'{plIk@nt 'spi:k@/, [N: [N: applicant][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker. Synonyms: current speaker. Def.: An applicant speaker is a speaker using a speaker recognition system at a given instant. (Gibbon et al. 1997, p. 413)

application domain
/æplɪ'keɪʃən dəʊ'meɪn/, /{plI'keIS@n d@U'meIn/, [N: [N: application][N: domain]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: domain. Def.: An application domain is a particular domain associated with particular topics and tasks, in which an application such as a spoken language dialogue system is designed to be used. E.g. training for air-traffic controllers; timetable information provision; radiology dictation.

application programming interface
/æplɪ'keɪʃən 'prəʊgræmɪŋ 'ɪntəfeɪs/, /{plI'keIS@n 'pr@Ugr{mIN 'Int@feIs/, [N: [N: application][N: programming][N: interface]], [plural: -s]. Domain: system design. Hyperonyms: software interface. Hyponyms: speech application programming interface, telephony application programming interface. Synonyms: API. Def.: An Application Programming Interface is an software interface between a utility such as an input/output device (for instance a speech recogniser or speech synthesiser) and an application programme. APIs are under development for automatic speech recognition (ASR), speaker verification and text-to-speech applications. Different API definitions seem to prevail for desk top applications and telephone applications.

application requirement profile
/æplɪ'keɪʃən rɪ'kwaɪəmənt 'prəʊfaɪl/, /{plI'keIS@n rI'kwaI@m@nt 'pr@UfaIl/, [N: [N: application][N: requirement][N: profile]], [plural: -s]. Domain: system design. Hyperonyms: requirement profile. Cohyponym: system capability profile. Def.: The application requirement profile indicates the task-related technology needed to satisfactorily meet the user expectations. It expresses what should be done by the system to be developed. (Gibbon et al. 1997, p. 32)

applications-oriented
/æplɪ'keɪʃənz 'ɔːrɪentɪd/, /{plI'keIS@nz 'O:rIentId/, [AJ: [N: applications][AJ: oriented]], Def.: A property of a particular piece of dialogue in a corpus which has been sampled with the specific aim of using it in the context of some application development; also a property of spoken language technology research and development in general.

approximant
/ə'prɒksɪmənt/, /@'prQksIm@nt/, [N: approximant], [plural: -s]. Hyperonyms: consonant; manner of articulation. Synonyms: glide, semi-vowel. Cohyponym: plosive, nasal, trill, tap, flap, fricative, lateral fricative, lateral approximant. Def.: Approximants are speech sounds which are classified phonetically on the basis of their manner of articulation: one articulator approaches another, but the degree of narrowing involved does not produce audible friction. (Crystal 1988, p. 20) E.g. /j/, /w/.

arbitrary representation
/'ɑːbɪtrəri reprɪzen'teɪʃən/, /'A:bItr@ri reprIzen'teIS@n/, [N: [AJ: arbitrary][N: representation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: output modality representation. Cohyponym: linguistic representation, analogue representation, iconic representation, static-dynamic representation. Def.: A representation which can be interpreted correctly only within a system of conventions of use. For example, a diagram should be accompanied with the necessary information to interpret it (such as name axis or scale).

archi-sign
/ɑːki'saɪn/, /A:ki'saIn/, [N: [N: archi-sign]], [plural: -s]. Domain: lexicon. Hyperonyms: lexical object. Synonyms: lexical sign class, abstract lemma, lexeme. Def.: An abstraction over a class of related lexical signs with variant representations. (Gibbon et al. 1997, p. 195)

articulation disorder
/ɑːtɪkjʊˈleɪʃən dɪsˈɔːdə/, /A:tIkjU'leIS@n dIs'O:d@/, [N: [N: articulation][N: disorder]], [plural: -s]. Domain: corpora. Hyperonyms: speech disorder. Cohyponym: resonance disorder, voice disorder, language disorder, rhythm disorder. Def.: An articulation disorder involves the distortion, deletion, or substitution of sounds or sound combinations. Usually such disorders are functional, but they may also result from lesions of the lips (e.g. a cleft lip), the palate (a cleft palate), the teeth, the tongue, the jaw, or the nose. Another possible cause of articulatory disorders is dysarthria, damage to the central of peripheral nervous system, manifested by neuromuscular disability. (Gibbon et al. 1997, p. 114)

articulator
/ɑːˈtɪkjuːleɪtə/, /A:'tIkju:leIt@/, [N: articulator], [plural: -s]. Hyperonyms: speech organ . Hyponyms: fixed articulator, movable articulator; lower lip, tongue tip, tongue blade, tongue back, velum, uvula, glottis. Cohyponym: airstream mechanism, phonation source. Def.: The articulators are vocal organs used in pairs (fixed and movable articulators) in the human production of speech sounds, co-determining the place of articulation and the manner of articulation of a speech sound.

articulatory phonetics
/ɑːˈtɪkjʊlətəri fəˈnetɪks/, /A:'tIkjUl@t@ri f@'netIks/, [N: [AJ: articulatory][N: phonetics]], [plural: none]. Cohyponym: auditory phonetics, acoustic phonetics. Meronym. sup.: phonetics. Def.: Articulatory phonetics is the study of the way speech sounds are made by the vocal organs. (Crystal 1988)

ASCII
/ˈæski/, /'{ski/, [N: ASCII], [plural: none]. Hyperonyms: numerical character code. Synonyms: American Standard Code for Information Interchange. Cohyponym: ANSI, EBCDIC . Def.: A standard 7-bit (decimal 0-127) numerical encoding scheme for alphabetic characters, decimal digits, punctuation marks and display control codes, widely used for plain unformatted texts, for non-proprietary information exchange via email, for standard programming languages, for text markup languages such as SGML and LaTeX, and for the SAMPA computer-readable phonetic alphabet.

ASR
/ˈeɪ ˈes ˈɑː/, /'eI 'es 'A:/, [N: ASR], [plural: none]. Hyperonyms: speech recognition. Synonyms: Automatic Speech Recognition. Def.: Computer input and decoding of human acoustic speech signals into sequences of word and sentence hypotheses by means of (1) an acoustic model, and (2) the selection of the optimal hypothesis by means of a language model. Both acoustic models and language models typically use statistically trained probabilistic automata such as Hidden Markov Models or Artificial Neural Nets.

assessment
/əˈsesmənt/, /@'sesm@nt/, [N: assessment], [plural: -s]. Domain: system assessment and evaluation. Hyperonyms: system development cycle. Hyponyms: comparative testing; subjective assessment, objective assessment. Synonyms: performance evaluation. Def.: Assessment is a quantitative procedure for determining the performance of a recognition system, and the evaluation of the use of the system for a particular application.

AU
/ˈeɪ ˈjuː/, /'eI 'ju:/, [N: AU], [plural: -s]. Domain: multimodal systems. Hyperonyms: unit. Synonyms: Action Unit. Def.: An Action Unit is the basic unit used in FACS. An AU corresponds to the action of a muscle or a group of related muscles. Each AU describes the direct effect of muscle contraction as well as any secondary effects due to movement propagation, wrinkles or bulges. A facial expression is the combination of AUs. Most of the AUs combine additively. But they may also be subject to rules of dominance (an AU disappears for the benefit of another AU), substitution (an AU is eliminated when others produce the same effect), alteration (AUs cannot combine).

audio signal header
/ˈɔːdɪəʊ ˈsɪgnəl ˈhedə/ , /'O:dI@U 'sIgn@l 'hed@/ , [N:[AJ: audio][N: signal][N: header]], [plural: -s]. Hyperonyms: audio file. Hyponyms: NIST-SPHERE header, SAM header; file header, RIFF/WAV file header.. Cohyponym: audio signal body. Meronym. sup.: audio signal file. Def.: An audio signal header is a global annotation of a speech signal file containing information about speaker, recording context such as microphone etc., and signal recording properties such as sampling rate, dynamic resolution and byte order. The header may be contained in the same file.

audio-driven face synthesis
/ˈɔːdɪəʊ ˈdrɪvən ˈfeɪs ˈsɪnθəsɪs/, /'O:dI@U 'drIv@n 'feIs 'sInT@sIs/, [N: [AJ: audio-driven][N: face][N: synthesis]], [plural: audio-driven face syntheses]. Domain: multimodal systems. Hyperonyms: face synthesis. Cohyponym: performance-driven face synthesis, puppeteer control face synthesis, text-to-visual-speech face synthesis . Def.: In order to drive a face synthesiser, pre-recorded speech is analysed and information about phonemes, pauses and their respective durations is extracted from speech; additional paralinguistic vocal features (e.g. speech rhythm, intonation, loudness) can also be analysed. Phonemes which have been identified are associated with facial control parameters to compute the appropriate mouth shape. Linear prediction analysis, sound segmentation, TDNN, HMM modelling and decoding techniques have been used to generate mouth shapes.

audiology
/ɔːdiˈɒlədʒi/, /O:di'Ql@dZi/, [N: audiology], [plural: none]. Domain: corpora. Hyperonyms: speech sciences. Cohyponym: speech therapy, phonetics, phonology . Def.: Audiology is the scientific study of hearing. (Gibbon et al. 1997, p. 91)

audiometer
/ɔːdiˈɒmɪtə/, /O:di'QmIt@/, [N: audiometer], [plural: -s]. Domain: corpora. Def.: An audiometer is a measuring instrument which is used to test the intensity and frequency range of pure tones that the human ear can detect. (Gibbon et al. 1997, p. 91)

Audiotex
/ˈɔːdɪəʊteks/, /'O:dI@Uteks/, [N: Audiotex], [plural: none]. Hyperonyms: telephone call service. Def.: A system providing an Audiotex service plays pre-recorded messages to telephone callers. The purpose of such services is to inform or to entertain. Audiotex services are usually made available with Premium Rate Tariffs. Audiotex services tend to be tightly regulated, and they are not available in some countries. E.g. weather forecasts, traffic information, horoscopes, joke lines.

auditory icon
/ˈɔːdɪtəri ˈaɪkən/, /'O:dIt@ri 'aIk@n/, [N: [AJ: auditory][N: icon]], [plural: -s]. Domain: multimodal systems. Cohyponym: visual icon, earcon. Def.: Sounds with a natural or intuitively obvious relation to some form of human-computer interaction.

auditory phonetics
/ˈɔːdɪtəri fəˈnetɪks/, /'O:dIt@ri f@'netIks/, [N: [AJ: auditory][N: phonetics]], [plural: none]. Hyperonyms: phonetics. Cohyponym: articulatory phonetics, acoustic phonetics. Meronym. sup.: phonetics. Def.: 1. Auditory phonetics is the use of auditory perception by a trained phonetician in order to analyse and transcribe speech sounds. 2. Auditory phonetics is the study of the perceptual response to speech sounds, as mediated by ear, auditory nerve and brain. (Crystal 1988)

automated speech output testing
/ɔːtəˈmeɪtɪd ˈspiːtʃ ˈaʊtpʊt ˈtestɪŋ/, /O:t@'meItId 'spi:tS 'aUtpUt 'testIN/, [N: [AJ: automated][N: speech][N: output][N: testing]], [plural: -s]. Domain: speech synthesis. Hyperonyms: speech output testing. Hyponyms: human speech output testing. Synonyms: objective assessment. Cohyponym: subjective assessment. Meronym. sub.: algorithm, formal assessment model. Def.: A speech output assessment procedure in which the human observer (listener in the case of audio output, or linguist in the case of symbolic output) has been replaced (modelled) by an algorithm. Automated assessment presupposes that we know exactly how human observers evaluate differences between two (acoustic or symbolic) realisations of the same linguistic message.

automatic dialogue system
/ɔːtəˈmætɪk ˈdaɪəlɒg ˈsɪstəm/, /O:t@'m{tIk 'daI@lQg 'sIst@m/, [N: [AJ: automatic][N: dialogue][N: system]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: dialogue system. Def.: A spoken language dialogue system application which functions without an intervening human operator.

automatic segmentation
/ɔːtəˈmætɪk segmenˈteɪʃən/, /O:t@'m{tIk segmen'teIS@n/, [N: [AJ: automatic][N: segmentation]], [plural: -s]. Domain: corpora. Hyperonyms: segmentation. Cohyponym: semi-automatic segmentation, manual segmentation. Def.: The segmentation of a speech signal into phonemes, diphones, demi-syllables, syllables, words or sentences by means of an algorithm which assigns time stamps to points or intervals in the speech signal and stores these in a segmentation annotation file.

automatic speech recognition system
/ɔːtəˈmætɪk ˈspiːtʃ rekəgˈnɪʃən ˈsɪstəm/, /O:t@'m{tIk 'spi:tS rek@g'nIS@n 'sIst@m/, [N: [AJ: automatic][N: speech][N: recognition][N: system]], [plural: -s]. Domain: speech recognition. Hyperonyms: speech recognition system. Cohyponym: human speech recognition faculty; automatic speech synthesis system. Def.: A computer system for automatic speech recognition.

automatic speech recognition
/ɔːtəˈmætɪk ˈspiːtʃ rekəgˈnɪʃən/, /O:t@'m{tIk 'spi:tS rek@g'nIS@n/, [N: [AJ: automatic][N: speech][N: recognition]], [plural: none]. Hyperonyms: speech recognition. Synonyms: ASR. Cohyponym: human speech recognition; automatic speech synthesis. Def.: The computer input and decoding of human acoustic speech signals into sequences of word and sentence hypotheses by means of (1) an acoustic model, and (2) the selection of the optimal hypothesis by means of a language model. Both acoustic models and language models typically use statistically trained probabilistic automata such as Hidden Markov Models or Artificial Neural Nets.

autosegmental phonology
/ɔːtəʊseg'mentəl fəˈnɒlədʒi/, /O:t@Useg'ment@l f@'nQl@dZi/, [N: [AJ: autosegmental][N: phonology]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: phonological theory, prosodic phonology. Cohyponym: generative phonology, metrical phonology, phonemic phonology. Def.: A phonological theory which represents phonological structures as a lattifce of parallel feature values, with a temporal partial ordering, rather than as matrices of phonemes and features. The theory was developed by Leben, Goldsmith, Clements and others in the 1970s to describe the association of prosodic features with basic timing units such as segments or syllables, and has since become one of the standard theories of phonology. The relevance of autosegmental phonology to future generations of automatic speech recognition systems in terms of partially synchronised independent feature streams, has been demonstrated in various studies by Moore, Carson-Berndsen, Kirchhoff.

average noise consumption
/ˈævərɪdʒ ˈnɔɪz kənˈsʌmpʃən/, /'{v@rIdZ 'nOIz k@n'sVmpS@n/, [N: [AJ: average][N: noise][N: consumption]], [plural: -s]. Domain: physical characterisation. Hyperonyms: measure. Def.: Average noise consumption is the kind and amount of noise a subject is frequently exposed to. It gives a clue to possible hearing losses and to the degree to which a subject is accustomed to noisy environments.

backchanneling
/ˈbækˈtʃænəlɪŋ/, /'b{k'tS{n@lIN/, [N: backchanneling], [plural: -s]. Hyperonyms: turn-taking. Synonyms: feedback utterance. Cohyponym: full turn. Def.: A type of dialogue turn or utterance whose function is to influence the turn-taking behaviour of one's interlocutor without conveying any propositional information; e.g. uh-huh, mhm.

background stationary noise
/ˈbækgraʊnd ˈsteɪʃənəri ˈnɔɪz/, /'b{kgraUnd 'steIS@n@ri 'nOIz/, [N: [N: background][AJ: stationary][N: noise]], [plural: none]. Domain: physical characterisation. Hyperonyms: noise. Def.: Background stationary noise is noise of constant characteristics in the environment of a speech signal.

backing-off distribution
/ˈbækɪŋ ˈɒf dɪstrɪˈbjuːʃən/, /'b{kIN 'Qf dIstrI'bju:S@n/, [N: [N: backing][PREP: off][N: distribution]], [plural: -s]. Domain: language modelling. Hyponyms: singleton backing-off distribution. Synonyms: singleton distribution. Def.: A distribution estimate based on a backing-off procedure.

backing-off
/ˈbækɪŋ ˈɒf/, /'b{kIN 'Qf/, [N: [N: backing][PREP: off]], [plural: none]. Domain: language modelling. Hyperonyms: model, procedure. Def.: Specific procedure for smoothing estimates of the probability of occurrence of phenomena that have not been observed often enough to make straightforward estimates.

backup vocabulary
/ˈbækʌp vəˈkæbjʊləri/, /'b{kVp v@'k{bjUl@ri/, [N: [N: backup][N: vocabulary]], [plural: y/-ies]. Hyperonyms: vocabulary. Cohyponym: active vocabulary. Def.: Backup vocabulary is the complement of the active vocabulary within the overall vocabulary of a spoken language system, i.e. the set of words which are not currently in active use in the system.

backward looking communicative function
/ˈbækwəd ˈlʊkɪŋ kəˈmjuːnɪkətɪv ˈfʌŋkʃən/, /'b{kw@d 'lUkIN k@'mju:nIk@tIv 'fVNkS@n/, [N: [AV: backward][V: looking][AJ: communicative][N: function]], [plural: -s]. Hyperonyms: discourse coherence. Synonyms: anaphoric function. Cohyponym: forward looking communicative function. Def.: A communicative function that refers back or relates to an item that has previously been mentioned.

badger
/ˈbædʒə/, /'b{dZ@/, [N: badger], [plural: -s]. Domain: speaker recognition. Hyperonyms: impostor. Synonyms: poor impostor. Cohyponym: skilled impostor, wolf. Def.: Impostor with a low success rate in claiming an identity averaged over each claimed identity. (Gibbon et al. 1997, p. 441)

band-pass filter
/ˈbændpɑːs ˈfɪltə/, /'b{ndpA:s 'fIlt@/, [N: [N: band][V: pass][N: filter]], [plural: -s]. Domain: physical characterisation. Hyperonyms: filter. Cohyponym: low-pass filter, high-pass filter, band-stop filter, notch filter, all-pass filter. Def.: A band-pass filter removes or reduces the amplitude of frequencies in a band between specified upper and lower frequency thresholds, for instance for removing both low and high frequency noise from a signal. It is equivalent to a cascade of a low-pass filter and a high-pass filter, in which the cut-off frequency of the low-pass filter is higher than that of the high-pass filter.

band-stop filter
/ˈbændstɒp ˈfɪltə/, /’b{ndstQp ’fIlt@/, [N: [N: band][V: stop][N: filter]], [plural: -s]. Domain: physical characterisation. Hyperonyms: filter. Synonyms: notch filter. Cohyponym: low-pass filter, high-pass filter, band-pass filter, notch filter, all-pass filter. Def.: A band-stop filter passes all frequencies except frequencies between specified high and low thresholds, which it attenuates.

bandwidth
/ˈbændwɪdθ/, /’b{ndwIdT/, [N: bandwidth], [plural: -s]. Domain: physical characterisation. Hyperonyms: frequency range. Def.: The bandwidth of a signal is determined by the range of the frequencies which constitute the signal. E.g. An analogue telephone signal has a bandwidth of about 3000 Hz, a GSM data signal 9600 Hz, an amplitude modulated shortwave radio signal about 5000 Hz, a video signal above 5 MHz depending on the picture resolution..

barge-in
/ˈbɑːdʒ ˈɪn/, /’bA:dZ ’In/, [N: [V: barge][PREP: in]], [plural: -s]. Hyperonyms: interruption, interrupt. Synonyms: cut-through, talkover. Def.: The ability for the human to speak over a system prompt or system output. Barge-in is assumed to be of great importance in spoken dialogue systems for frequent users. Two types of barge-in must be distinguished: one in which the human can only interrupt the system output, but without being understood; and another in which the human can stop the system output by starting to speak and the speech is understood.

baseline reference condition
/ˈbeɪslaɪn ˈrefərəns kɒnˈdɪʃən/, /’beIslaIn ’ref@r@ns kQn’dIS@n/, [N:[N: baseline][N: reference][N: condition]], [plural: -s]. Hyponyms: reference condition. Cohyponym: topline reference condition. Def.: A baseline reference condition is a standardised minimal reference condition, such as the output of a spoken language system that contains no specific intelligence.

Bayes decision rule
/ˈbeɪz dɪˈsɪʒən ˈruːl/, /’beIz dI’sIZ@n ’ru:l/, [N: [N: Bayes][N: decision] [N: rule]], [plural: none]. Domain: language modelling, statistical decision theory. Hyperonyms: decision procedure. Def.: The Bayes decision rule defines the conditional probability of an event in a given context in terms of its prior and posterior probabilities. In mainstream approaches to automatic speech recognition, the Bayes decision rule is used to select the best word or sentence hypothesis by finding the hypothesis to which the largest product of the probabilities defined by the language model (treated as prior probability) and the acoustic model (treated as posterior probability) is assigned.

benchmark test
/ˈbentʃmɑːk ˈtest/ , /’bentSmA:k ’test/ , [N: [N: benchmark][N: test]], [plural: -s]. Domain: speech synthesis, speech recognition, assessment methodologies, multimodal systems. Hyperonyms: evaluation method, test. Def.: An efficient, easily administered quantitative testing procedure or set of tests that can be used to express the performance of a speech output system (or some system module) in numerical terms with reference to some pre-defined standard of performance.

benchmark
/ˈbentʃmɑːk/, /’bentSmA:k/, [N: benchmark], [plural: -s]. Domain: assessment methodologies, speech synthesis, speech recognition. Meronym. sup.: benchmark test. Def.: The value that characterises some reference system against which a newly developed system is (implicitly) set off.

bidirectional microphone
/baɪdaɪ'rekʃənəl 'maɪkrəfəʊn/, /baIdaI'rekS@n@l 'maIkr@f@Un/, [N: [AJ: bidirectional][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: microphone, directional microphone. Cohyponym: unidirectional microphone, omnidirectional microphone, ultradirectional microphone, pressure zone microphone, headset microphone; handheld microphone, table-top microphone, room microphone, headmounted microphone. Def.: Bidirectional microphones are most sensitive at the front and at the rear. There is a plane of minimum sensitivity perpendicular to the direction of maximum sensitivity. This behaviour makes bidirectional microphones most suited for the recording of more than one speaker. Bidirectional microphones should not be used to produce speech recordings from one speaker. The bidirectional microphone also exhibits the proximity effect. The effect is approximately 6 db stronger as compared to cardioid microphones. (Gibbon et al. 1997, p. 305)

bigram count
/'baɪgræm 'kaʊnt/, /'baIgr{m 'kaUnt/, [N:[N: bigram][N: count]], [plural: -s]. Hyperonyms: count. Cohyponym: trigram count, n-gram count. Def.: The number of bigrams (i.e. pairs of adjacent tokens) in a corpus. Bigrams can be further defined in terms of any kind of linguistic unit, but are usually taken to be words.

bigram grammar
/'baɪgræm 'græmə/, /'baIgr{m 'gr{m@/, [N: [N: bigram][N: grammar]], [plural: -s]. Domain: language modelling. Hyperonyms: grammar. Cohyponym: trigram grammar. Def.: A probabilistic grammar based on transition probabilities of words, predicting the probability of a word in a given context from the product of the a priori probability of the word and the probability of its predecessor. The transition probabilities of words are calculated from their distribution in a corpus. Analogously, the probability of a word in a trigram or n-gram grammar is calculated using the probabilities of the preceding two or (n-1) words.

bigram language model
/'baɪgræm 'læŋgwɪdʒ 'mɒdəl/, /'baIgr{m 'l{NgwIdZ 'mQd@l/, [N: [N: bigram][N: language][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: language model. Synonyms: bigram model, bigram grammar. Cohyponym: unigram language model, trigram language model, n-gram language model. Def.: A bigram language model is a language model which consists of a bigram grammar.

bigram
/'baɪgræm/, /'baIgr{m/, [N: bigram], [plural: -s]. Domain: language modelling. Cohyponym: zerogram, unigram, trigram. Def.: In language modelling a bigram is a sequence of two words. (Gibbon et al. 1997, p. 94)

bilabial consonant
/baɪ'leɪbɪəl 'kɒnsənənt/, /baI'leIbI@l 'kQns@n@nt/, [N: [AJ: bilabial][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: labiodental consonant, dental consonant, alveolar consonant, postalveolar consonant, retroflex consonant, palatal consonant, velar consonant, uvular consonant, pharyngeal consonant, glottal consonant. Def.: A bilabial consonant is a consonant sound classified on the basis of the place of articulation: it refers to a sound made by touching or closely approximating the upper and lower lips. (Crystal 1988, p. 33)

binary search
/'baɪnəri 'sɜːtʃ/, /'baIn@ri 's3:tS/, [N: [AJ: binary][N: search]], [plural: -es]. Domain: language modelling. Hyperonyms: search algorithm. Cohyponym: linear search, exhaustive search, heuristic search . Def.: Binary search is a strategy which structures the search space into a balanced binary tree, yielding reduced complexity - e.g. n*log(n) instead of n*n - and consequently reduced search time and greater efficiency.

black box approach
/'blæk 'bɒks ə'prəʊtʃ/, /'bl{k 'bQks @'pr@UtS/, [N: [AJ: black][N: box][N: approach]], [plural: -es]. Domain: speech synthesis. Hyperonyms: testing procedure, evaluation, assessment. Cohyponym: glass box approach. Def.: Performance evaluation of a system as a whole, typically used to compare systems developed by different manufacturers, or to establish the improvement of one system relative to an earlier edition (comparative testing). Black box evaluations consider the overall performance of a system without reference to any internal components or behaviours. Evaluations of this kind address large questions such as "How good is it as an integrated system?" rather than detailed questions of the "What is its word recognition rate?" variety. (Gibbon et al. 1997, p. 840)

bound morph
/'baʊnd 'mɔːf/, /'baUnd 'mO:f/, [N: [AJ: bound][N: morph]], [plural: -s]. Domain: lexicon. Hyperonyms: morph. Cohyponym: free morph. Meronym. sub.: phoneme. Def.: A bound morph is a morph (generally an affix) which always occurs together with at least one other morph (typically a stem) in the same word. (Gibbon et al. 1997, p. 215) E.g. English 's' as in 'sees', 'ed' as in 'limited', 'pre' as in 'preselect'..

boundary position assignment
/'baʊndəri pə'zɪʃən ə'saɪnmənt/, /'baUnd@ri p@'zIS@n @'saInm@nt/, [N:[N: boundary][N: position][N: assignment]], [plural: -s]. Domain: . Hyperonyms: annotation. Synonyms: segmentation. Meronym. sup.: linguistic interface. Def.: Assignment of a boundary annotation to a speech signal by a segmentation procedure.

Braille
/'breɪl/, /'breIl/, [N: Braille], [plural: none]. Domain: multimodal systems. Def.: A form of printing by embossing groups of raised point marks which blind people can read by touching. (Longman 1992, p.137))

break index
/'breɪk 'ɪndeks/, /'breIk 'Indeks/, [N: [N: break][N: index]], [plural: break indices]. Domain: corpora. Hyperonyms: annotation tier. Def.: A tier in ToBI annotation indicating a perceived juncture between words transcribed on the orthographic tier.

breath noise
/'breθ 'nɔɪz/, /'breT 'nOIz/, [N: [N: breath][N: noise]], [plural: -s]. Domain: physical characterisation. Hyperonyms: noise. Def.: Noise caused by exhaled air passing directly over the microphone.

breathy voice
/'breθi 'vɔɪs/, /'breTi 'vOIs/, [N: [AJ: breathy][N: voice]], [plural: -s]. Domain: physical characterisation. Def.: A breathy voice results from slow, sometimes incomplete closure of the vocal folds during the laryngeal cycle.

broad phonetic transcription
/'brɔːd fə'netɪk træn'skrɪpʃən/, /'brO:d f@'netIk tr{n'skrIpS@n/, [N: [AJ: broad][AJ: phonetic][N: transcription]], [plural: -s]. Domain: lexicon, corpora. Hyperonyms: phonetic transcription. Cohyponym: narrow phonetic transcription. Def.: A transcription with less detail than a full phonetic transcription. The broadest possible transcription is a phonemic transcription, i.e. a transcription in which only those phonetic segments are notated which correspond to contrastive, i.e. functionally distinctive units in the language. Some phoneticians use 'broad' exclusively in the sense of 'phonemic'. (see also Crystal 1988, p. 313)

bus
/'bʌs/, /'bVs/, [N: bus], [plural: buses]. Domain: system design. Hyperonyms: interface. Hyponyms: 8-bit bus, 16-bit bus, 32-bit bus, 64-bit bus; SCSI bus, ISA bus, EISA bus . Def.: A hardware interface between the components of a computer enabling low-level parallel data interchange between processors, memory, and input/output devices.

C++
/'siː 'plʌs 'plʌs/, /'si: 'plVs 'plVs/, [N: C++], [plural: none]. Hyperonyms: object-oriented programming language. Def.: C++ is the standard object-oriented programming language for standalone applications both for spoken language processing and other purposes. C++ is being standardised by the ISO.

canned speech
/'kænd 'spiːtʃ/, /'k{nd 'spi:tS/, [N: [AJ: canned][N: speech]], [plural: none]. Domain: speech synthesis. Hyperonyms: recorded speech. Synonyms: pre-recorded speech. Cohyponym: speech synthesis. Def.: Speech which has been recorded for use directly in the prompts or information play-outs of a dialogue system is referred to as canned speech or canned messages. Canned messages can be concatenated to create a single system utterance, with careful attention to prosodic issues in order to produce a high quality, natural-sounding interface. Speech synthesis, though less natural-sounding, is more flexible and thus more appropriate when lengthy, lexically rich, or numerous complex system utterances are required.

cardioid microphone
/'kɑːdɪɔɪd 'maɪkrəfəʊn/, /'kA:dIOId 'maIkr@f@Un/, [N: [AJ: cardioid][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: unidirectional microphone. Cohyponym: hypercardioid microphone, supercardioid microphone. Def.: Cardioid microphones are unidirectional microphones which show best ambient noise suppression for incident sound from the back. Sensitivity loss is about 6 db at the sides of the microphone and 15-25 db at the rear. (Gibbon et al. 1997, p. 304)

CART
/'kɑːt/, /'kA:t/, [N: CART], [plural: -s]. Domain: language modelling. Hyperonyms: classification. Synonyms: classification and regression tree. Def.: A procedure used, for example in text to speech synthesis, for inferring grammars from a training data set.

casual impostor
/'kæʒʊəl ɪm'pɒstə/, /'k{ZU@l Im'pQst@/, [N: [AJ: casual][N: impostor]], [plural: -s]. Domain: speaker recognition. Hyperonyms: impostor. Def.: Speaker who is used as impostor in an evaluation, but who was not recorded with the explicit instruction to try to defeat the system. (Gibbon et al. 1997, p. 422)

casual registered speaker
/'kæʒʊəl 'redʒɪstəd 'spiːkə/, /'k{ZU@l 'redZIst@d 'spi:k@/, [N: [AJ: casual][AJ: registered][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: registered speaker. Def.: Registered speaker who has not received an explicit instruction to succeed in being identified or verified positively. Or who is not even aware that he is being recorded. (Gibbon et al. 1997, p. 422)

categorical estimation
/kætə'gɒrɪkəl estɪ'meɪʃən/, /k{t@'gQrIk@l estI'meIS@n/, [N: [AJ: categorical] [N: estimation]], [plural: -s]. Domain: speech synthesis. Hyperonyms: rating method. Def.: Categorial estimation is a rating method where the subject has to assign to (some aspect of) a speech output system a value from a limited range of prespecified values, e.g. "1" representing extremely poor and "10" excellent intelligibility.

CD-ROM
/'siː 'diː 'rɒm/, /'si: 'di: 'rQm/, [N: CD-ROM], [plural: -s]. Hyperonyms: disk. Hyponyms: ISO 9660 CD-ROM, Joliet CD-ROM, Mac CD-ROM, Adaptec CD-ROM . Synonyms: Compact-Disk Read Only Memory. Cohyponym: Audio Compact Disk, audio CD. Def.: Compact disk for the storage of data, e.g. software, signal data, etc., with different file structure from audio CD-ROMs. The international CD-ROM standard is ISO 9660 (restricted directory depth, short '8.3' file names, restricted filename character set), but more flexible de facto specifications (e.g. Joliet) have been developed for PC and Mac environments, as well as specifications for hybrid audio and data file CD-ROMs.

CELP coding
/ˈkelp ˈkəʊdɪŋ/, /'kelp 'k@UdIN/, [N: [AJ: CELP] [N: coding]], [plural: none]. Synonyms: Codebook-Excited Linear Predictive coding. Def.: A form of speech coding using linear prediction in which the excitation of the linear predictive filter is drawn from a codebook of possibilities.

cepstrum
/ˈkepstrəm/, /'kepstr@m/, [N: cepstrum], Hyperonyms: speech signal transformation. Def.: The cepstrum is defined as the inverse Fourier transform of the log of the short-term power spectrum. It is widely used in speech recognition as the basis of a method for extracting the pitch track (F0 trajectory) from the speech signal.

CGU
/ˈsiː ˈdʒiː ˈjuː/, /'si: 'dZi: 'ju:/, [N: CGU], [plural: -s]. Hyperonyms: annotation category. Synonyms: Common Ground Unit. Def.: An abstract category of pragmatic meso-level annotation, comprising all units of speech that are relevant to developing mutual understanding of a topic in a dialogue.

channel characteristic
/ˈtʃænəl kærəktəˈrɪstɪk/, /'tS{n@l k{r@kt@'rIstIk/, [N: [N: channel][N: characteristic], [plural: -s]. Domain: physical characterisation. Def.: Information relating to the physical recording conditions for speech data, i.e. type of microphone, number of recording channels, etc.

CHILDES
/ˈtʃaɪldəs/, /'tSaIld@s/, [N: CHILDES], [plural: none]. Synonyms: child language data exchange system. Def.: A corpus annotation system for the exchange of computer-readable language data. It was originally developed within the field of child language to foster the sharing of transcribed language data of children's spontaneous speech. CHILDES consists of an archive of recordings and annotations, and of an annotation and processing software to create and access CHILDES data.

chroma-key technique
/ˈkrəʊmə ˈkiː tekˈniːk/, /'kr@Um@ 'ki: tek'ni:k/, [N: [N: chroma-key][N: technique]], [plural: -s]. Domain: multimodal systems. Def.: Technique used in face recognition in order to detach the lips from the image background.

chunk
/ˈtʃʌŋk/, /'tSVNk/, [N: chunk], [plural: -s]. Domain: corpora. Hyperonyms: speech unit. Synonyms: segment. Def.: An general term for units into which speech can be segmented or according to which it is organised.

circumfix
/ˈsɜːkəmfɪks/, /'s3:k@mfIks/, [N: circumfix], [plural: -es]. Domain: lexicon. Hyperonyms: affix. Cohyponym: prefix, suffix, infix, interfix, superfix. Def.: An inflectional or derivational morpheme realised as a combination of prefix and suffix. E.g. German past participle 'gewartet': prefix 'ge-' + stem 'wart' + suffix '-et', infinitive 'warten' to wait, to service a machine.

citation form
/saɪˈteɪʃən ˈfɔːm/, /saI'teIS@n 'fO:m/, [N: [N: citation][N: form]], [plural: -s]. Domain: lexicon. Def.: A representation of the pronunciation of a word in isolation, often used as the canonical representation of a word in a lexicon. (Gibbon et al. 1997, p. 205)

classification and regression tree
/klæsɪfɪˈkeɪʃən ˈænd rɪˈgreʃən ˈtriː/, /kl{sIfI'keIS@n '{nd rI'greS@n 'tri:/, [N: [N: classification][C: and][N: regression][N: tree]], [plural: -s]. Domain: statistics. Hyperonyms: classification. Synonyms: CART. Def.: A procedure used, for example in text to speech synthesis, for inferring grammars from a training data set.

classification
/klæsɪfɪ'keɪʃən/, **/kl{sIfI'keIS@n/**, [N: classification], [plural: -s]. Def.: The procedure of classifying segmented data, for example the classification of phones as allophones on the grounds of distinctiveness, minimality, phonetic similarity and complementary distribution (i.e. their occurrence in complementary contexts as contextual variants of that phoneme). (Gibbon et al. 1997, p. 206)

clitic
/'klɪtɪk/, **/'klItIk/**, [N: clitic], [plural: -s]. Domain: lexicon. Hyperonyms: functional word. Def.: A clitic is a functional word which merges at the boundaries of a lexical word to form a sequence which behaves as a phonological unit, i.e. as a functional unit. (Gibbon et al. 1997, p. 220) E.g. English 'I'm coming' /aIm kVmIN/; 'he's' /hi:z/ for 'he is'. .

closed vocabulary
/'kləʊzd və'kæbjʊlərɪ/, **/'kl@Uzd v@'k{bjUl@ri/**, [N: [AJ: closed][N: vocabulary]], [plural: y/-ies]. Domain: lexicon, language modelling. Hyperonyms: vocabulary. Cohyponym: open vocabulary. Def.: 1. A fixed finite vocabulary, for example in a speech recogniser or speech synthesiser. 2. The fixed finite subset of the grammatical words (function words) of a language.

cluttering
/'klʌtərɪŋ/, **/'klVt@rIN/**, [N: cluttering], [plural: none]. Domain: corpora. Hyperonyms: rhythm disorder. Cohyponym: stuttering, stammering. Def.: A speech defect. The primary characteristic of cluttering is that the patient tries to talk too quickly, and as a result introduces distortions into his rhythm and articulation. (Gibbon et al. 1997, p. 115)

co-hyponym
/kəʊ'haɪpənɪm/, **/k@U'haIp@nIm/**, [N: co-hyponym], [plural: -s]. Domain: lexicon, semantics. Hyperonyms: hyponym. Hyponyms: synonym, antonym. Def.: A lexical item standing in the lexical semantic relation of co-hyponymy to a given lexical item. E.g. 'Manual' and 'novel' are co-hyponyms in relation to 'book'..

co-hyponymy
/kəʊhaɪ'pɒnəmi/ , **/k@UhaI'pQn@mi/** , [N: [N: co-hyponymy]], [plural: y/-ies]. Domain: lexicon, semantics. Hyperonyms: lexical semantic relation. Hyponyms: synonymy, antonymy. Def.: A lexical semantic relation between two words that have the same hyperonym or superordinate term (in the same meaning of the hyperonym). (Gibbon et al. 1997, p. 201)

coarticulation
/kəʊɑːtɪkjʊ'leɪʃən/, **/k@UA:tIkjU'leIS@n/**, [N: coarticulation], [plural: -s]. Hyponyms: anticipatory coarticulation, perseverative coarticulation, left-to-right coarticulation, right-to-left coarticulation. Def.: An articulation of a speech sound which involves simultaneous or overlapping movements of more than one articulator and at more than one point in the vocal tract. (cf. also Crystal 1988, p. 52)

coda
/'kəʊdə/, **/'k@Ud@/**, [N: coda], [plural: -s]. Domain: speech synthesis. Hyperonyms: syllable constituent. Synonyms: margin; slope; trough. Cohyponym: nucleus, crest, peak; onset. Meronym. sup.: syllable. Meronym. sub.: consonant. Def.: The consonant or consonant sequence which occurs after the vowel or vowel-like nucleus in a syllable.

code table
/'kəʊd 'teɪbəl/, **/'k@Ud 'teIb@l/**, [N: [N: code][N: table]], [plural: -s]. Def.: Indexed list of codes, e.g. the characters of an alphabet, and their interpretations.

codebook
/ˈkəʊdbʊk/, **/’k@UdbUk/**, [N: codebook], [plural: -s]. Domain: multimodal systems. Def.: A codebook is a store of codes, often of representations of signal segments, e.g. phones in the acoustic domain or of lip images in the visual domain. A codebook for lip representation is based on diphone clustering (e.g. /bb/, /ba/, /br/). The input text is decomposed into a sequence of diphones and the closest image chosen from the codebook is displayed. Image interpolation techniques smooth the transition between successive images.

Codebook-Excited Linear Predictive coding
/ˈkəʊdbʊk ɪkˈsaɪtɪd ˈlɪneə prɪˈdɪktɪv ˈkəʊdɪŋ/, **/’k@UdbUk Ik’saItId ’lIne@ prI’dIktIv ’k@UdIN/**, [N: [N: Codebook][AJ: Excited][AJ: Linear][AJ: Predictive][N: Coding]], [plural: none]. Synonyms: CELP coding. Def.: A form of speech coding using linear prediction in which the excitation of the linear predictive filter is drawn from a codebook of possibilities.

codec
/ˈkəʊdek/, **/’k@Udek/**, [N: codec], [plural: -s]. Hyperonyms: interface. Synonyms: coder-decoder. Def.: Codecs are interfaces between a computer system and a signal source, implemented either in hardware, e.g. in mobile phones, or in software, e.g. as plug-ins for WWW browsers to encode or decode a signal, e.g. to achieve compression.

coder-decoder
/ˈkəʊdə ˈdiːkəʊdə/, **/’k@Ud@ ’di:k@Ud@/**, [N: [N: coder][N: decoder]], [plural: -s]. Hyperonyms: interface. Synonyms: codec. Def.: Codecs are implemented either in hardware, e.g. in mobile phones, or in software, e.g. as plug-ins for WWW browsers to encode or decode a signal, e.g. to achieve compression.

command and control system
/kəˈmɑːnd ənd kənˈtrəʊl ˈsɪstəm/, **/k@’mA:nd @nd k@n’tr@Ul ’sIst@m/**, [N: [N: command][CONJ: and][N: control][N: system]], [plural: -s]. Domain: consumer off-the-shelf products. Hyperonyms: system. Cohyponym: document generation system. Def.: These systems contain a speech recognition system as interface for controlling the environment of the user. This can be as simple as the graphical shell of the user’s computer, and as complicated as controlling all operational functions of a fast fighter aircraft.

command system
/kəˈmɑːnd ˈsɪstəm/, **/k@’mA:nd ’sIst@m/**, [N: [N: command][N: system]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: dialogue system. Cohyponym: interactive dialogue system. Def.: In command systems, the interaction is direct and deterministic: to one stimulus from one agent corresponds one unique response from the other agent, the response being independent of the state or context of each agent. For example, you press a key on a keyboard and the expected character appears on the screen. With command systems, the human has direct control over the machine. This form, not normally considered a variety of human communication, is usually referred to as the tool metaphor. (Gibbon et al. 1997, p. 569)

common ground unit
/ˈkɒmən ˈgraʊnd ˈjuːnɪt/, **/’kQm@n ’graUnd ’ju:nIt/**, [N: [AJ: common][N: ground][N: unit]], [plural: -s]. Synonyms: CGU. Def.: An abstract category of pragmatic meso-level annotation, comprising all units of speech that are relevant to developing mutual understanding of a topic in a dialogue.

common-password speaker recognition system
/ˈkɒmən ˈpɑːswɜːd ˈspiːkə rekəgˈnɪʃən ˈsɪstəm/, **/’kQm@n ’pA:sw3:d ’spi:k@ rek@g’nIS@n ’sIst@m/**, [N: [AJ: common][N: password][N: speaker][N: recognition][N: system]], [plural: -s]. Domain: speaker recognition. Hyperonyms: text-dependent speaker recognition system. Cohyponym: personal-password speaker recognition system. Def.: A text-dependent speaker recognition system for which all registered speakers have the same voice password.

communicating agent
/kə'mjuːnɪkeɪtɪŋ 'eɪdʒənt/, /k@'mju:nIkeItIN 'eIdZ@nt/, [N: [AJ: communicating][N: agent]], [plural: -s]. Domain: multimodal systems. Synonyms: conversational agent. Def.: Agent capable of being semi-autonomous, taking decisions and conversing with a user. The agent is also able to show emotions and have a personality.

communication media
/kəmjuːnɪ'keɪʃən 'miːdɪə/, /k@mju:nI'keIS@n 'mi:dI@/, [N: [N: communication][N: media]], [plural: none]. Domain: interactive dialogue systems. Hyperonyms: media. Synonyms: communication means. Def.: Materials or devices used by an interactive dialogue system to communicate with the user.

communication mode
/kəmjuːnɪ'keɪʃən 'məʊd/ , /k@mju:nI'keIS@n 'm@Ud/ , [N: [N: communication][N: mode]], [plural: -s] . Domain: interactive dialogue systems. Hyperonyms: perception sense. Hyponyms: vocal communication mode, visual communication mode, auditive communication mode, tactile communication mode, olfactive communication mode. Def.: Perception sense which allows for communication.

communication
/kɒmjuːnɪ'keɪʃən/, /kQmju:nI'keIS@n/, [N: communication], [plural: -s]. Domain: . Hyponyms: tactile communication mode, interaction. Def.: The transfer of a message as a data stream between a sender and a receiver via a channel.

communicative status
/kə'mjuːnɪkətɪv 'steɪtəs/, /k@'mju:nIk@tIv 'steIt@s/, [N: [AJ: communicative][N: status]], [plural: none]. Domain: dialogue annotation. Def.: A pragmatic utterance tag that indicates whether an utterance is intelligible or complete.

Compact-Disk Read Only Memory
/'kɒmpækt 'dɪsk 'riːd 'əʊnli 'meməriː/, /'kQmp{kt 'dIsk 'ri:d '@Unli 'mem@ri:/, [N: [AJ: Compact][N: Disk][V: Read][AV: Only][N: Memory]], [plural: y/-ies]. Hyperonyms: disk. Synonyms: CD-ROM. Cohyponym: Audio CD . Def.: Compact disk for the storage of data, e.g. software, signal data, etc., with different file structure from audio CD-ROMs. The international CD-ROM standard is ISO 9660 (restricted directory depth, short '8.3' file names, restricted filename character set), but more flexible de facto specifications (e.g. Joliet) have been developed for PC and Mac environments, as well as specifications for hybrid audio and data file CD-ROMs.

comparative testing
/kəm'pærətɪv 'testɪŋ/, /k@m'p{r@tIv 'testIN/, [N: [AJ: comparative][N: testing]], [plural: -s] . Domain: speech synthesis. Hyperonyms: assessment technique. Synonyms: benchmarking assessment. Cohyponym: diagnostic testing, diagnostic evaluation, diagnostic assessment. Def.: a) Performance evaluation of a system as a whole. b) Comparative or benchmarking assessment is used to select the best available system, or just to determine the state of the art of the technology.

competence
/'kɒmpɪtəns/, /'kQmpIt@ns/, [N: competence], [plural: -s]. Domain: interactive dialogue systems. Cohyponym: performance. Def.: Competence is the (unconscious) mental knowledge about a certain mother tongue that an 'ideal' speaker/hearer belonging to a homogenous speech community, i.e. a speech community free of dialectal or sociolectal speech variants, has. (Bussmann, p. 396) E.g. Though the competence of an English speaker tells him that the past tense of the English verb 'go' is 'went', a host of factors including fatigue, distraction, or word-play may result in his performance production of the ill-formed *goed..

complementarity
/kɒmplɪmən'tærɪti/, /kQmplIm@n't{rIti/, [N: complementarity], [plural: none]. Domain: multimodal systems. Hyperonyms: cooperation type. Cohyponym: redundancy, equivalence, specialisation, concurrency, transfer. Def.: Different chunks of information belonging to the same command are transmitted over more than one modality. E.g. Saying "put-that-there", while pointing at an object, and then at a location..

component evaluation
/kəm'pəʊnənt ɪvæljʊ'eɪʃən/, /k@m'p@Un@nt Iv{ljU'eIS@n/, [N: [N: component][N: evaluation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: evaluation. Def.: Evaluation of the components of a system such as a multimodal system. Evaluation methodologies that are accepted in the various subfields can be reused, including evaluation of speech recognition, handwriting recognition, and gesture recognition, as well as the evaluation of talking heads. In addition, the quality of the integration of the components in a multimodal system may have to be evaluated, for example, the accuracy of automatically assigning multimodal input to the appropriate (specialised) recognisers.

composite word
/'kɒmpɒsɪt 'wɜːd/ , /'kQmpQsIt 'w3:d/ , [N:[AJ: composite][N: word]], [plural: -s]. Domain: word formation. Hyperonyms: word. Hyponyms: compound word, derived word. Cohyponym: simplex word. Def.: A composite word is either a compound word or a derived word.

compound
/'kɒmpaʊnd/, /'kQmpaUnd/, [N: compound], [plural: -s]. Domain: lexicon. Hyperonyms: word. Synonyms: composite word. Cohyponym: simplex word. Def.: A compound is a word morphologically concatenated with a word or stem and thus contains at least two stems.

compounding
/'kɒmpaʊndɪŋ/, /'kQmpaUndIN/, [N: compounding], [plural: none]. Domain: lexicon. Hyperonyms: morphological operation. Synonyms: composition. Cohyponym: derivation. Meronym. sup.: word formation. Def.: Compounding deals with the construction of words by concatenating words or stems. (Gibbon et al. 1997, p. 214) E.g. wind + mill = windmill.

comprehension test
/kɒmprɪ'henʃən 'test/, /kQmprI'henS@n 'test/, [N: [N: comprehension][N: test]], [plural: -s]. Domain: speech synthesis. Hyperonyms: testing procedure. Hyponyms: off-line comprehension test, on-line comprehension test. Cohyponym: identification test. Def.: Procedure testing a listener's understanding of a speech stimulus at the sentence or text level (often by asking the listener to answer content questions).

compression scheme
/kɒm'preʃən 'skiːm/, /kQm'preS@n 'ski:m/, [N: [N: compression][N: scheme]], [plural: -s]. Hyponyms: MPEG-1, MPEG-2, MPEG-3, MPEG-4, Huffman compression. Def.: A compression scheme is an algorithm for reducing the size of a data packet, for example by searching for and indexing repeated identical sequences.

concatenation technique
/kənkætə'neɪʃən tek'niːk/, /k@nk{t@'neIS@n tek'ni:k/, [N: [N: concatenation][N: technique]], [plural: -s]. Domain: speech synthesis, consumer off-the-shelf products. Hyperonyms: speech synthesis technique. Cohyponym: production model, playback technique. Def.: The technique of concatenating and smoothing pre-defined segments of speech for speech synthesis. By playing back sub-word units of pre-recorded speech contiguously, whole words and phrases can be synthesised. Mostly, the units chosen are diphones, i.e. the period of the last half of the previous phone up to the first half of the next phone. Usually, the voice quality of these systems is high. By using techniques such as PSOLA the pitch of the pre-recorded waveforms can be changed and thus controlled intonation and stress is possible. A genuine change of voice characteristics is not possible. The vocabulary is limited by pronunciation rules.

concatenative synthesis
/kɒn'kætɪnətɪv 'sɪnθəsɪs/, /kQn'k{tIn@tIv 'sInT@sIs/, [N: [AJ: concatenative][N: synthesis]], [plural: concatenative syntheses]. Domain: speech synthesis. Hyperonyms: speech synthesis. Cohyponym: parametric synthesis. Def.: Speech synthesis where samples of predefined segments of speech are concatenated, usually with some smoothing applied so that the boundaries are less audible.

concept-to-speech system
/'kɒnsept 'tə 'spi:tʃ 'sɪstəm/, /'kQnsept 't@ 'spi:tS 'sIst@m/, [N: [N: concept][PREP: to][N: speech][N: system]], [plural: -s]. Domain: speech synthesis. Hyperonyms: speech output system. Synonyms: CTS, meaning to speech, MTS. Cohyponym: text-to-speech system, TTS. Def.: A concept-to-speech system is a speech output system that converts some abstract representation of a communicative intention to speech rather than a text representation.

conceptual appropriateness
/kən'septjʊəl ə'prəʊprɪətnɪs/, /k@n'septjU@l @'pr@UprI@tnIs/, [N: [AJ: conceptual][N: appropriateness]], [plural: none]. Domain: interactive dialogue systems. Hyperonyms: measure. Def.: Conceptual appropriateness is a measure of the appropriateness of a system utterance in its immediate dialogue context. This is a five-valued measure, with values drawn from one set: - TF (total failure) - AP (appropriate) - IA (inappropriate) - AI (appropriate / inappropriate)- IC (incomprehensible). (Gibbon et al. 1997, p. 606/607)

concurrency
/kən'kʌrənsi/, /k@n'kVr@nsi/, [N: concurrency], [plural: none]. Domain: multimodal systems. Hyperonyms: cooperation type. Cohyponym: complementarity, redundancy, equivalence, specialisation, transfer. Def.: Independent chunks of information are transmitted using different modalities and overlap in time. Concurrency means parallel use of different modalities to initiate different actions. E.g. Talking over speaker phone while editing a document..

condenser microphone
/kɒn'densə 'maɪkrəfəʊn/, /kQn'dens@ 'maIkr@f@Un/, [N: [N: condenser][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: microphone. Cohyponym: dynamic microphone. Def.: Condenser microphones basically consist of a capacitor, one of the electrodes of which is formed by a conductive membrane. This membrane is exposed to the incident sound and, when moved back and forth by the sound pressure, slightly changes the capacitance of the capacitor. When the load on the capacitor is kept constant the capacitance changes will, for the voltage across the electrodes, follow the movements of the membrane as long as the voltage changes are small compared to the total voltage across the electrodes. Since the membrane can be manufactured from very thin plastic film material with a conductive layer of vapourised gold or aluminium, it will follow the sound pressure quite exactly and the signal produced by the microphone will be a rather precise reproduction of the original course of the sound pressure. For high-quality studio recordings most microphones used are condenser microphones. (Gibbon et al. 1997, p. 302/303)

confidence interval
/'kɒnfɪdəns 'ɪntəvəl/, /'kQnfId@ns 'Int@v@l/, [N: [N: confidence][N: interval]], [plural: -s]. Domain: assessment methodologies. Hyperonyms: statistical measure. Def.: A statistical measure specifying the proportion of independent estimates of some population value that are likely to fall within this interval.

confidence measure
/'kɒnfɪdəns 'meʒə/, /'kQnfId@ns 'meZ@/, [N: [N: confidence][N: measure]], [plural: -s]. Hyperonyms: value, measure. Def.: A value computed, for instance, by an automatic speech recogniser indicating the degree of belief that the word, phrase or sentence to which the confidence measure was assigned has been recognised correctly.

confidence
/'kɒnfɪdəns/, /'kQnfId@ns/, [N: confidence], [plural: none]. Domain: speaker recognition. Def.: The degree of belief in the correctness of a decision made by a system.

conformant speaker
/kən'fɔːmənt 'spiːkə/, /k@n'f0:m@nt 'spi:k@/, [N: [AJ: conformant][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker. Def.: A speaker who belongs to one of the classes of speakers for a given speaker classification system. (Gibbon et al. 1997, p. 413) E.g. For a spoken language identification system that discriminates between languages spoken in Switzerland, a conform speaker is a speaker who speaks either German, French, Italian or Romansch, but not some other language the system does not expect..

congruence
/'kɒŋgrʊəns/, /'kQNgrU@ns/, [N. congruence], [plural: none]. Domain: lexicon. Synonyms: agreement. Cohyponym: word formation. Def.: Agreement between two or more elements of a sentence with regard to their morpho-syntactic categories (case, person, number, gender). (Bussmann, p. 404) E.g. In English 'He sings a song' there is congruence between subject and predicate with regard to person and number..

conjunction
/kən'dʒʌŋkʃən/, /k@n'dZVNkS@n/, [N: conjunction], [plural: -s]. Domain: lexicon. Hyperonyms: grammatical category. Cohyponym: article, preposition, interjection, pronoun. Def.: A word whose function is to signal a link between syntactic units of equivalent status; conjunctions are normally subdivided into coordinating (and, or, etc.) and subordinating (if, because, etc.) types.

connected word speech recognition system
/kə'nektɪd 'wɜːd 'spiːtʃ rekəg'nɪʃən 'sɪstəm/, /k@'nektId 'w3:d 'spi:tS rek@g'nIS@n 'sIst@m/, [N: [AJ: connected][N: word][N: speech][N: recognition][N: system]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: speech recognition system. Cohyponym: isolated word speech recognition system, continuous speech recognition system. Def.: A connected word recognition system uses isolated words as speech models, but is capable of recognising these words when they are connected as in free running speech.

connected word
/kə'nektɪd 'wɜːd/, /k@'nektId 'w3:d/, [N: [AJ: connected][N: word]], [plural: always plural]. Domain: speech recognition. Hyperonyms: speech style. Cohyponym: isolated word, continuous speech. Def.: A style of speech where the word form a continuous signal, i.e. the words follow each other fluently. The distinction between "connected words" and "continuous speech" is somewhat technical. A connected word recogniser uses words as recognition units, which can be trained in an isolated word mode. Continuous speech is generally associated with large vocabulary recognisers that use phones as recognition units and can be trained with continuous speech.

consonant cluster
/'kɒnsənənt 'klʌstə/, /'kQns@n@nt 'klVst@/, [N: [N: consonant][N: cluster]], [plural: -s]. Hyperonyms: consonant. Def.: Consonant cluster is a term used in the analysis of connected speech to refer to any sequence of adjacent consonants occurring initially or finally in a syllable. (Crystal 1988) E.g. initially: [br-] in 'bread'; finally: [-st] in 'best'. (Crystal 1988, p. 52).

consonant
/'kɒnsənənt/, /'kQns@n@nt/, [N: consonant], [plural: -s]. Hyperonyms: speech sound. Cohyponym: vowel. Def.: A consonant is produced with constriction or blockage of the airflow in the oral cavity.

constitutive dialogue strategy
/kən'stɪtjʊtɪv 'daɪəlɒg 'strætədʒi/, /k@n'stItjUtIv 'daI@lQg 'str{t@dZi/, [N: [AJ: constitutive][N: dialogue][N: strategy]], [plural: y/-ies]. Domain: interactive dialogue systems. Hyperonyms: dialogue strategy. Cohyponym: adaptive dialogue strategy, cooperative dialogue strategy, deterministic dialogue strategy. Def.: The constitutative dialogue strategy implies that (for educational systems) the system has to learn new notions in its normal operation.(Gibbon et al. 1997, p. 598)

context free grammar
/'kɒntekst 'fri: 'græmə/, /'kQntekst 'fri: 'gr{m@/, [N: [N: context][AJ: free][N: grammar]], [plural: -s]. Domain: language modelling. Hyperonyms: Chomsky formal grammar hierarchy. Hyponyms: deterministic context-free grammar, nondeterministic context-free grammar, metalinear grammar, regular grammar. Synonyms: phrase structure grammar. Cohyponym: unrestricted rewrite grammar, context-sensitive grammar. Def.: A context-free grammar is a set of rules which defines constituent structure trees over strings of symbols from a vocabulary V. It is defined formally as a quadrulple <N,T,S,R>, where N is a finite set of nonterminal symbols in V, T is a finite set of terminal symbols in V, S is a start symbol (defining the root of the tree structures) in N, and R is a set of rules of the form XAY -> XGY, where XAY and XGY are strings, X and Y are (possibly zero) strings of symbols from V, A is an element of N, G is a non-zero string of symbols from V. The constant context of context-free rules are commonly factored out and annotated as A -> G / X_Y. Context-free grammars are also known as Type 2 grammars in the Chomsky hierarchy of formal grammars. Context-free grammars are weakly equivalent to and processed by push-down automata. Rules in stochastic context-free grammars are annotated with application probabilities on the basis of corpus analyses. Enhanced context-free grammars such as Definite Clause Grammars permit the annotation of constituent structure trees with feature matrices. (Crystal 1988, p. 71) E.g. Elementary examples of context-free rewrite rules are S -> NP VP, NP -> N, NP -> Det N, V -> V -> V NP, Det -> this, N -> CD, N -> floppy, V -> replaces. A grammar with these rules defines trees over the sentences 'this CD replaces this floppy', 'this floppy replaces this CD', 'this CD replaces this CD', 'this floppy replaces this floppy'..

contextual appropriateness
/kən'tekstjʊəl ə'prəʊprɪətnɪs/, /k@n'tekstjU@l @'pr@UprI@tnIs/, [N: [AJ: contextual][N: appropriateness]], [plural: none]. Domain: interactive dialogue systems. Hyperonyms: measure. Def.: Contextual appropriateness is a measure of the appropriateness of a system utterance in its immediate dialogue context.

contextual fusion
/kən'tekstjʊəl 'fju:ʒən/, /k@n'tekstjU@l 'fju:Z@n/, [N: [AJ: contextual][N: fusion]], [plural: none]. Domain: multimodal systems. Hyperonyms: fusion. Cohyponym: microtemporal fusion, macrotemporal fusion. Def.: Contextual fusion combines information units based on semantic constraints.

continuous speech recognition system
/kən'tɪnjʊəs 'spi:tʃ rekəg'nɪʃən 'sɪstəm/, /k@n'tInjU@s 'spi:tS rek@g'nIS@n 'sIst@m/, [N: [AJ: continuous][N: speech][N: recognition][N: system]], [plural: -s]. Domain: interactive dialogue systems, multimodal systems, speech recognition, consumer off-the-shelf products. Hyperonyms: speech recognition system. Cohyponym: discrete speech recognition system, isolated word recognition system, connected word recognition system. Def.: A speech recognition system that recognises fluent speech, which is more difficult to recognise than isolated words as the end of a word is not easily distinguished from the beginning of the next word. A continuous speech recognition system is trained (possibly in the factory) by continuous speech. Some systems are hybrid, they are word recognition systems, but can cope for instance with continuous digit strings.

continuous speech
/kən'tɪnjʊəs 'spiːtʃ/, /k@n'tInjU@s 'spi:tS/, [N: [AJ: continuous][N: speech]], [plural: -es]. Domain: speech recognition. Hyperonyms: speech style. Def.: A style of speech where the words form a continuous signal, i.e. the words follow each other fluently. The distinction between "connected words" and "continuous speech" is somewhat technical. A connected word recogniser uses words as recognition units, which can be trained in a isolated word mode. Continuous speech is generally associated with large vocabulary recognisers that use phones as recognition units and can be trained with continuous speech.

conversational agent
/kɒnvə'seɪʃənəl 'eɪdʒənt/, /kQnv@'seIS@n@l 'eIdZ@nt/, [N: [AJ: conversational][N: agent]], [plural: -s]. Domain: multimodal systems. Synonyms: communicating agent. Def.: Agent capable of semi-autonomous decision-taking and dialogue with a user. Current research also addresses the issue of developing conversational agents which show emotions and have a personality.

conversational game
/kɒnvə'seɪʃənəl 'geɪm/, /kQnv@'seIS@n@l 'geIm/, [N: [AJ: conversational][N: game]], [plural: -s]. Def.: A category of pragmatic meso-level annotation, encompassing all utterances following an initiating move up to the point where the goal of the initiating move has been fulfilled or abandoned.

cooperation type
/kəʊɒpə'reɪʃən 'taɪp/, /k@UQp@'reIS@n 'taIp/, [N: [N: cooperation][N: type]], [plural: -s]. Domain: multimodal systems. Hyponyms: complementarity, redundancy, equivalence, specialisation, concurrency, transfer. Def.: Way several modalities work together to improve the (human-computer) interaction.

cooperative dialogue strategy
/kəʊ'ɒpərətɪv 'daɪəlɒg 'strætədʒi/, /k@U'Qp@r@tIv 'daI@lQg 'str{t@dZi/, [N: [AJ: cooperative][N: dialogue][N: strategy]], [plural: y/-ies]. Domain: interactive dialogue systems. Hyperonyms: dialogue strategy. Cohyponym: adaptive dialogue strategy, constitutive dialogue strategy, deterministic dialogue strategy. Def.: A cooperative dialogue strategy includes correction and prediction mechanisms, shares initiative with the user, accepts interruptions or negotiation, and is capable of clarifying the system's choices and responses, (turn-taking is balanced between the user and the system). (Gibbon et al. 1997, p. 598)

cooperative speaker
/kəʊ'ɒpərətɪv 'spiːkə/ , /k@U'Qp@r@tIv 'spi:k@/ , [N: [AJ: cooperative][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: registered speaker. Cohyponym: uncooperative speaker. Def.: A cooperative speaker can be defined as an authorised applicant who is willing to be identified or as a genuine speaker who intends to be verified positively. (Gibbon et al. 1997, p. 421)

corpus
/'kɔːpəs/, /'kO:p@s/, [N: corpus], [plural: corpora]. Domain: corpora. Hyponyms: spoken language corpus, written language corpus. Def.: 1. A collection of texts in machine readable form, comprising either written or spoken data or both. 2. In the context of Spoken Language, a body of spoken language data which has been recorded, has been transcribed and/or annotated (in part or in toto) and documented for use in the development of LE systems, and in general is available for use by more than one team in the research and development community.

correction rate
/kə'rekʃən 'reɪt/, /k@'rekS@n 'reIt/, [N: [N: correction][N: rate]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: ratio of turns. Synonyms: CR. Def.: Percentage of all turns in a dialogue which are correction turns.

count
/'kaʊnt/, /'kaUnt/, [N: count], [plural: -s]. Domain: language modelling. Hyponyms: bigram count, trigram count. Def.: Counts are elementary quantifications of categories in data, and are used for example to describe training data. For example, trigram counts are obtained by counting how often a particular word trigram occurs in the training data. (Gibbon et al. 1997, p. 249)

CR
/'siː 'ɑː/, /'si: 'A:/, [N: CR], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: ratio of turns. Synonyms: correction rate. Def.: Percentage of all turns which are correction turns.

creaky voice
/'kriːki 'vɔɪs/, /'kri:ki 'vOIs/, [N: [AJ: creaky][N: voice]], [plural: -s]. Domain: physical characterisation. Def.: A creaky voice results from irregular laryngeal vibrations often with a cycle of 'normal' duration being followed by a cycle of roughly twice the normal duration.

cross-validation
/'krɒs vælɪ'deɪʃən/, /'krQs v{lI'deIS@n/, [N: [AJ: cross][N: validation]], [plural: -s]. Domain: language modelling. Hyperonyms: statistical technique. Hyponyms: leaving-one-out. Meronym. sup.: statistical estimation. Def.: Cross-validation is a technique in statistical estimation by which the parameters of a model are optimised on a new unseen test set. In the context of stochastic language modelling, for example, cross-validation is used to estimate smoothing parameters.

CSLU Toolkit
/'siː 'es 'el 'juː 'tuːlkɪt//, /'si: 'es 'el 'ju: 'tu:lkIt//, [N: CSLU Toolkit] , [plural: none]. Hyperonyms: toolkit. Def.: A software environment for research, development, and education of spoken language systems, developed at the Center for Spoken Language Understanding, Oregon Graduate Institute. It integrates a set of core technologies including speech recognition, speech synthesis, facial animation and speaker recognition. It also features authoring and analysis tools enabling quick and easy development of desktop and telephone-based speech applications.

cued speech
/'kjuːd 'spiːtʃ/, /'kju:d 'spi:tS/, [N: [AJ: cued][N: speech]], [plural: none]. Domain: multimodal systems. Def.: Cued speech is a mode of communication by manual gesture, and differs from sign languages in that handshapes and hand placements are used to signal articulatory movements during speech: one hand is placed close to the lips and changes shape in synchronisation with speech. Cued speech is used by hearing-impaired people.

cut-through
/'kʌt 'θruː/, /'kVt 'Tru:/, [N: [V: cut][PREP: through]], [plural: -s]. Domain: system design. Synonyms: talkover, barge-in. Def.: The system hears and understands simultaneously (single step). (see also definition of 'barge-in')

DAMSL
/'dæmzəl/, /'d{mz@l/, [N: DAMSL], [plural: none]. Domain: dialogue representation/annotation. Hyperonyms: coding scheme. Synonyms: Dialogue Act Mark-up in Several Layers. Def.: DAMSL is a scheme for annotating dialogs. It marks important characteristics of utterances that indicate their role in the dialog and their relationship to each other. The annotation scheme has been defined in order to provide a top-level structure for annotating a range of dialogs for many different purposes.

data glove
/'deɪtə 'glʌv/, /'deIt@ 'glVv/, [N: [N: data][N: glove]], [plural: -s]. Domain: multimodal systems. Hyperonyms: position tracker. Def.: A data glove is a position tracker for hand and finger movements that is worn like a glove.

database management system
/'deɪtəbeɪs 'mænɪdʒmənt 'sɪstəm/, /'deIt@beIs 'm{nIdZm@nt 'sIst@m/, [N: [N: database][N: management][N: system]], [plural: -s]. Hyperonyms: software. Synonyms: DBMS. Meronym. sub.: database, termbank, termbase. Def.: A database management system is a program that allows the secure management and storage of large amounts of data and provides controlled access to this data. E.g. Oracle, MS Access, mSQL, Shoebox.

database
/'deɪtəbeɪs/, /'deIt@beIs/, [N: database], [plural: -s]. Hyponyms: speech database, termbank, termbase. Meronym. sup.: DBMS, database management system. Def.: A collection of data structured according to a data model such as the relational or the object-oriented model, and stored in a database management system (DBMS).

DAVID
/'deɪvɪd/, /'deIvId/, [N: DAVID], [plural: none]. Domain: multimodal systems. Hyperonyms: database. Synonyms: Digital Audio-Visual Integrated Database. Def.: DAVID was developed by the British Telecom Laboratories and the Department of Electrical and Electronic Engineering of the University of Wales in Swansea, UK. The purpose of DAVID is to offer a database for research in speech or person recognition, synthesis of talking heads, facial image segmentation, visual speech feature assessment, and voice control of video-conferencing resources. The database contains material including isolated digits, the English-alphabet E-set, some "VCVCV" nonsense utterances, and some full sentences. Some of the speakers have been recorded over six months. Others had only one recording session. Most recordings were performed with plain background, but some were done in complex scenes. Some of the database elements show both front and profile images of the speaker, others are a frontal and profile close-up view of the speaker's lip only. This last set is useful for assessing automatic lip segmentation systems. The database contains data of about 100 persons.

dB
/'di: 'bi:/, /'di: 'bi:/, [N: dB], [plural: -s]. Domain: physical characterisation. Hyperonyms: acoustic measure. Synonyms: decibel. Def.: A measure, on a log scale, of the difference in power of two acoustic signals: 10log 10 (P1/P2). Thus, 10 dB corresponds to a power difference of a factor of 10, and 20 dB to a factor of 100.

DBMS
/'di: 'bi: 'em 'es/, /'di: 'bi: 'em 'es/, [N: DBMS], [plural: -s]. Hyperonyms: software. Synonyms: DBMS. Meronym. sub.: database, termbank, termbase. Def.: A database management system is a program that allows the secure management and storage of large amounts of data and provides controlled access to this data. E.g. Oracle, MS Access, mSQL, Shoebox.

DDL
/'di: 'di: 'el/, /'di: 'di: 'el/, [N: DDL], [plural: -s]. Synonyms: Dialogue Description Language. Def.: An annotation scheme for dialogue markup whose graphical component is based on SDL (Specification and Description Language). (Gibbon et al. 1997, p.573)

decibel
/'desɪbəl/, /'desIb@l/, [N: decibel], [plural: -s]. Domain: physical characterisation. Hyperonyms: acoustic measure. Synonyms: dB. Def.: A measure, on a log scale, of the difference in power of two acoustic signals: 10log 10 (P1/P2). Thus, 10 dB corresponds to a power difference of a factor of 10, and 20 dB to a factor of 100.

decision outcome
/dɪ'sɪʒən 'aʊtkʌm/, /dI'sIZ@n 'aUtkVm/, [N:[N: decision][N: outcome]], [plural: -s]. Domain: speaker recognition. Hyponyms: identitiy assignment, rejection. Def.: The result of a decision procedure or decision task.

declarative lexicon
/dɪ'klærətɪv 'leksɪkən/, /dI'kl{r@tIv 'leksIk@n/, [N: [AJ: declarative][N: lexicon]], [plural: -s; declarative lexica]. Domain: lexicon. Hyperonyms: lexicon. Cohyponym: procedural lexicon. Def.: A lexicon based on a neutral abstract lemma concept, in which the structure of the lexicon is not dictated by requirements of specific types of lexical access but by general logical principles. (Gibbon et al. 1997, p. 200)

definition
/defɪ'nɪʃən/, /defI'nIS@n/, [N: definition], [plural: -s]. Domain: terminology. Hyperonyms: phrase. Def.: The verbal description of a word (nominal definition) or concept (real definition). The main classical types definition are by genus proximum and differentia specifica (general category and specific differences), by ostension (showing particular cases or examples), and by contextualisation (showing the use of a word in context).

deformable template matching
/dɪ'fɔːməbəl 'templeɪt 'mætʃɪŋ/, /dI'fO:m@b@l 'templeIt 'm{tSIN/, [N: [AJ: deformable][N: template][N: matching]], [plural: none]. Domain: multimodal systems. Hyperonyms: template matching. Cohyponym: Principal Component Analysis, PCA, geometric template matching, optical flow technique, neural network based approach. Def.: Deformable templates are used to model lip shapes and recognise faces. They are constructed based on a priori knowledge about the feature shapes as parameterised curves that can deform during model fitting. The curves follow the outline of the facial features and their final shapes can be used to recognise a particular lip shape or face. When multiple templates are used in the recognition process the results of correct recognition increases; for example, 16% classification accuracy with one template is reported, and 33% accuracy with six templates.

deictic gesture
/'daɪktɪk 'dʒestʃə/, /'daIktIk 'dZestS@/, [N: [AJ: deictic][N: gesture]], [plural: -s]. Domain: Spoken Language Technolgy: multimodal systems. Hyperonyms: gesture; deixis. Cohyponym: iconic gesture, metaphoric gesture, symbolic gesture; deictic word, deictic meaning. Def.: Deictic gestures (pointing gestures) refer to objects or events in the surrounding environment, for example the famous "put-that-there" accompanied by pointing with the mouse or fingers, often accompanying deictic pronouns or deictic adverbs in an utterance, as in this example.

deleted interpolation
/dɪ'liːtɪd ɪntɜːpə'leɪʃən/, /dI'li:tId Int3:p@'leIS@n/, [N: [AJ: deleted][N: interpolation]], [plural: -s]. Domain: language modelling. Hyperonyms: interpolation. Def.: Specific method for smoothing the estimates of the frequency of occurrence of phenomena that do not occur often enough in the training data so as to make straightforward estimates.

deletion
/dɪ'liːʃən/, /dI'li:S@n/, [N: deletion], [plural: -s]. Domain: corpora, lexicon, language modelling, speech synthesis, lexicon. Cohyponym: substitution, insertion. Def.: A word in the utterance that is not recognised.

demisyllable
/demi'sɪləbəl/, /demi'sIl@b@l/, [N: demisyllable], [plural: -s]. Domain: speech synthesis, language modelling, lexicon. Hyperonyms: syllable. Def.: A demisyllable is defined as the interval between the beginning of a syllable and the centre of the nucleus (usually a vowel) of the syllable. Demisyllables are used in some automatic speech recognition systems in order to capture more coarticulation variants than is possible with diphone or triphone based approaches.

dental consonant
/'dentəl 'kɒnsənənt/, /'dent@l 'kQns@n@nt/, [N: [AJ: dental][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, labiodental consonant, alveolar consonant, postalveolar consonant, retroflex consonant, palatal consonant, velar consonant, uvular consonant, pharyngeal consonant, glottal consonant. Def.: A dental consonant is a consonant sound classified phonetically on the basis of its place of articulation, and produced by the tongue tip and rims against the teeth. (cf. also Crystal 1988, p. 88)

dependable speaker
/dɪ'pendəbəl 'spiːkə/, /dI'pend@b@l 'spi:k@/, [N: [AJ: dependable][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: registered speaker. Synonyms: sheep. Cohyponym: unreliable speaker. Def.: A registered speaker with a low misclassification rate. (Gibbon et al. 1997, p. 432)

derivation
/derɪ'veɪʃən/, /derI'veIS@n/, [N: derivation], [plural: -s]. Domain: lexicon. Hyperonyms: morphological operation. Cohyponym: compounding. Meronym. sup.: word formation. Def.: 1. Derivation is a branch of morphology which deals with the construction of words by the concatenation of stems with affixes. (Gibbon et al. 1997, p. 214) 2. A derivation is a word formed by the morphological process of derivation. 3. A derivation is a chain of inferences based on formal logical or grammatical rules.

derivational affixation
/derɪ'veɪʃənəl æfɪk'seɪʃən/, /derI'veIS@n@l {fIk'seIS@n/, [N: [AJ: derivational][N: affixation]], [plural: -s]. Domain: lexicon. Hyperonyms: affixation. Hyponyms: derivational prefixation, derivational suffixation. Cohyponym: inflectional affixation. Def.: Morphological concatenation of a stem with a derivational affix. (Gibbon et al. 1997, p. 215) E.g. English 'algorithm' + 'ic', 'algorithm' + 'ic' + 'al' + 'ly', 'non' + 'algorithm' + 'ic' + 'al' + 'ly', etc. (Gibbon et al. 1997, p. 215).

deterministic dialogue strategy
/dɪtɜːmɪ'nɪstɪk 'daɪəlɒg 'strætədʒi/, /dIt3:mI'nIstIk 'daI@lQg 'str{t@dZi/, [N: [AJ: deterministic][N: dialogue][N: strategy]], [plural: y/-ies]. Domain: interactive dialogue systems. Hyperonyms: dialogue strategy. Cohyponym: adaptive dialogue strategy, constitutive dialogue strategy, cooperative dialogue strategy. Def.: A dialogue strategy with fully determined decision sequences, in which no initiative is left to the user. Interactive Voice Response (IVR) systems typically fall into this category. (Gibbon et al. 1997, p. 598)

DFT
/'diː 'ef 'tiː/, /'di: 'ef 'ti:/, [N: DFT], [plural: -s]. Domain: signal processing. Hyperonyms: Fourier transform. Hyponyms: FFT, Fast Fourier Transform. Synonyms: Discrete Fourier Transform. Cohyponym: Continuous Fourier Transform. Def.: A transformation of a sampled digital signal into the frequency domain in order to reveal the amplitude or energy of its component frequencies in the form of a spectrogram or sonagram. For speech signals this means that the signal is transformed from the time domain (signal amplitude as a function of time) to the frequency domain (amplitude of signal components as a function of frequency) and, strictly speaking, also to the phase domain. The DFT is essentially a representation of the correlations of the time-domain signal with a finite set of simple (pure sine-wave) time-domain signals in a given range of frequencies.

diagnostic evaluation
/daɪə'gnɒstɪk ɪvæljʊ'eɪʃən/, /daI@'gnQstIk Iv{ljU'eIS@n/, [N: [N: diagnostic][N: evaluation]], [plural: -s]. Domain: multimodal systems, consumer off-the-shelf products, speech recognition. Hyperonyms: assessment technique. Cohyponym: performance evaluation, adequacy evaluation, comparative assessment, benchmarking assessment. Def.: Diagnostic evaluations obtain a profile of system performance with respect to some taxonomy of possible uses of a system. It requires the specification of an appropriate test suite. It is typically used by system developers. Diagnostic assessment involves setting up a framework for testing the product with the intention of giving feedback to the developer, hopefully resulting in an improved system.

dialect
/'daɪəlekt/, /'daI@lekt/, [N: dialect], [plural: -s]. Hyperonyms: language variety. Cohyponym: sociolect, style, functional variant, register. Meronym. sup.: natural language. Def.: Dialect refers to a regionally or distinctive variety of a language, identified by a particular set of words and grammatical structures. (Crystal 1988)

dialogue act
/'daɪəlɒg 'ækt/, /'daI@lQg '{kt/, [N: [N: dialogue][N: act]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: speech act. Def.: A dialogue act is a simplification of the notion of speech act in terms of domain and task specific questions, confirmations, statements, etc., determined by the particular needs of a given dialogue application.

dialogue control particle
/'daɪəlɔːg kən'trəʊl 'pɑːtɪkəl/, /'daI@lO:g k@n'tr@Ul 'pA:tIk@l/, [N: [N: dialogue][N: control][N: particle]], [plural: -s]. Domain: lexicon. Hyperonyms: lexical unit. Def.: A discourse particle used in human dialogue for influencing turn-taking procedures. E.g. er, uhm, aha.

dialogue control
/'daɪəlɒg kən'trəʊl/, /'daI@lQg k@n'tr@Ul/, [N: [N: dialogue][N: control]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Def.: Dialogue control is necessary to provide for a fully automated service. Dialogue control is responsible for the interaction between the user and the service. It must handle events triggered by the user, but also triggers the user to provide the system with information. It must send requests to the information retrieval engines and instruct the text-to-speech engines.

dialogue corpus
/'daɪəlɒg 'kɔːpəs/, /'daI@lQg 'kO:p@s/, [N: [N: dialogue][N: corpus]], [plural: dialogue corpora]. Domain: corpora. Def.: A spoken language corpus consisting of a collection of dialogue recordings, transcriptions and annotations.

dialogue duration
/'daɪəlɒg djuː'reɪʃən/, /'daI@lQg dju:'reIS@n/, [N: [N: dialogue][N: duration]], [plural: none]. Domain: interactive dialogue systems. Hyperonyms: measure. Def.: Dialogue duration is a measure applied to a dialogue corpus.

dialogue grammar
/'daɪəlɒg 'græmə/, /'daI@lQg 'gr{m@/, [N: [N: dialogue][N: grammar]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: grammar. Def.: A grammar for describing a set of well-formed dialogues. The terminal symbols in a dialogue grammar are speech act or dialogue act labels (though for convenience these labels may also be treated as the start symbol for more conventional sentences or utterance grammars). A dialogue grammar might, for example, contain a rule which says that a simple information request consists of two turns, the first of which is a question, and the second of which is an answer. The philosophical roots of dialogue grammars lie in the field of discourse analysis.

dialogue history
/ˈdaɪəlɒg ˈhɪstəri/ , /'daI@lQg 'hIst@ri/ , [N: [N: dialogue][N: history]], [plural: y/-ies]. Domain: interactive dialogue systems. Hyperonyms: record. Def.: A system-internal record of what has happened in a dialogue so far. The dialogue history provides the immediate context within which interpretation takes place.

dialogue manager
/ˈdaɪəlɒg ˈmænɪdʒə/, /'daI@lQg 'm{nIdZ@/, [N: [N: dialogue][N: manager]], [plural: -s]. Domain: interactive dialogue systems, system design. Meronym. sup.: interactive dialogue system. Def.: The component in an interactive dialogue system which is responsible for maintaining dialogue coherence. Functions typically undertaken by a dialogue manager include the following: 1. maintaining a model of the current dialogue context; 2. interpreting utterances in context; 3. linking interpretations to actions; 4. thinking of something to say next; 5. generating topdown predictions of the next user utterance; 6. keeping track of who knows what; 7. generating utterances which are cooperative; 8. selecting an appropriate dialogue strategy; 9. recovering from dialogue breakdowns.

dialogue participant
/ˈdaɪəlɒg pəˈtɪsɪpənt/, /'daI@lQg p@'tIsIp@nt/, [N: [N: dialogue][N: participant]], [plural: -s]. Meronym. sup.: dialogue role. Def.: Each of the participants involved in a dialogue - those speaking and those listening - is a dialogue participant. Participant roles include speaker (first person), addressee (second person), and listener, eavesdropper (third person).

dialogue system
/ˈdaɪəlɒg ˈsɪstəm/, /'daI@lQg 'sIst@m/, [N: [N: dialogue][N: system]], [plural: -s]. Def.: A dialogue system is an interface between a human being and an application system which may include several other systems. The dialogue system processes two kinds of information, from the user and from the task itself, through specialised interfaces, one for the speech technologies, one for the application. A central role of a is to maintain coherence between user and system. A dialogue system whose reactions depend both on environmental and internal conditions is often referred to as an agent or advisor system. The components of a dialogue system include a recogniser, a parser, an interpretation module, a domain model, a partner model, a dialogue manager, a planner, a formulator and a sythesiser. Each of the modules requires associated knowledge databases (lexica, rules and models concerning the language used, the system, the task, the user, the environment, the dialogue itself). An important dynamic component is the dialogue history which keeps track of system state changes. (Gibbon et al. 1997, p. 570)

dialogue
/ˈdaɪəlɒg/, /'daI@lQg/, [N: dialogue], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: interaction. Hyponyms: task-oriented dialogue; system-driven dialogue, system-led dialogue. Def.: A type of discourse taking place between two or more human participants or between human participants and a computer.

dictation speech
/dɪkˈteɪʃən ˈspiːtʃ/, /dIk'teIS@n 'spi:tS/, [N: [N: dictation][N: speech]], [plural: none]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: speaking style. Synonyms: dictation style speech. Cohyponym: spontaneous speech, read speech. Def.: The way a skilled user of an automatic dictation system speaks to the computer: intonation is similar to that in read speech, but grammatical constructions and error corrections are more like in spontaneous speech.

dictation style speech
/dɪkˈteɪʃən ˈstaɪl ˈspiːtʃ/, /dIk'teIS@n 'staIl 'spi:tS/, [N: [N: dictation][N: style][N: speech]], [plural: none]. Domain: speech recognition, consumer off-the-shelf product. Hyperonyms: speaking style. Synonyms: dictation speech. Cohyponym: spontaneous speech, read speech. Def.: The way a skilled user of an automatic dictation system speaks to the computer: intonation is similar to that in read speech, but grammatical constructions and error corrections are more like in spontaneous speech.

dictionary
/'dɪkʃənəri/, /'dIkS@n@ri/, [N: dictionary], [plural: y/-ies]. Synonyms: lexicon. Cohyponym: grammar. Def.: 1. A book containing a list of lexical entries (usually words) and their properties. 2. A component of a spoken or written language processing system containing a database with the words used in the system and properties such as pronunciation, part of speech, meaning, which are relevant for processing by the system.

diffuse-field equalised headphone
/dɪ'fjuːs 'fiːld 'iːkwəlaɪzd 'hedfəʊn/ , /dI'fju:s 'fi:ld 'i:kw@laIzd 'hedf@Un/ , [N: [AJ: diffuse][N: field][AJ: equalised][N: headphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: headphone. Cohyponym: free-field equalised headphone. Def.: A diffuse-field equalised headphone, when fed with white noise, produces the same spectral distribution of sound at the ear drum of the listener as appears in a diffuse field. In a diffuse sound field the direction of incidence is evenly distributed over all directions (e.g. in a reverberation chamber). (Gibbon et al. 1997, p. 325)

Digital Audio-Visual Integrated Database
/'dɪdʒɪtəl 'ɔːdɪəʊ 'vɪʒʊəl 'ɪntɪgreɪtɪd 'deɪtəbeɪs/, /'dIdZIt@l 'O:dI@U 'vIZU@l 'IntIgreItId 'deIt@beIs/, [N: [AJ: Digital][AJ: Audio-Visual][AJ: Integrated][N: Database]], [plural: none]. Domain: multimodal systems. Hyperonyms: database. Synonyms: DAVID. Def.: DAVID was developed by the British Telecom Laboratories and the Department of Electrical and Electronic Engineering of the University of Wales in Swansea, UK. The purpose of DAVID is to offer a database for research in speech or person recognition, synthesis of talking heads, facial image segmentation, visual speech feature assessment, and voice control of video-conferencing resources. The database contains material including isolated digits, the English alphabet E-set, some "VCVCV" nonsense utterances, and some full sentences. Some of the speakers have been recorded over six months. Others had only one recording session. Most recordings were performed with plain background, but some were done in complex scenes. Some of the database elements show both front and profile images of the speaker, others are a frontal and profile close-up view of the speaker's lip only. This last set is useful for assessing automatic lip segmentation systems. The database contains data of about 100 persons.

digital signal processor
/'dɪdʒɪtəl 'sɪgnəl 'prəʊsesə/, /'dIdZIt@l 'sIgn@l 'pr@Uses@/, [N: [AJ: digital][N: signal][N: processor]], [plural: -s]. Domain: physical characterisation. Hyperonyms: signal processor. Synonyms: DSP. Def.: A processor whose instruction set and memory configuration are designed to be well suited to efficient processing of operations commonly used in digital signal processing such as FFT or filtering.

digital versatile disk
/'dɪdʒɪtəl 'vɜːsətaɪl 'dɪsk/, /'dIdZIt@l 'v3:s@taIl 'dIsk/, [N: [AJ: digital][AJ: versatile][N: disk]], [plural: -s]. Hyperonyms: disk. Synonyms: DVD. Def.: A successor to the CD with up to 18.4 GB of capacity. Data is stored in up to two layers on both sides of a disk. DVD was originally devised for entertainment purposes (full size video films) and thus has the same structural problems as CDs (helical track, constant angular velocity, i.e. variable disk rotation speed). For entertainment media content, DVDs can be marked with a country code that allows this medium to be played only in a region with the correct code. DVD is backward compatible so that DVD drives can read DVD, DVD-ROM, and traditional CD-ROMs.

diphone
/'daɪfəʊn/, /'daIf@Un/, [N: diphone], [plural: -s]. Hyperonyms: transition unit. Synonyms: phone bigram. Cohyponym: triphone. Def.: A diphone is a segment of an utterance from the centre of phone to the centre of the immediately following phone in an utterance, and is used as one of the simplest possible strategies for capturing transition and coarticulation phenomena between phones. (Gibbon et al. 1997, p. 93)

diphthong
/'dɪfθɒŋ/, /'dIfTQN/, [N: diphthong], [plural: -s]. Hyperonyms: vowel. Def.: A diphthong is a vowel sound in a single syllable with one vowel quality in the first part and another in the second. Depending on the analysis criteria, it may be analysed as one phoneme or as two phonemes. E.g. /aU/, /aI/, /U@/, ...

discordant speaker
/dɪs'kɔːdənt 'spiːkə/, /dIs'kO:d@nt 'spi:k@/, [N: [AJ: discordant][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker. Def.: In speaker classification: A speaker whose real identity is different from his claimed identity. (Gibbon et al. 1997, p. 413) E.g. In age verification: a child claiming that he is an adult..

discounting
/'dɪskaʊntɪŋ/, /'dIskaUntIN/, [N: discounting], [plural: none]. Domain: language modelling. Hyperonyms: smoothing technique. Hyponyms: linear discounting, absolute discounting. Cohyponym: (linear) interpolation. Meronym. sup.: language modelling. Def.: Discounting is a technique in the context of language model smoothing by which the relative frequencies are discounted to allow for unseen events.

discourse analysis
/'dɪskɔːs ə'næləsɪs/, /'dIskO:s @'n{l@sIs/, [N: [N: discourse][N: analysis]], [plural: discourse analyses]. Domain: corpora. Meronym. sup.: linguistics. Def.: 1. The original approach to the linguistic analysis of naturally occurring connected spoken interaction pioneered by Sinclair and his associates at the University of Birmingham. 2. Subsequently used more generally to include systematic approaches to conversation analysis, an ethnomethodological (sociological) approach to the analysis of spoken interaction.

discourse function
/'dɪskɔːs 'fʌŋkʃən/, /'dIskO:s 'fVNkS@n/, [N: [N: discourse][N: function]], [plural: -s]. Domain: dialogue representation. Def.: The role played by a unit of language in discourse, including take-up, backchannel communication, framing, hesitation and repair marking, answering, checking, modal (possibility, probability, attitudinal) meanings.

discourse marker
/'dɪskɔːs 'mɑːkə/, /'dIskO:s 'mA:k@/, [N: [N: discourse][N: marker]], [plural: -s]. Domain: dialogue representation. Def.: A word (or fixed phrase) which is loosely attached to a larger structure in a stretch of speech and which has a discoursally defined role such as indicating a change in the direction of the discourse, or signalling the speaker's stance towards what has been said. E.g. 'well', 'right'.

discourse particle
/'dɪskɔːs 'pɑːtɪkəl/, /'dIskO:s 'pA:tIk@l/, [N: [N: discourse][N: particle]], [plural: -s]. Domain: lexicon. Hyperonyms: lexical unit. Def.: A short word-like segment of discourse, often with non-typical phonotactic structure, used to signal hesitation or attitudinal feedback, or exercise control over turn-taking. E.g. um, hmm, mm, etc..

Discourse Resource Initiative
/'dɪskɔːs rɪ'sɔːs ɪ'nɪʃətɪv/, /'dIskO:s rI'sO:s I'nIS@tIv/, [N: [N: discourse][N: resource][N: initiative]], [plural: none]. Synonyms: DRI. Def.: A group of researchers working in the area of dialogue annotation with the goal of creating standards in dialogue annotation based on consensus ideas reached at annual workshops.

Discrete Fourier Transform
/dɪs'kriːt 'fʊrɪeɪ 'trænsfɔːm/, **/dIs'kri:t 'fUrIeI 'tr{nsfO:m/**, [N: [AJ: Discrete][N: Fourier][N: Transform]], [plural: -s]. Domain: signal processing. Hyperonyms: Fourier transform. Hyponyms: FFT, Fast Fourier Transform. Synonyms: Discrete Fourier Transform. Cohyponym: Continuous Fourier Transform. Def.: A transformation of a sampled digital signal into the frequency domain in order to reveal the amplitude or energy of its component frequencies in the form of a spectrogram or sonagram. For speech signals this means that the signal is transformed from the time domain (signal amplitude as a function of time) to the frequency domain (amplitude of signal components as a function of frequency) and, strictly speaking, also to the phase domain. The DFT is essentially a representation of the correlations of the time-domain signal with elements of a finite series of simple (pure sine-wave) time-domain signals with a given range of frequencies.

discrete speech recognition system
/dɪs'kriːt 'spiːtʃ rekəg'nɪʃən 'sɪstəm/, **/dIs'kri:t 'spi:tS rek@g'nIS@n 'sIst@m/**, [N: [AJ: discrete][N: speech][N: recognition][N: system]], [plural: -s]. Domain: speech recognition, multimodal systems. Hyperonyms: speech recognition system. Synonyms: isolated word speech recognition system. Cohyponym: continuous speech recognition system. Def.: Speech recogniser which recognises individual words, i.e. the speaker has to separate each word by a pause which makes it easier for the system to recognise the enunciated word.

discriminative training
/dɪs'krɪmɪnətɪv 'treɪnɪŋ/, **/dIs'krImIn@tIv 'treInIN/**, [N: [AJ: discriminative][N: training]], [plural: none]. Hyperonyms: training. Hyponyms: minimal error training, maximum mutual information training. Def.: Training of a pattern recognition system, such as a speech recogniser, by using a training algorithm in which the discrimination between adjacent categories is used as the optimisation criterion. Thus, examples from the class being modelled as well as from other classes are taken account of.

distortion
/dɪs'tɔːʃən/, **/dIs'tO:S@n/**, [N: distortion], [plural: -s]. Domain: physical characterisation. Hyponyms: linear distortion, non-linear distortion. Def.: Distortion is a measure of the non-linearity of a system transfer function, e.g. a transmisssion line, an amplifier, a digital signal processor.

document generation system
/'dɒkjʊmənt dʒenə'reɪʃən 'sɪstəm/, **/'dQkjUm@nt dZen@'reIS@n 'sIst@m/**, [N: [N: document][N: generation][N: system]], [plural: -s]. Domain: consumer off-the-shelf products. Hyperonyms: system. Cohyponym: command and control system. Def.: A system which permits the automatic generation of documents such as manuals and user guides from a database. Some semi-automatic document generation systems incoporate dictation and readback I/O in order to allow fast and flexible generation of documents, forms and reports.

document type definition
/'dɒkjʊmənt 'taɪp defɪ'nɪʃən/, **/'dQkjUm@nt 'taIp defI'nIS@n/**, [N: [N: document][N: type][N: definition]], [plural: -s]. Hyperonyms: SGML, Standard Generalized Markup Language. Synonyms: DTD. Def.: An SGML or XML template (a form of grammar) outlining and constraining the structure of SGML or XML documents. E.g. The definition of HTML is an HTML DTD; the Text Encoding Initiative (TEI) has developed DTDs for the description of a wide variety of document types..

domain
/də'meɪn/, **/d@'meIn/**, [N: domain], [plural: -s]. Def.: The content and task area of language usage for which a recognition system is designed to be used, a (possibly ill-defined) subset of general activity (such as business, avionics, aeronautics, medicine, transport, etc.) in which some coherent collection of tasks may be carried out. A variable defining the type of dialogue according to its subject-matter, e.g. travel, transport, appointment scheduling, etc.

doubleton event
/'dʌbəltən ɪ'vent/, /'dVb@lt@n I'vent/, [N: [N: doubleton][N: event]], [plural: -s]. Domain: language modelling. Hyperonyms: event. Cohyponym: singleton event, unseen event. Def.: Event that was observed exactly twice. (Gibbon et al. 1997, p. 249)

DRI
/'di: 'ɑ:r 'aɪ/, /'di: 'A:r 'aI/, [N: DRI], [plural: none]. Synonyms: Discourse Resource Initiative. Def.: A group of researchers working in the area of dialogue annotation that is attempting to create standards in dialogue annotation based on consensus ideas reached at annual workshops.

DSP
/'di: 'es 'pi:/, /'di: 'es 'pi:/, [N: DSP], [plural: -s]. Domain: physical characterisation. Hyperonyms: signal processor. Synonyms: digital signal processor. Def.: A processor whose instruction set and memory configuration are designed to be well suited to efficient processing of operations commonly used in digital signal processing such as FFT or filtering.

DTD
/'di: 'ti: 'di:/, /'di: 'ti: 'di:/, [N: DTD], [plural: -s]. Synonyms: document type definition. Def.: An SGML or XML template (a form of grammar) outlining and constraining the structure of SGML or XML documents. E.g. The definition of HTML is an HTML DTD; the Text Encoding Initiative (TEI) h as developed DTDs for the description of a wide variety of document types..

DTMF
/'di: 'ti: 'em 'ef/, /'di: 'ti: 'em 'ef/, [N: DTMF], [plural: -s]. Domain: system design. Synonyms: dual tone multi frequency. Def.: The system of pairs of tones used to signal key presses in touch-tone telephone dialling.

dual tone multi frequency
/'dju:əl 'təʊn 'mʌlti 'fri:kwənsi/, /'dju:@l 't@Un 'mVlti 'fri:kw@nsi/, [N: [AJ: dual][N: tone][AJ: multi][N: frequency]], [plural: y/-ies]. Domain: system design. Synonyms: DTMF. Def.: The system of pairs of tones used to signal key presses in touch-tone telephone dialling.

duplex
/'dju:pleks/, /'dju:pleks/, [N: duplex], [plural: duplexes]. Domain: system design. Hyperonyms: property of a communication system. Synonyms: full duplex. Cohyponym: half duplex. Def.: A property of a communication system allowing simultaneous transmission of signals in both directions.

duration
/dju:reɪʃən/, /dju:reIS@n/, [N: duration], [plural: none]. Domain: . Hyponyms: segment duration. Def.: The temporal interval during which a property of a speech signal occurs.

DVD
/'di: 'vi: 'di:/, /'di: 'vi: 'di:/, [N: DVD], [plural: -s]. Hyperonyms: disk. Synonyms: digital versatile disk. Def.: Successor to the CD with up to 18.4 GB of capacity. Data is stored in up to two layers on both sides of a disk. DVD was originally devised for entertainment purposes (full size video films) and thus has the same structural problems as CDs (helical track, constant angular velocity, i.e. variable disk rotation speed). For entertainment media content, DVDs can be marked with a country code that allows this medium to be played only in a region with the correct code. DVD is backward compatible so that DVD drives can read DVD, DVD-ROM, and traditional CD-ROMs.

dynamic microphone
/daɪˈnæmɪk ˈmaɪkrəfəʊn/, /daI'n{mIk 'maIkr@f@Un/, [N: [AJ: dynamic][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: microphone. Cohyponym: condenser microphone. Def.: Dynamic microphones use a constant magnetic field to induce voltage in a moving coil mechanically coupled to the diaphragm. Since the output voltage of the microphone is directly generated by the conversion process, no external power supplies are required. Dynamic microphones are quite robust and may be exposed even to high sound pressure levels, which makes them suited for close-talking applications, for example in headsets. The major disadvantage of the dynamic operation principle is that in addition to the diaphragm the comparably heavy moving coil also has to be moved by the sound pressure, resulting in a poorer transient response of the microphone. For this reason dynamic microphones are, with some exceptions, rarely used as top quality studio microphones. (Gibbon et al. 1997, p. 302)

dysfluency
/dɪsˈfluːənsi/, /dIs'flu:@nsi/, [N: dysfluency], [plural: y/-ies]. Domain: corpora. Synonyms: disfluency. Def.: Generally unintended and inconspicuous break in continuous speech production that occurs in spoken language as a result of on-line production pressures, such as hesitation, false starts, interruption, some kinds of repetition, and filled pauses; sometimes used strategically to influence the addressee.

dysfluent repetition
/dɪsˈfluːənt repəˈtɪʃən/, /dIs'flu:@nt rep@'tIS@n/, [N: [AJ: dysfluent][N: repetition]], [plural: -s]. Def.: A dysfluency phenomenon that consists of the speaker repeating part of an utterance.

earcon
/ˈiɪəkɒn/, /'i:@kQn/, [N: earcon], [plural: -s]. Domain: multimodal systems. Cohyponym: auditory icons, visual icons. Def.: Earcons are defined as abstract (i.e. stylised, not directly imitative) sounds for signalling; they are short-lasting sound samples of a stylised or caricatural nature (in comparison with everyday sound events), by analogy with (visual) icons.

early integration
/ˈɜːli ɪntɪˈgreɪʃən/, /'3:li IntI'greIS@n/, [N: [AJ: early][N: integration]], [plural: -s]. Domain: multimodal systems. Cohyponym: late integration. Def.: Integration of audio and visual information in HMMs where recognition is done using the combination of both signals.

echo cancellation
/ˈekəʊ kænsəˈleɪʃən/, /'ek@U k{ns@'leIS@n/, [N: [N: echo][N: cancellation]], [plural: -s]. Domain: system design. Def.: In a duplex audio system, a procedure to eliminate the electric and acoustic feedback of the speech output of a system that might be present on the return input channel. Echo cancellation is needed in order to allow barge-in and talk through.

echo suppression
/ˈekəʊ səˈpreʃən/, /'ek@U s@'preS@n/, [N: [N: echo][N: suppression]], [plural: -s]. Def.: A procedure to avoid echo to occur by switching off the input of a hands free phone (speaker phone) while the other party is speaking. The return channel is opened only if speech input is detected. Echo suppression and its attendant delayed switching may cause the first syllable of the user's input to be discarded. In command and control applications where commands may be monosyllabic echo suppression may cause the loss of complete commands.

echo
/ˈekəʊ/, /'ek@U/, [N: echo], [plural: -es]. Domain: system design, physical characterisation. Def.: The time-delayed reflection of radiated acoustic or electric signals which is superimposed on signals radiated later. Echo tends to blur the temporal structure of speech utterances, thereby degrading its intelligibility. Electrical echo in analogue telephone networks occurs due to an impedance mismatch at forks where two wire and four wire connections are joined. In spoken dialogue systems echo of the system's output will interfere with user input, to such an extent that it can disable a full duplex connection.

electrical characteristic
/ɪ'lektrɪkəl kærəktə'rɪstɪk/, /I'lektrIk@l k{r@kt@'rIstIk/, [N: [AJ: electrical][N: characteristic]], [plural: -s]. Domain: physical characterisation, speech recognition. Meronym. sup.: physical characterisation. Def.: An electrical characteristic is a parameter (post-production factor) in the physical characterisation of the properties of a speech signal.

embedded programming language
/em'bedɪd 'prəʊgræmɪŋ 'læŋgwɪdʒ/, /em'bedId 'pr@Ugr{mIN 'l{NgwIdZ/, [N: [AJ: embedded][N: programming][N: language]], [plural: -s]. Hyperonyms: programming language. Cohyponym: script language. Def.: Embedded (programming) languages run inside an application, e.g. languages such as Basic which are often used to program macros in order to automatise complex sequences of user operations in word processors and PC database systems, or such as Java or JavaScript in a web browser. Embedded languages may be integrated into an application, or additional modules sometimes called add-ons or plug-ins).

EMU
/'iːmjuː/, /'i:mju:/, [N: EMU], [plural: none]. Hyperonyms: software. Def.: EMU is a collection of software for the creation, manipulation and analysis of speech databases. At the core of EMU is a database search engine which allows the researcher to find various speech segments based on the sequential and hierarchical structure of the utterances in which they occur. EMU includes an interactive labeller which can display spectrograms and other speech waveforms, and which allows the creation of hierarchical, as well as sequential, labels for a speech utterance.

environment
/en'vaɪrənmənt/, /en'vaIr@nm@nt/, [N: environment], [plural: -s]. Hyperonyms: context of recognition. Def.: 1. The environment is the total context in which a system such as a recognition or interactive dialogue system is located. For example, a dashboard control system operates in an in-car environment. Environments may be characterised in many different ways. Most commonly, however, factors which might affect the performance of the system (such as high background noise) are singled out to describe environments. 2. The software and hardware within which a application is developed or used.

equivalence
/ɪ'kwɪvələns/, /I'kwIv@l@ns/, [N: equivalence], [plural: none]. Domain: multimodal systems. Hyperonyms: cooperation type. Cohyponym: complementarity, redundancy, specialisation, concurrency, transfer. Def.: 1. Formally, a relation between two expressions with the same meaning or function. 2. A relation between two or more modalities transmitting the same chunk of information , for instance an option to choose from a menu by either mouse or voice selection.

error rate
/'erə 'reɪt/, /'er@ 'reIt/, [N: [N: error][N: rate]], [plural: -s]. Domain: speech recognition, speaker recognition. Hyperonyms: ratio. Hyponyms: word error rate, sentence error rate. Meronym. sup.: statistics. Def.: The fraction of errors made by a recognition system, i.e. the number of errors divided by the number of words to be recognised. Often expressed as a percentage.

error recovery
/'erə rɪ'kʌvəri/, /'er@ rI'kVv@ri/, [N: [N: error][N: recovery]], [plural: y/-ies]. Domain: speech synthesis, speech recognition, consumer off-the-shelf products. Hyperonyms: performance measure. Cohyponym: recognition accuracy, OOV-rejection, response time, situational awareness. Def.: Error recovery is property of a system based on a procedure by which a system returns to a defined state after the occurrence of an error, either directly or by undoing previous actions (user or system backtracking).

ESPS/waves+
/ˈiː ˈes ˈpiː ˈes ˈweɪvz ˈplʌs/, /'i: 'es 'pi: 'es 'weIvz 'plVs/, [N: ESPS/waves+], [plural: none]. Hyperonyms: software. Def.: ESPS/waves+ is widely used high end commercial suite of programs used for the analysis and display of speech signal data. It includes a library of signal processing programs to assist in computing spectra, analysing speech, converting data, and applying time-referenced labels, and synchronised windows for display of parallel data representations.

estimator
/ˈestɪmeɪtə/, /'estImeIt@/, [N: estimator], [plural: -s]. Domain: speech recognition, assessment methodologies. Hyperonyms: statistical value. Def.: A mathematical expression for estimating the value of a statistical property, such as the mean or the variance of a set of numerical data values.

evaluation methodology
/ɪvæljʊˈeɪʃən meθəˈdɒlədʒi/, /Iv{ljU'eIS@n meT@'dQl@dZi/, [N: [N: evaluation][N: methodology]], [plural: y/-ies]. Domain: multimodal systems. Hyponyms: user-based evaluation, theory-based evaluation, expert-based evaluation, benchmarking. Def.: A set of well-defined methods for quantifying the properties of a system in terms of the system requirements and design specifications.

evaluation
/ɪvæljʊˈeɪʃən/, /Iv{ljU'eIS@n/, [N: evaluation], [plural: -s]. Hyperonyms: measuring. Cohyponym: validation. Def.: A process of measuring whether a given resource or the results of a given activity fit the requirements or design specifications.

event
/ɪˈvent/, /I'vent/, [N: event], [plural: -s]. Domain: language modelling. Hyponyms: unseen event, singleton event, doubleton event. Def.: 1. A state or change of state of a system, or a value or change of value of a parameter, at a point in time or over an interval in time. 2. In graphical user interface (GUI) design, one of a set of specified events such as mouse movements or keyboard inputs which may be used to influence the behaviour of the system.

event-dependent speaker recognition system
/ɪˈvent dɪˈpendənt ˈspiːkə rekəgˈnɪʃən ˈsɪstəm/, /I'vent dI'pend@nt 'spi:k@ rek@g'nIS@n 'sIst@m/, [N: [N: event][AJ: dependent][N: speaker][N: recognition][N: system]], [plural: -s]. Domain: speaker recognition. Hyperonyms: text-independent speaker recognition system. Hyponyms: keyword spotting, concept spotting. Cohyponym: text-dependent speaker recognition system. Def.: A text-independent speaker recognition system for which test utterances must contain a certain linguistic event (or class of events) while the rest of the acoustic material is discarded. This approach requires a preliminary step for spotting and localising the relevant events.

exception vocabulary size
/ekˈsepʃən vəˈkæbjʊləri ˈsaɪz/, /ek'sepS@n v@'k{bjUl@ri 'saIz/, [N: [N: exception][N: vocabulary][N: size]], [plural: -s]. Domain: consumer off-the-shelf products, speech recognition. Hyperonyms: vocabulary size. Synonyms: user vocabulary size, extension vocabulary size. Cohyponym: active vocabulary size, passive vocabulary size. Def.: The number of words a user may add to the lexicon of a speech recogniser.

exchange
/eksˈtʃeɪndʒ/, /eks'tSeIndZ/, [N: exchange], [plural: -s]. Domain: interactive dialogue systems. Meronym. sup.: dialogue. Def.: A pair of contiguous and related turns, one spoken by each party in the dialogue.

expansion model
/eks'pænʃən 'mɒdəl/, /eks'p{nS@n 'mQd@l/, [N: [N: expansion][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: coarticulation model. Cohyponym: time-locked model, look-ahead model, hybrid model. Def.: The expansion model is based on the fact that the protrusion effect of a vowel can be expanded. The zone of influence depends on the number of consonants to the next (or from the previous) vowel but it cannot arise in less than a constant time.

experimental technique
/eksperɪ'mentəl tek'niːk/, /eksperI'ment@l tek'ni:k/, [N: [AJ: experimental][N: technique]], [plural: -s]. Domain: multimodal systems. Hyperonyms: evaluation method. Hyponyms: benchmark evaluation, user study, simulation study, iterative design, rapid prototyping. Cohyponym: predictive model, expert evaluation. Def.: Experimental techniques deal with real data observed from real users accomplishing real tasks with an actual system.

expert-based evaluation
/'ekspɜːt 'beɪst ɪvæljʊ'eɪʃən/, /'eksp3:t 'beIst Iv{ljU'eIS@n/, [N: [N: expert][AJ: based][N: evaluation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: evaluation method. Synonyms: expert evaluation. Cohyponym: theory-based evaluation, user-based evaluation. Def.: Expert-based evaluation involves an expert using the system in a more or less structured way, to determine whether the system matches predefined criteria or guidelines, or whether it violates some established design guidelines and heuristics. The evaluation yields the evaluator's subjective judgement on the system's conformity to general human factors, principles and approved guidelines.

eXtended Markup Language
/eks'tendɪd 'mɑːkʌp 'læŋgwɪdʒ/, /eks'tendId 'mA:kVp 'l{NgwIdZ/, [N: [AJ: eXtended][N: Markup][N: Language]], [plural: none]. Hyperonyms: Standard Generalized Markup Language, SGML. Synonyms: XML. Def.: XML is a simplified and flexible SGML (Standard Generalized Markup Language, ISO 8879) derivate which many expect to become the standard language for describing WWW documents.

extension vocabulary size
/eks'tenʃən və'kæbjʊlərɪ 'saɪz/, /eks'tenS@n v@'k{bjUl@ri 'saIz/, [N: [N: extension][N: vocabulary][N: size]], [plural: -s]. Domain: consumer off-the-shelf products, speech recognition. Hyperonyms: vocabulary size. Synonyms: user vocabulary size, exception vocabulary size. Cohyponym: active vocabulary size, passive vocabulary size. Def.: The number of words a user may add to the lexicon of a speech recogniser.

E_ToBI
/'iː 'təʊbi/, /'i: 't@Ubi/, [N: [N: E_ToBI]], [plural: none]. Domain: corpora. Hyperonyms: ToBI. Cohyponym: GlaToBI, J_ToBI, G_ToBI. Def.: English Tone and Break Index prosodic annotation scheme, a variant of ToBI, developed for US American English, and applied to Southern Standard British and Standard Australian English.

F0
/'ef 'ziːrəʊ/, /'ef 'zi:r@U/, [N: F0], [plural: -s]. Domain: physical characterisation. Hyperonyms: frequency, acoustic measure. Synonyms: fundamental frequency, F zero. Cohyponym: harmonic; amplitude, intensity. Def.: The fundamental frequency of an utterance, the more or less direct realisation of a tone sequence or intonation contour, corresponding quite directly to the glottal phonation rate and mapped by a non-linear function to perceived pitch. An acoustic measurement, rather than a perceptual category.

face detection
/'feɪs dɪ'tekʃən/, /'feIs dI'tekS@n/, [N: [N: face][N: detection]], [plural: none]. Domain: multimodal systems. Hyperonyms: recognition task. Synonyms: face locating. Cohyponym: face recognition, face tracking. Def.: Determining whether a scene has any faces; may include , locating the positions of faces.

face locating
/ˈfeɪs ləˈkeɪtɪŋ/, /'feIs l@'keItIN/, [N: [N: face][N: locating]], [plural: none]. Domain: multimodal systems. Hyperonyms: recognition task. Synonyms: face detection. Cohyponym: face recognition, face tracking. Meronym. sup.: recognition process. Def.: When a face is detected in the input scene, determining its exact location.

face recognition
/ˈfeɪs rekəgˈnɪʃən/, /'feIs rek@g'nIS@n/, [N: [N: face][N: recognition]], [plural: none]. Domain: multimodal systems. Hyperonyms: recognition task. Hyponyms: template matching, feature-based recognition. Cohyponym: face locating, face tracking. Def.: Assigning the face in an input image to a visual recognition system to one of a set of known faces.

face synthesis
/ˈfeɪs ˈsɪnθəsɪs/, /'feIs 'sInT@sIs/, [N: [N: face][N: synthesis]], [plural: face syntheses]. Domain: multimodal systems. Hyperonyms: synthesis task. Hyponyms: performance-driven face synthesis, audio-driven face synthesis, puppeteer control face synthesis, text-to-visual-speech face synthesis. Def.: The generation of visual images of the appearance and movements of the face used in communication, using formal models of bone and soft tissue structure.

face tracking
/ˈfeɪs ˈtrækɪŋ/, /'feIs 'tr{kIN/, [N: [N: face][N: tracking]], [plural: -s]. Domain: multimodal systems. Hyperonyms: recognition task. Cohyponym: face locating, face recognition. Def.: Locating a face in the first image of a sequence of (consecutive) images and keeping track of it in following images. Face tracking is distinguished from face recognition in that local rather than global search techniques are sufficient: since the movement of a head is typically slow relative to the frame rate, a head moves only a small distance from one frame to the next, and simple tracking algorithms can follow a person's motion in a video sequence. However, to track faces outside close proximity to the camera, the tracking system has to control the camera, including panning, tilting, and zooming. Face tracking algorithms first apply a face recognition algorithm to locate a face, and then local search algorithms to follow face motion within a sequence of video images.

FACS
/ˈfæks/, /'f{ks/, [N: FACS], [plural: none]. Domain: multimodal systems. Hyperonyms: software. Synonyms: Facial Action Coding System. Def.: The Facial Action Coding System (FACS) was described by P. Ekman and W. Friesen. It is designed to describe visible facial actions but it does not look at which muscles are activated to produce the facial actions. It is based on anatomical studies in which positions of points on the face during production of six emotion stereotypes (fear, anger, sadness, disgust, joy, surprise) were measured. FACS is composed of 44 basic units called Action Units (AU). An AU corresponds to the action of a muscle or a group of related muscles. Each AU describes the direct effect of muscle contraction as well as any secondary effects due to movement propagation, and apparition of wrinkles or bulges. A facial expression is the combination of AUs. Most of the AUs combine additively. But they may also be subject to rules of dominance (an AU disappears for the benefit of another AU), substitution (an AU is eliminated when others produce the same effect), alteration (AUs cannot combine).

false acceptance
/ˈfɒls əkˈseptəns/, /'fQls @k'sept@ns/, [N: [AJ: false][N: speaker][N: acceptance]], [plural: -s]. Domain: speaker recognition, system design. Hyperonyms: acceptance; error. Synonyms: false speaker acceptance, type-II error, false positive. Cohyponym: false rejection. Def.: Erroneous acceptance of an impostor in open-set speaker identification or in speaker verification.

false rejection
/ˈfɔls rɪˈdʒekʃən/ , **/’fQls rI’dZekS@n/** , [N: [AJ: false][N: rejection]] , [plural: -s] . Domain: speaker recognition, system design. Hyperonyms: rejection; error. Synonyms: false speaker rejection, type-I error, false negative. Cohyponym: false acceptance. Def.: Erroneous rejection of a registered speaker or of a genuine speaker in open-set speaker identification or speaker verification.

false start
/ˈfɔls ˈstɑːt/, **/’fQls ’stA:t/**, [N: [AJ: false][N: start]], [plural: -s]. Domain: corpora. Def.: A dysfluency phenomenon where the speaker interrupts an utterance he has already begun and corrects or reformulates it.

FAQ
1. /ˈef ˈeɪ ˈkjuː/ 2. /ˈfæk/, **1. /’ef ’eI ’kju:/ 2. /’f{k/**, [N: FAQ], [plural: -s]. Synonyms: Frequently Asked Questions. Def.: A compiled list (FAQ, FAQ list) of questions frequently asked by new users of a product or members of a newsgroup, and the answers to these questions.

far end echo
/fɑːrˈend ˈekəʊ/, **/fA:r’end ’ek@U/**, [N: [AJ: far][N: end][N: echo]], [plural: -es]. Hyperonyms: echo. Cohyponym: near end echo. Def.: The echo of the system output generated at the site of the caller. Far end echo can be acoustic and/or electric. Acoustic echo is especially likely to occur with speaker phones.

Fast Fourier Transform
/ˈfɑːst ˈfuːrɪeɪ ˈtrænsfɔːm/, **/’fA:st ’fu:rIeI ’tr{nsfO:m/**, [N: [AJ: Fast][N: Fourier][N: Transform]], [plural: -s]. Hyperonyms: DFT, Discrete Fourier Transform. Synonyms: FFT. Def.: Commonly used optimisation of the discrete Fourier transform (DFT) algorithm using binary structuring of the signal data to reduce complexity and increase processing speed. (Clark & Yallop 1995, p. 259) The Fast Fourier Transform is used to analyse the speech spectrum.

feature-based face recognition
/ˈfiːtʃə ˈbeɪst ˈfeɪs rekəgˈnɪʃən/, **/’fi:tS@ ’beIst ’feIs rek@g’nIS@n/**, [N: [N: feature][AJ: based][N: face][N: recognition]] , [plural: none]. Domain: multimodal systems. Hyperonyms: face recognition . Cohyponym: template matching. Def.: Face recognition using a set of geometrical features such as the relative position and size of the nose, eyes, mouth and chin. The distance between features of input signals is measured. Features are primitives of input obtained in a step of preprocessing the input. Geometric features correspond to parameters such as angles, distances and curvatures of the eyes, nose, mouth. Anthropometric features and profiles are also used. Parameters can be extracted by first reducing the information from the video image: a binary image is generated using a threshold value; the chroma-key technique is used to detach the lips from the image background; reflective markers are placed onto and around the lip area. Next, the face is identified by comparing its features with features of faces stored in a database. Before features can be compared, scale normalisation ensures that face images are of the same scale.

feature-based model
/ˈfiːtʃə ˈbeɪst ˈmɒdəl/, **/’fi:tS@ ’beIst ’mQd@l/**, [N: [N: feature][AJ: based][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: look-ahead model. Cohyponym: target-based model, goal-based model. Def.: A feature-based model is a model of speech production or perception in which the basic unit is not the temporal segment but properties of stretches (intervals) of the signal. Coarticulation in feature-based models starts as soon as features involved at the articulatory level in segments are compatible with the features used to realise the current segment.

FEM
/'ef 'iː 'em/, /'ef 'i: 'em/, [N: FEM], [plural: -s]. Domain: multimodal systems. Synonyms: Finite Element Method. Def.: Finite Element Methods (FEM) are used in many areas of physics, chemistry and engineering, and have been applied to simulate the visco-elasticity properties of the skin. These models have mainly been applied to facial surgery simulation. They model with good approximation the skin and muscle actions but the complexity and duration of the computation forbids its use in interactive applications for now. The computation time even on very powerful machines does not allow for real-time animation. FEM have been used, for example, to model lip shapes during speech.

FERET
/'ferət/, /'fer@t/, [N: FERET], [plural: none]. Domain: multimodal systems. Hyperonyms: database. Def.: US Army database offering a very large collection of images of faces. The images have been made under different lighting conditions, backgrounds, locations and times. The distance between the camera and the subject varies. For each individual, the database contains frontal and a variety of profile views taken at different times, whith changed background and lighting conditions.

FFT
/'ef 'ef 'tiː/, /'ef 'ef 'ti:/, [N: FFT], [plural: -s]. Domain: Spoken Language technology. Hyperonyms: DFT, Discrete Fourier Transform. Synonyms: Fast Fourier Transform. Def.: Commonly used version of the discrete Fourier transform (DFT) algorithm using binary structuring of the signal data to reduce complexity and increase processing speed. (Clark & Yallop 1995, p. 259) The Fast Fourier Transform is used to analyse the speech spectrum.

field testing
/'fiːld 'testɪŋ/, /'fi:ld 'testIN/, [N: [N: field][N: testing]], [plural: none]. Domain: speech synthesis. Hyperonyms: assessment technique. Cohyponym: laboratory testing. Meronym. sup.: speech output testing. Def.: Speech output testing procedure entirely run in the actual application, using the real-life situation with the actual end-users.

filter bank
/'fɪltə 'bæŋk/, /'fIlt@ 'b{Nk/, [N: [N: filter][N: bank]], [plural: -s]. Domain: physical characterisation. Def.: A set of band-pass filters that together cover the frequency range of interest to, for example, a speech recogniser.

filter
/'fɪltə/, /'fIlt@/, [N: filter], [plural: -s]. Domain: physical characterisation. Hyponyms: low-pass filter, high-pass filter, band-pass filter, band-stop filter, notch filter, all-pass filter. Def.: 1. A filter is a system whose output is a subset or substructure of its input. 2. A system which attenuates (weakens) certain frequencies in a signal. 3. A program which selects a section of a search space or a subset of data, for example a query filter or an output filter in a database management system (DBMS); a UNIX filter.

Finite Element Method
/'faɪnaɪt 'eləmənt 'meθəd/, /'faInaIt 'el@m@nt 'meT@d/, [N: [AJ: Finite][N: Element][N: Method]], [plural: -s]. Domain: multimodal systems. Synonyms: FEM. Def.: Finite Element Methods (FEM) are widely used in physics, chemistry and engineering, and have been applied to simulate the visco-elasticity properties of the skin. These models have mainly been applied to facial surgery simulation. They model with good approximation the skin and muscle actions but the complexity and duration of the computation forbids its use in interactive applications for now. The computation time even on very powerful machines does not allow for real-time animation. FEM has also been used to model lip shapes during speech.

finite state grammar
/ˈfaɪnaɪt ˈsteɪt ˈgræmə/, **/’faInaIt ’steIt ’gr{m@/**, [N: [AJ: finite][N: state][N: grammar]], [plural: -s]. Domain: language modelling. Hyperonyms: formal grammar, context-free grammar. Synonyms: regular grammar, Chomsky Type 3 grammar. Cohyponym: finite state automaton. Def.: Finite state grammars (more correctly: regular grammars) define (‘generate’) regular languages and are processed by finite state automata. In a finite state grammar, the occurrence of an item depends at most on the occurrence of an immediately neighbouring item (either left or right, but not mixed left and right in the same grammar).

finite state language model
/ˈfaɪnaɪt ˈsteɪt ˈlæŋgwɪdʒ ˈmɒdəl/, **/’faInaIt ’steIt ’l{NgwIdZ ’mQd@l/**, [N: [AJ: finite][N: state][N: language][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: language model. Def.: A language model based on a (usually probabilistic) finite state automaton. The set of legal word sequences is represented as a finite state network (or regular grammar) whose edges stand for the spoken words, i.e. each path through the network results in a legal word sequence. To make this approach correct from a probabilistic point of view, the edges have to be assigned probabilities. A Hidden Markov Model (HMM) language model is a variety of probabilistic finite state language model. (Gibbon et al. 1997, p. 243)

fixed-vocabulary speaker recognition system
/ˈfɪkst vəˈkæbjʊləri ˈspiːkə rekəgˈnɪʃən ˈsɪstəm/, **/’fIkst v@’k{bjUl@ri ’spi:k@ rek@g’nIS@n ’sIst@m/**, [N: [AJ: fixed][N: vocabulary][N: speaker][N: recognition][N: system], [plural: -s]. Domain: speaker recognition. Hyperonyms: text-independent speaker recognition system. Def.: A text-independent speaker recognition system for which test utterances are composed of words, the order of which varies across speakers and sessions, but for which all the words are pronounced at least once by the speaker when he registers to the system.

flap
/ˈflæp/, **/’fl{p/**, [N: flap], [plural: -s]. Hyperonyms: consonant; manner of articulation. Cohyponym: plosive, nasal, trill, fricative, lateral fricative, approximant, lateral approximant. Def.: A flap consonant sound classified on the basis of its manner of articulation: it refers to any sound produced by a single rapid contact between two organs of articulation (excluding vocal cord vibration); sometimes used synonymously with . (Crystal 1988, p. 123) E.g. The realisation of intervocalic /t/ in some dialects of US American English, as in ’butter’..

flawless speech
/ˈflɔːləs ˈspiːtʃ/, **/’flO:l@s ’spi:tS/**, [N: [AJ: flawless][N: speech]], [plural: none]. Domain: physical characterisation. Hyperonyms: benchmark. Def.: The unweighted reproducible 1:1 transduction of an acoustical signal emitted by a speaker into a sequence of 2 byte numbers that is free of any room or environment information, exhibits a sufficient signal-to-noise ratio of at least 50 dB, and has been produced under recording conditions that do not impose any stress upon the speaker in addition to what might be intended for a given talking situation.

flexible vocabulary
/ˈfleksɪbəl vəˈkæbjʊləri/, **/’fleksIb@l v@’k{bjUl@ri/**, [N: [AJ: flexible][N: vocabulary]], [plural: y/-ies]. Domain: system design. Hyperonyms: vocabulary. Cohyponym: fixed vocabulary. Def.: The feature in a speech recognition system allowing the vocabulary to be changed easily, by, for example, supplying the orthography of the words in the new vocabulary, or their phonetic transcriptions.

font
/ˈfɒnt/, **/’fQnt/**, [N: font], [plural: -s]. Def.: A definition of a set of homogeneously designed graphical forms representing the the characters of an alphabet. E.g. Courier, Times Roman, Helvetica, Arial, Computer Modern.

formal language
/'fɔːməl 'læŋgwɪdʒ/, /'fO:m@l 'l{NgwIdZ/, [N: [AJ: formal][N: language]], [plural: -s]. Domain: language modelling. Hyperonyms: language. Cohyponym: natural language. Def.: An artificial language with a mathematical definition, such as a logic or an algebra, usually developed primarily for purposes of representation and manipulation of symbols and numbers (for example, in mathematics, logic or semantics) and not for the purpose of everyday communication. Formal languages may be operationalised in high level functional or logical programming languages such as LISP or Prolog.

formant extraction
/'fɔːmənt ek'strækʃən/, /'fO:m@nt ek'str{kS@n/, [N: [N: formant][N: extraction]], [plural: -s]. Domain: physical characterisation. Def.: The identification of formant tracks in voiced stretches of a speech signal, for example by identifying peaks in the spectrum.

formant
/'fɔːmənt/, /'fO:m@nt/, [N: formant], [plural: -s]. Domain: physical characterisation. Hyponyms: first formant, second formant, third formant; vowel formant, nasal formant; singer's formant. Def.: A formant is a concentration of acoustic energy in the spectrum, reflecting the way air from the lungs vibrates in the vocal tract, as it changes its shape, and corresponding to resonances in the speech production organs. For any vowel, the air vibrates at many different frequencies all at once, and the most dominant frequencies combine to produce the distinctive vowel qualities. Each dominant band of frequencies constitutes a formant, which shows up clearly on a sound spectrograph as a thick black line. Three main formants provide the basis of vowel description: the 'first formant' is the lowest, and the 'second' and 'third formants' are respectively higher. Other formants are less significant for linguistic analysis. (Crystal 1988, p. 125)

forward looking communicative function
/'fɔːwəd 'lʊkɪŋ kə'mjuːnɪkətɪv 'fʌŋkʃən/, /'fO:w@d 'lUkIN k@'mju:nIk@tIv 'fVNkS@n/, [N: [AV: forward][V: looking][AJ: communicative][N: function]], [plural: -s]. Cohyponym: backward looking communicative function. Def.: A communicative function that either establishes the background for verbal or non-verbal action that is to follow or constrains it.

free form deformation
/'friː 'fɔːm diːfə'meɪʃən/, /'fri: 'fO:m di:f@'meIS@n/, [N: [AJ: free][N: form][N: deformation]] , [plural: -s]. Domain: multimodal systems. Hyperonyms: animation control technique, synthetic model. Cohyponym: procedural model, parametric model. Def.: The technique of free form deformation and rational free form deformation can be applied to model muscle action. A deformation box is set to act on a set of points. The box can stretch, squash or bend. The points inside the box are moved according to the next shape of the box.

free morph
/'friː 'mɔːf/, /'fri: 'mO:f/, [N: [AJ: free][N: morph]], [plural: -s]. Domain: lexicon. Hyperonyms: morph. Cohyponym: bound morph. Def.: A free morph is a morph which can occur on its own with no affixes or prosodic modifications as a separate word. (Gibbon et al. 1997, p. 215) E.g. English 'tree', 'cut', 'find' .

free-field equalised headphone
/'friː 'fiːld 'iːkwəlaɪzd 'hedfəʊn/, /'fri: 'fi:ld 'i:kw@laIzd 'hedf@Un/, [N: [AJ: free][N: field][AJ: equalised][N: headphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: headphone. Cohyponym: diffuse-field equalised headphone. Def.: A free-field equalised headphone produces the same spectral distribution of sound at the ear drum of the listener as does an ideal loudspeaker placed under free-field conditions (e.g. comparable with an anechoic chamber) in front of the listener... A free-field equalised headphone is equalised with respect to the forward direction (whereas the diffuse-field equalised headphone is equalised with respect to an average over all directions of incidence). (Gibbon et al. 1997, p. 324/325)

frequency response
/ˈfriːkwənsi rɪsˈpɒns/, /'fri:kw@nsi rIs'pQns/, [N: [N: frequency][N: response]], [plural: -s]. Domain: physical characterisation. Def.: The frequency response of a system is the transfer function expressing the relation between the frequency range in the input and the frequency range in the output of the system.

frequency
/ˈfriːkwənsi/, /'fri:kw@nsi/, [N: frequency], [plural: y/-ies]. Domain: physical characterisation. Hyperonyms: acoustic measure. Hyponyms: fundamental frequency (F0). Def.: Frequency is the property of a signal which defines the number of repetitions of similar portions of the signal in a given period of time such as 1 second. The Fourier theorem states that any signal can be defined as the point by point sum of a series of simple (pure sinusoid) signals. Frequency is measured in Hertz, i.e. cycles per second.

Frequently Asked Questions
/ˈfriːkwəntli ˈɑːskt ˈkwestʃənz/, /'fri:kw@ntli 'A:skt 'kwestS@nz/, [N: [AV: Frequently][V: Asked][N: Questions]], [plural: always plural]. Synonyms: FAQ. Def.: A compiled list of questions frequently asked by new users of a product or members of a newsgroup, and the answers to these questions.

fricative
/ˈfrɪkətɪv/, /'frIk@tIv/, [N: fricative], [plural: -s]. Hyperonyms: consonant; manner of articulation. Synonyms: spirant. Cohyponym: plosive, nasal, trill, tap, flap, lateral fricative, approximant, lateral approximant. Def.: A fricative is a consonant sound classified phonetically on the basis of its manner of articulation: a sound made when two organs come so close together that the air moving between them produces audible friction. (Crystal 1988, p. 128)

full synonym
/ˈfʊl ˈsɪnənɪm/, /'fUl 'sIn@nIm/, [N: [AJ: full][N: synonym]], [plural: -s]. Domain: lexicon. Hyperonyms: synonym. Cohyponym: partial synonym. Def.: A word in the lexical semantic relation of full synonymy to another word. E.g. FFT - Fast Fourier Transform.

full synonymy
/ˈfʊl sɪˈnɒnəmi/, /'fUl sI'nQn@mi/, [N: [AJ: full][N: synonymy]], [plural: none]. Domain: lexicon. Hyperonyms: synonymy. Cohyponym: partial synonymy. Def.: Full synonymy is a lexical semantic relation between two words which have no reading which they do not share. Full synonymy is generally restricted to a particular semantic domain. The clearest cases of full synonymy are found in abbreviations and the expressions they abbreviate. (Gibbon et al. 1997, p. 850)

functional speech disorder
/ˈfʌŋkʃənəl ˈspiːtʃ dɪsˈɔːdə/, /'fVNkS@n@l 'spi:tS dIs'O:d@/, [N: [AJ: functional][N: speech][N: disorder]], [plural: -s]. Domain: corpora. Hyperonyms: speech disorder. Cohyponym: organic speech disorder. Def.: Speech disorder where there is no clear organic cause. (Gibbon et al. 1997, p. 114)

functional testing
/ˈfʌŋkʃənəl ˈtestɪŋ/, /'fVNkS@n@l 'testIN/, [N: [AJ: functional][N: testing], [plural: none]. Domain: speech synthesis. Hyperonyms: testing procedure. Cohyponym: analytic testing, opinion testing, judgment testing. Def.: Assessment of speech output in terms of how well a system actually performs (some aspect of) its communicative purpose.

functional unit
/ˈfʌŋkʃənəl ˈjuːnɪt/, /'fVNkS@n@l 'ju:nIt/, [N: [AJ: functional][N: unit]], [plural: -s]. Domain: lexicon. Def.: Sequence of functional words which behave as a phonological unit, usually with cliticisation. (Gibbon et al. 1997, p. 220) E.g. French 'n'est-ce pas' /nespa/), English 'Ic'n' /aIkN/ for 'I can' in informal, fast speech or particularly unstressed contexts. (Gibbon et al. 1997, p. 220).

functional utterance
/ˈfʌnkʃənəl ˈʌtərəns/, **/'fVnkS@n@l 'Vt@r@ns/**, [N: [AJ: functional][N: utterance]], [plural: -s]. Def.: A classification of an utterance based on its functional content, rather than its structural, syntactic properties.

functional word
/ˈfʌŋkʃənəl ˈwɜːd/, **/'fVNkS@n@l 'w3:d/**, [N:[AJ: functional][N: word]], [plural: -s]. Domain: lexicon. Cohyponym: lexical word. Def.: A word belonging to a small closed set of words with the function of a grammatical connective or phrase-building operator, and not denoting properties of events and objects in the world. E.g. articles, pronouns, prepositions, conjunctions.

fundamental frequency
/fʌndəˈmentəl ˈfriːkwənsi/, **/fVnd@'ment@l 'fri:kw@nsi/**, [N: [AJ: fundamental][N: frequency]], [plural: y/-ies]. Domain: physical characterisation. Hyperonyms: acoustic measure; frequency. Synonyms: F0. Cohyponym: harmonic; amplitude, intensity. Def.: The (lowest) frequency component in a harmonic sound, of which the frequencies of the harmonics are integer multiples. Frequency refers to the number of complete cycles (opening and closing movements) of vocal cord vibration in a unit of time (per second). The fundamental frequency or F0 ('f nought', 'f zero'), is of particular importance in studies of intonation, where it displays a reasonably close correspondence with the pitch movements involved. It is measured in hertz (Hz),... (cf. also Crystal 1988, p. 131)

fusion
/fjuːʒən/, **/fju:Z@n/**, [N: fusion], [plural: -s]. Domain: multimodal systems. Hyponyms: microtemporal fusion, macrotemporal fusion, contextual fusion. Def.: Extracting a meaningful representation of multiple (potentially multimodal) input events.

garbage model
/ˈgɑːbɪdʒ ˈmɒdəl/, **/'gA:bIdZ 'mQd@l/**, [N: [N: garbage][N: model]], [plural: -s]. Hyperonyms: speech model. Def.: A general model of units of speech which used in speech recognisers to match spoken input that cannot be matched to words in the vocabulary.

general lexicon theory
/ˈdʒenərəl ˈleksɪkən ˈθiːəri/, **/'dZen@r@l 'leksIk@n 'Ti:@ri/**, [N: [AJ: general][N: lexicon][N: theory]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: lexicon theory. Cohyponym: specific lexicon theory. Def.: A general lexicon theory is a general theory of lexical objects and information, for instance a theory of lexical signs and their representation, rather than a lexicon as a specific theory of, for instance, the words of a particular language and their properties. (Gibbon et al. 1997, p. 193)

generic concept hierarchy
/dʒəˈnerɪk ˈkɒnsept ˈhaɪərɑːki/, **/dZ@'nerIk 'kQnsept 'haI@rA:ki/**, [N: [AJ: generic][N: concept][N: hierarchy]], [plural: y/-ies]. Domain: terminology. Hyperonyms: hierarchy. Synonyms: taxonomy, logical concept hierarchy, ISA hierarchy, implication hierarchy, inheritance hierarchy. Cohyponym: meronomy, mereonomy, PARTOF hierarchy. Def.: Hierarchy of concepts holding an ISA relation, i.e. a hierarchy defined by the relation of generalisation and its inverse, specialisation.

genuine speaker
/ˈdʒenjuɪn ˈspiːkə/, **/'dZenjuIn 'spi:k@/**, [N: [AJ: genuine][N: speaker]], [plural: -s] . Domain: speaker recognition. Hyperonyms: applicant speaker. Synonyms: authentic speaker; true speaker; correct speaker. Def.: A speaker whose real identity is in accordance with the claimed identity. By extension: a speaker whose actual character and claimed class are in accordance. (Gibbon et al. 1997, p. 413) E.g. In sex verification: a female speaker claiming that she is a female speaker..

geometric parameter
/dʒiːəˈmetrɪk pəˈræmɪtə/, /dZi:@'metrIk p@'r{mIt@/, [N: [AJ: geometric][N: parameter]], [plural: -s]. Domain: multimodal systems. Cohyponym: acoustic parameter. Def.: Geometric parameters are used to represent the vocal tract in the articulatory domain. E.g. tongue body center, jaw angle, lip height, lip protrusion.

geometric template matching
/dʒiːəˈmetrɪk ˈtempleɪt ˈmætʃɪŋ/, /dZi:@'metrIk 'templeIt 'm{tSIN/, [N: [AJ: geometric][N: template][N: matching]], [plural: none]. Domain: multimodal systems. Hyperonyms: template matching. Cohyponym: Principal Component Analysis, PCA, deformable template matching, optical flow technique, neural network based approach. Def.: Geometric templates of specific facial features (such as eyes and lips) are built to describe and then recognise faces. These templates are constructed based on a priori knowledge about the feature shapes. Templates are parameterised curves that can deform during model fitting. The curves follow the outline of the facial features, and their final shapes can be used to verify if the observed object is an eye, lip, or face. An appropriate distance metric has to be defined, for example a potential energy function. Minimising the potential energy is equivalent to forcing the templates toward salient features (valleys, edges, peaks and intensity). A problem with this technique is its relative dependency on position and lighting.

gesture
/ˈdʒestʃə/, /'dZestS@/, [N: gesture], [plural: -s]. Domain: multimodal systems. Hyponyms: single-stroke gesture, multi-stroke gesture; hand gesture, body gesture; pointing, 2D gesture, 3D gesture; symbolic gesture, deictic gesture, metaphoric gesture, iconic gesture. Cohyponym: 2D gesture, 3D gesture, posture. Def.: A movement of a part of the body, generally with a communicative function. Gesture input is captured using dedicated input devices, e.g. for 2D gestures and pointing: mouse and stylus, for 3D gestures: data glove, position trackers, or cameras. Pattern classification and computer vision algorithms have been developed to automatically recognise gestures. Articulatory phonology defines properties of speech production in terms of gestures of the articulatory organs.

gesture-based interaction
/ˈdʒestʃə ˈbeɪst ɪntəˈrækʃən/, /'dZestS@ 'beIst Int@'r{kS@n/, [N: [N: gesture][AJ: based][N: interaction]], [plural: -s]. Domain: multimodal systems. Hyperonyms: human-computer interaction. Def.: Human-computer interaction using gestures as a modality.

glass box approach
/ˈglɑːs ˈbɒks əˈprəʊtʃ/, /'glA:s 'bQks @'pr@UtS/, [N: [N: glass][N: box][N: approach]] , [plural: -es] . Domain: speech synthesis. Hyperonyms: testing procedure. Cohyponym: black box approach. Def.: A testing procedure relying on the input/output function of a system, in which the effects of all modules in a text-to-speech system but one are kept constant, and the characteristics of the free module are systematically varied, so that any difference in the assessment of the system's output must be caused by the variations in the target module (diagnostic testing). Glass box testing presupposes that the researcher has control over the input and output of each individual module.

GlaToBI
/glɑːˈtəʊbi/, /glA:'t@Ubi/, [N: GlaToBI], [plural: none]. Domain: corpora, dialogue representation. Hyperonyms: ToBI . Cohyponym: J_ToBI, E_ToBI, G_ToBI. Def.: A variant of the ToBI prosodic annotation scheme developed for the transcription of the English of Glasgow, Scotland, UK.

Global System for Mobile Communication
/ˈgləʊbəl ˈsɪstəm ˈfɔː ˈməʊbaɪl kəmjuːnɪˈkeɪʃən/, /'gl@Ub@l 'sIst@m 'fO: 'm@UbaIl k@mju:nI'keIS@n/, [N: [AJ: Global][N: System][PREP: for][AJ: Mobile][N: Communication]], [plural: none]. Hyperonyms: standard; speech coding algorithm; digital telephony. Synonyms: GSM. Cohyponym: ISDN, Integrated Services Digital Network. Def.: A family of public-domain speech coding algorithms (GSM algorithms) used for digital mobile telephony used in Europe and many other parts of the world.

global testing
/'gləʊbəl 'testɪŋ/, /'gl@Ub@l 'testIN/, [N: [AJ: global][N: testing]], [plural: -s] . Domain: speech synthesis. Hyperonyms: testing procedure. Cohyponym: analytic testing. Def.: A testing procedure in which the listener is instructed to attend to the general performance of a speech output system, e.g. in terms of listening effort, acceptability, and naturalness.

glottal consonant
/'glɒtəl 'kɒnsənənt/ , /'glQt@l 'kQns@n@nt/ , [N: [AJ: glottal][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, labiodental consonant, dental consonant, alveolar consonant, postalveolar consonant, retroflex consonant, palatal consonant, velar consonant, uvular consonant, pharyngeal consonant. Def.: A glottal consonant is classified on the basis of the place of articulation: it is made in the larynx, due to the closure or narrowing of the glottis. (Crystal 1988, p. 136)

glottal stop
/'glɒtəl 'stɒp/, /'glQt@l 'stQp/, [N: [AJ: glottal][N: stop]], [plural: -s]. Hyperonyms: glottal consonant. Def.: A glottal stop is the audible release of a complete closure at the glottis. (Crystal 1988)

glottal-to-noise excitation parameter
/'glɒtəl tʊ 'nɔɪz eksaɪ'teɪʃən pə'ræmɪtə/, /'glQt@l tU 'nOIz eksaI'teIS@n p@'r{mIt@/, [N: [AJ: glottal][PREP: to][N: noise][N: excitation][N: parameter]], [plural: -s]. Domain: physical characterisation. Hyperonyms: measure. Def.: The glottal-to-noise parameter defines quantitatively to what extent vocal excitation is mainly due to glottal vibration or rather turbulent noise.

glottis
/'glɒtɪs/, /'glQtIs/, [N: glottis], [plural: -es]. Hyperonyms: articulator. Def.: The aperture between the vocal folds in the larynx. (Clark & Yallop 1995, p. 15)

goal-based model
/'gəʊl 'beɪst 'mɒdəl/, /'g@Ul 'beIst 'mQd@l/, [N: [N: goal][AJ: based][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: look-ahead model. Cohyponym: feature-based model, target-based model. Def.: A goal-based model defines strategies for problem solving or action by inferring backwards from goals to the means whereby the goals may be reached, for example the sequence of goals to be achieved in computing articulator behaviours.

goat
/'gəʊt/, /'g@Ut/, [N: goat], [plural: -s]. Domain: speaker recognition. Hyperonyms: registered speaker. Synonyms: unreliable speaker. Cohyponym: sheep, dependable speaker. Def.: A speaker who obtains particularly bad performance with a speech recognition system. A speaker with a high misclassification rate. (Gibbon et al. 1997, p. 432)

grammar based language model
/'græmə 'beɪst 'læŋgwɪdʒ 'mɒdəl/, /'gr{m@ 'beIst 'l{NgwIdZ 'mQd@l/, [N: [N: grammar][AJ: based][N: language][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: language model. Cohyponym: stochastic language model, finite state language model. Meronym. sub.: link grammar. Def.: Typically, grammar based language models are based on variants of stochastic context free grammars or other phrase structure grammars. (Gibbon et al. 1997, p. 243)

grammar
/'græmə/, /'gr{m@/, [N: grammar] , [plural: -s] . Domain: language modelling. Hyperonyms: linguistic characterisation. Hyponyms: formal grammar, stochastic grammar . Cohyponym: lexicon . Def.: 1. The study of the principles by which units of language, such as sentences and words, are constructed. 2. A set of rules that define how the basic units of a particular language, such as words and sentences, are constructed. Traditionally, a grammar deals with syntax, the composition of sentences, and morphology, the composition of words, but also generally includes treatment of spelling, punctuation and pronunciation. the words in a language can follow each other. 3. A component of a language processing system, such as a parser or a sentence generator, which defines the order of words in sentences.

grammatical category
/grə'mætɪkəl 'kætəgəri/, /gr@'m{tIk@l 'k{t@g@ri/, [N: [AJ: grammatical][N: category]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: part of speech (POS). Hyponyms: pronoun, article, interjection, conjunction, preposition. Cohyponym: lexical category. Def.: Grammatical categories (as opposed to lexical categories) are the closed classes or parts of speech (POS) which express syntactic and indexical relations. (Gibbon et al. 1997, p. 218) E.g. pronoun and Article (anaphoric and deictic relations), Preposition (spatial, temporal, personal relations etc.), Conjunction (propositional relations), Interjection (dialogue relations).

grammatical morph
/grə'mætɪkəl 'mɔːf/, /gr@'m{tIk@l 'mO:f/, [N:[AJ: grammatical][N: morph]], [plural: -s]. Domain: lexicon. Cohyponym: lexical morph. Def.: Orthographic or phonological realisation of a grammatical morpheme.

grammatical morpheme
/grə'mætɪkəl 'mɔːfiːm/, /gr@'m{tIk@l 'mO:fi:m/, [N: [AJ: grammatical][N: morpheme]] , [plural: -s]. Domain: lexicon. Hyperonyms: morpheme. Cohyponym: lexical morpheme. Def.: A grammatical morpheme is characterised by membership of a closed class, defined by its distribution with respect to larger units such as sentences or complex words. A grammatical morpheme indicates a grammatical or indexical relation between lexical morphemes. (Gibbon et al. 1997, p. 215) E.g. inflectional and derivational endings; function words such as prepositions, articles.

grapheme-phoneme conversion
/'græfiːm 'fəʊniːm kən'vɜːʃən/, /'gr{fi:m 'f@Uni:m k@n'v3:S@n/, [N: [N: grapheme][N: phoneme][N: conversion]], [plural: -s]. Domain: speech synthesis. Meronym. sup.: text-to-speech system. Def.: Grapheme-to-phoneme conversion is the process of mapping an orthographic text on to a phonemic representation, generally as a component in Text-To-Speech Synthesis (TTS). A grapheme-to-phoneme converter may also include a parser to provide stress marks, (sentence) accent positions, and boundaries. The main strategies used in grapheme-to-phoneme conversion are table lookup (particularly for vocabularies with large numbers of irregular spellings), grapheme-to-phoneme rules, and stochastic finite state conversion based on training with large amounts of aligned orthographic texts and phonemic transcriptions data.

graphemic word
/grə'fiːmɪk 'wɜːd/, /gr@'fi:mIk 'w3:d/, [N:[AJ: graphemic][N: word]], [plural: -s]. Domain: lexicon. Cohyponym: morphological word, orthographic word, phonetic word. Def.: Word consisting of a minimal contrastive unit in the writing system of a language.

graphic mark
/'græfɪk 'mɑːk/, /'gr{fIk 'mA:k/, [N: [AJ: graphic][N: mark]] , [plural: -s]. Domain: multimodal systems. Synonyms: 2D gesture. Cohyponym: pointing, 3D gesture. Def.: Graphic marks refer to the reflexes of movements which exert pressure on a surface, for example marks drawn with a pen on a touch-sensitive display.

greedy search
/'griːdi 'sɜːtʃ/, /'gri:di 's3:tS/, [N: [AJ: greedy][N: search]], [plural: -es]. Domain: language modelling. Hyperonyms: search. Synonyms: greedy algorithm, hill-climbing algorithm. Def.: Greedy search is a strategy for fast search using an algorithm which trades off between performing a possibly exponential exhaustive search of a given search space and returning a non-optimal solution. A greedy search algorithm always takes a locally optimal solution and incurs the risk that this might not be globally optimal. Some search problems cannot be solved using greedy search. In the Artificial Intelligence literature, the strategy is generally referred to as hill-climbing: to reach the goal of getting to the highest possible point as fast (as greedily) as possible, one climbs the highest hill in the vicinity; this is of course not necessarily the highest point from a global point of view.

grunt detection
/'grʌnt dɪ'tekʃən/, /'grVnt dI'tekS@n/, [N: [N: grunt][N: detection]], [plural: -s]. Def.: Property of an interactive, usually telephone-based, system responding to the presence of input speech rather than to its lexical content.

GSM
/'dʒiː 'es 'em/, /'dZi: 'es 'em/, [N: GSM], [plural: none]. Hyperonyms: standard; speech coding algorithm; digital telephony. Synonyms: Global System for Mobile Communication. Cohyponym: ISDN, Integrated Services Digital Network. Def.: A family of public-domain speech coding algorithms (GSM algorithms) used for digital mobile telephony used in Europe and many other parts of the world.

Guidelines for Electronic Text Encoding and Interchange
/'gaɪdlaɪnz 'fɔːr elek'trɒnɪk 'tekst ɪŋ'kəʊdɪŋ 'ænd ɪntə'tʃeɪndʒ/, /'gaIdlaInz 'fO:r elek'trQnIk 'tekst IN'k@UdIN '{nd Int@'tSeIndZ/, [N:[N: Guidelines][PREP: for][AJ: Electronic][N: Text][N: Encoding][C: and][N: Interchange]], [plural: none]. Hyperonyms: TEI. Synonyms: TEI P3. Def.: These Guidelines are the result of over five years' effort on the standardisation of markup conventions for different types of text in an international cooperative project called the Text Encoding Initiative (TEI). The TEI was established in 1987 under the joint sponsorship of the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing.

G_ToBI
/'dʒiː 'təʊbi/, /'dZi: 't@Ubi/, [N: G_ToBI], [plural: none]. Domain: dialogue representation. Hyperonyms: ToBI. Cohyponym: E_ToBI, GlaToBI, J_ToBI. Def.: German adaptation of the ToBI (tone and break indices) system.

half duplex
/'hɑːf 'djuːpleks/, /'hA:f 'dju:pleks/, [N: [AJ: half][N: duplex]], [plural: -es]. Domain: system design. Hyperonyms: property of a communication system. Cohyponym: full duplex, duplex. Def.: In a half-duplex communication system the signals flowing in the two directions may not be simultaneous.

haptic output device
/'hæptɪk 'aʊtpʊt dɪ'vaɪs/, /'h{ptIk 'aUtpUt dI'vaIs/, [N: [AJ: haptic][N: output][N: device]], [plural: -s]. Domain: multimodal systems. Hyperonyms: output device. Cohyponym: visual output device, acoustic output device. Def.: Devices stimulating the tactile sense. Complex devices are expensive (minimum 10,000 dollars, and more), but simpler systems are also found, for example in mobile telephones as a ringing signal. Vibration generation seems the most adequate way to stimulate the tactile sense. Work is underway to develop electrotactile stimulation. But most haptic devices do not act directly on the somatic sense of users (for example force feedback devices). DataGlove can also be equipped to send feedback to the user. A simple 2D mouse can be transformed to produce force feedback and predict the user's next actions.

harsh voice
/ˈhɑːʃ ˈvɔɪs/, /'hA:S 'vOIs/, [N: [AJ: harsh][N: voice]], [plural: -s]. Domain: physical characterisation. Hyperonyms: voice characteristic. Cohyponym: breathy voice, creaky voice. Def.: A harsh voice may result from a very fast glottal closing gesture, and a high closed/open phase ratio.

headphone
/ˈhedfəʊn/, /'hedf@Un/, [N: headphone], [plural: -s]. Domain: physical characterisation. Hyponyms: closed headphone, open headphone, earphone. Def.: An acoustic output device worn on the head, with separate transducers for each ear which are generally supported by a curved connecting bar.

headset microphone
/ˈhedset ˈmaɪkrəfəʊn/, /'hedset 'maIkr@f@Un/, [N: [AJ: headset][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: microphone. Cohyponym: omnidirectional microphone, unidirectional microphone, bidirectional microphone, ultradirectional microphone, pressure zone microphone; handheld microphone, table-top microphone, room microphone, headmounted microphone. Def.: A microphone mounted on a frame worn on the head, often as part of a headphone, which is supported at a constant distance from and angle to the mouth. The use of a headset microphone is recommended in all situations where a high ambient noise rejection is needed. The noise rejection properties are mainly due to the extremely close talking distance which allows preamplifier gain to be greatly reduced. Additional noise rejection can be achieved by choosing microphone capsules with directional properties. The good noise rejection behaviour has to be traded off by a degraded frequency response at low frequencies, which leads to an effect already referred to as proximity effect. (Gibbon et al. 1997, p. 306)

health state identification
/ˈhelθ ˈsteɪt aɪdentɪfɪˈkeɪʃən/, /'helT 'steIt aIdentIfI'keIS@n/, [N: [N: health][N: state][N: identification]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker classification task. Cohyponym: sex identification, age identification, mood identification, accent identification, speaker cluster identification. Def.: The task of detecting pathologies using voice samples, for instance, vocal cord disfunctionings, is called health state identification. This concept could be extended to the characterisation of voices modified by external temporary factors that affect speech production, such as alcohol for instance. (Gibbon et al. 1997, p. 408)

heterograph
/ˈhetərəgrɑːf/, /'het@r@grA:f/, [N: heterograph], [plural: -s]. Domain: lexicon. Hyperonyms: word, lexical item. Cohyponym: homograph. Def.: Two orthographic forms of the same word are heterographs. (Gibbon et al. 1997, p. 201) E.g. standardise - standardize /'st{nd@daIz/.

heterography
/hetəˈrɒgrəfi/, /het@'rQgr@fi/, [N: heterography], [plural: y/-ies]. Domain: lexicon. Hyperonyms: orthographic variation. Cohyponym: homography. Def.: Relation between heterographs, i.e. orthographic forms of the same word. E.g. standardise - standardize /st{nd@daIz/.

heterophone
/ˈhetərəfəʊn/, /'het@r@f@Un/, [N: heterophone], [plural: -s]. Domain: lexicon. Hyperonyms: phonological variant of a word. Cohyponym: homophone, heterograph. Def.: Two phonological forms of the same word are heterophones. (Gibbon et al. 1997, p. 201) E.g. either /aID@/ - /i:D@/.

heterophony
/hetəˈrɒfəni/, /het@'rQf@ni/, [N: heterophony], [plural: y/-ies]. Domain: lexicon. Hyperonyms: phonological variation. Cohyponym: homophony. Def.: Relation between two phonological forms of the same word. E.g. either /aID@/ - /i:D@/.

Hidden Markov Model
/'hɪdən 'mɑːkɒf 'mɒdəl/, /'hId@n 'mA:kQf 'mQd@l/, [N: [AJ: Hidden][N: Markov][N: Model]], [plural: -s]. Domain: language modelling. Hyperonyms: probabilistic language model. Synonyms: HMM. Def.: A statistical approach to extracting symbolic data from signal data, e.g. phonemes from speech. Basically, an HMM is a finite automaton with probability values for every arc and arc label.

Hidden Markov Toolkit
/'hɪdən 'mɑːkɒf 'tuːlkɪt/, /'hId@n 'mA:kQf 'tu:lkIt/, [N: [AJ: Hidden][N: Markov][N: Toolkit]], [plural: none]. Hyperonyms: toolkit. Synonyms: HTK. Def.: A commercial toolkit for building Hidden Markov Models (HMM).

hierarchical lexicon
/haɪə'rɑːkɪkəl 'leksɪkən/, /haI@'rA:kIk@l 'leksIk@n/, [N: [AJ: hierarchical][N: lexicon]], [plural: -s; hierarchical lexica]. Domain: lexicon. Hyperonyms: lexicon. Def.: Lexicon in which fully regular information (e.g. in compounds) can be inherited from elsewhere in the lexicon (e.g. from the parts of the compounds), while idiosyncratic information is specified locally. Thus modern computational lexicographic practice attempts to reduce the redundancy in a lexicon as far as possible. (Gibbon et al. 1997, p. 195)

hierarchy
/'haɪərɑːki/, /'haI@rA:ki/, [N: hierarchy], [plural: y/-ies]. Domain: terminology. Def.: A kind partial ordering over a set of categories, generally represented by a tree graph in which the root is interpreted as being highest, or most dominant, and the leaves are interpreted as being lowest, or least dominant with respect to some empirical parameter.

high-pass filter
/'haɪpɑːs 'fɪltə/, /'haIpA:s 'fIlt@/, [N: [AJ: high][V: pass][N: filter]], [plural: -s]. Domain: physical characterisation. Hyperonyms: filter. Cohyponym: low-pass filter, band-pass filter, band-stop filter, notch filter, all-pass filter. Def.: A high-pass filter removes or reduces the amplitude of low frequencies a signal, i.e. it attenuates frequencies below a specified threshold or cut-off frequency.

HMM
/'eɪtʃ 'em 'em/, /'eItS 'em 'em/, [N: HMM], [plural: -s]. Domain: language modelling. Hyperonyms: statistical model. Synonyms: Hidden Markov Model. Def.: A statistical approach to extracting symbolic data from signal data, e.g. phonemes from speech. Basically, an HMM is a finite automaton with probability values for every arc and arc label.

hoarse voice
/'hɔːs 'vɔɪs/, /'hO:s 'vOIs/, [N: [AJ: hoarse][N: voice]], [plural: -s]. Domain: corpora, physical characterisation. Hyperonyms: phonetic feature; voice characteristic. Def.: A hoarse voice is a mixture of laryngeal irregularity with breathiness.

holistic approach
/həʊ'lɪstɪk ə'prəʊtʃ/, /h@U'lIstIk @'pr@UtS/, [N: [AJ: holistic][N: approach]], [plural: -es]. Domain: multimodal systems. Cohyponym: analytic approach. Def.: Approach to, for example, speech recognition based on global information (the whole input signal).

homograph
/'hɒməgrɑːf/, /'hQm@grA:f/, [N: homograph], [plural: -s]. Domain: lexicon. Hyperonyms: word, lexical item. Def.: Two lexical items are homographs if they have the same spelling but different pronunciation (heterophonous homographs) and/or meaning (Gibbon et al. 1997, p. 185) E.g. English 'read' /ri:d/ (infinitive) - /red/ (past); English 'row' /r@U/ 'horizontal sequence' - /raU/ 'quarrel'.

homography
/hə'mɒgrəfi/, /h@'mQgr@fi/, [N: homography], [plural: y/-ies]. Domain: lexicon. Hyperonyms: surface-meaning relation. Cohyponym: heterography. Def.: Relation between two words with the same orthographic form and different phonological forms. (Gibbon et al. 1997, p. 201) E.g. row /r@U/ 'horizontal sequence', /raU/ 'noise, quarrel'.

homonym
/'hɒmənɪm/, /'hQm@nIm/, [N: homonym], [plural: -s]. Domain: lexicon. Hyperonyms: word, lexical item. Def.: Two words with the same orthographic and phonological forms, but different syntactic categories and/or meanings are homonyms. (Gibbon et al. 1997, p. 201) E.g. 'mate' /meIt/ 'friend' or 'a final win-lose state of play in a chess game in which the loser's king is in check (threatened) and cannot be moved without continuing to be in check; check mate'.

homonymy
/hə'mɒnɪmi/, /h@'mQnImi/, [N: homonymy], [plural: y/-ies]. Domain: lexicon. Hyperonyms: surface-meaning relation. Def.: Relation between two words with the same orthographic and phonological forms, but different syntactic categories and/or meanings. (Gibbon et al. 1997, p. 201) E.g. mate /meIt/ 'friend' or 'state of play in a chess game'.

homophone
/'hɒməfəʊn/, /'hQm@f@Un/, [N: homophone], [plural: -s]. Domain: lexicon. Hyperonyms: word, lexical item. Cohyponym: homograph, heterophone. Def.: Two words with the same phonological form and different orthographic forms are (heterographic) homophones. (Gibbon et al. 1997, p. 201) E.g. meet /mi:t/ 'encounter' - meat /mi:t/ 'edible animal tissue'.

homophony
/hə'mɒfəni/, /h@'mQf@ni/, [N: homophony], [plural: y/-ies]. Domain: lexicon. Hyperonyms: surface-meaning relation. Cohyponym: heterophony. Def.: Relation between two words that have the same phonological form and different orthographic forms. (Gibbon et al. 1997, p. 201) E.g. meet /mi:t/ 'encounter' - meat /mi:t/ 'edible animal tissue'.

HTK
/'eɪtʃ 'tiː 'keɪ/, /'eItS 'ti: 'keI/, [N: HTK], [plural: none]. Hyperonyms: toolkit. Synonyms: Hidden Markov Toolkit. Def.: A high-end commercial toolkit for building Hidden Markov Models (HMM).

human-computer interaction
/'hjuːmən kəm'pjuːtər ɪntə'rækʃən/, /'hju:m@n k@m'pju:t@r Int@'r{kS@n/, [N: [N: human][N: computer][N: interaction]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: interaction. Hyponyms: gesture-based interaction. Synonyms: HCI, man-machine interaction, MMI, human-machine interaction, HMI. Def.: Any interaction between a person and a computer. Some writers use human-computer dialogue as a synonym for HCI, while others use it to identify a subtype of HCI in which natural language is used as the primary or the only medium of communication..

human-human interaction
/'hjuːmən 'hjuːmən ɪntər'ækʃən/ , /'hju:m@n 'hju:m@n Int@r'{kS@n/ , [N: [N: human][N: human][N: interaction]], [plural: -s] . Hyperonyms: interaction. Synonyms: HHI. Def.: Any encounter between two (or more) people is a human-human interaction. Thus, a conversation is a human-human interaction. Human-human interactions are interesting to interactive dialogue technologists because of the light they may shed on human-computer interactions. However, a body of findings is being growing which shows that human-human and human-computer natural language dialogues differ systematically. Lessons for system design based on human-human dialogues must be interpreted in the light of these.

hybrid model
/ˈhaɪbrɪd ˈmɒdəl/, /'haIbrId 'mQd@l/, [N: [AJ: hybrid][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: coarticulation model. Cohyponym: look-ahead model, time-locked model, expansion model. Def.: The hybrid model combines aspects of the look-ahead and the time-locked model. Coarticulation effects occur in two phases of influence. The first phase starts as predicted by the look-ahead model, while the second phase begins at a locked time. In the first phase the appearance of movement due to the influence of a certain vowel is slow and in the next phase the appearance of movement is faster.

hypercardioid microphone
/ˈhaɪpəkɑːdɪɔɪd ˈmaɪkrəfəʊn/, /'haIp@kA:dIOId 'maIkr@f@Un/, [N: [AJ: hypercardioid][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: unidirectional microphone. Cohyponym: cardioid microphone, supercardioid microphone. Def.: Hypercardioid microphones are least sensitive at 110 degrees off-axis, 12 db down at the sides and approximately 6 db down at the rear. (Gibbon et al. 1997, p. 304)

hyperonymy
/haɪpəˈrɒnəmi/, /haIp@'rQn@mi/, [N: hyperonymy], [plural: y/-ies]. Domain: lexicon. Hyperonyms: semantic relation. Cohyponym: hyponymy. Def.: A semantic relation in which the meaning of one word (the hyperonym or superordinate) is entailed by the meaning of another (the hyponym or subordinate). (Gibbon et al. 1997, p. 201) E.g. 'Book' is a hyperonym of 'manual' as the meaning of book is implied by the meaning of manual (in one of its meanings)..

hyponymy
/haɪˈpɒnəmi/ , /haI'pQn@mi/ , [N: hyponymy] , [plural: y/-ies] . Domain: lexicon. Hyperonyms: relation of function between lexical signs; semantic relation. Cohyponym: hyperonymy. Def.: A semantic relation in which the meaning of one word (the hyponym or subordinate) entails the meaning of another (the hyperonym or superordinate). (Gibbon et al. 1997, p. 201) E.g. 'Manual' is a hyponym of 'book' as the meaning of 'manual' implies the meaning of 'book'..

iconic gesture
/aɪˈkɒnɪk ˈdʒestʃə/, /aI'kQnIk 'dZestS@/, [N: [AJ: iconic][N: gesture]], [plural: -s]. Domain: Spoken Language Technolgy: multimodal systems. Hyperonyms: gesture. Cohyponym: deictic gesture, metaphoric gesture, symbolic gesture. Def.: Iconic gestures refer to objects, spatial relations, or actions by describing them visually using a representation which is familiar to everyone, similar to icons representing applications in graphic user interfaces.

iconic representation
/aɪˈkɒnɪk reprɪzenˈteɪʃən/, /aI'kQnIk reprIzen'teIS@n/, [N: [AJ: iconic][N: representation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: output modality representation. Synonyms: analogue representation. Cohyponym: linguistic representation, arbitrary representation, static-dynamic representation. Def.: 1. A representation in terms of an model in which values of variables vary along a continuous scale and correlate with the values of the continuous empirical variables they represent, as opposed to a digital or digitised representation, in which the continuous empirical variables are modelled by variables with values on a discrete scale. 2. A representation which is complementary to a symbolic linguistic representation, based on the particular physical characteristics of the object it represents. Image, sound, graphics and haptic devices may be used to give such a representation. A picture of a book may give information on the title of the book, the author, the collection, but it will not tell you who the book belongs to.

identification test
/aɪdentɪfɪˈkeɪʃən test/, /aIdentIfI'keIS@n test/, [N: [N: identification][N: test]] , [plural: -s] . Domain: speech synthesis. Hyperonyms: testing procedure. Hyponyms: off-line identification test, on-line identification test. Def.: Procedure by which the listener is asked to identify a speech stimulus in terms of some (closed or open) set of response alternatives (e.g. some or all of the phonemes in the language).

identity assignment
/aɪ'dentɪti ə'saɪnmənt/, `/aI'dentIti @'saInm@nt/`, [N: [N: identity][N: assignment]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: decision outcome. Def.: Decision outcome which consists in attributing an identity to an applicant speaker, in the context of speaker identification. For speaker classification, the term class assignment should be used instead.

idiolect
/'ɪdɪəlekt/, `/'IdI@lekt/`, [N: idiolect], [plural: -s]. Hyperonyms: language variety. Meronym. sup.: natural language. Def.: Idiolect is a term used in linguistics to refer to the linguistic system of an individual speaker - his personal dialect. (Crystal 1988) An idiolect is heterogeneous, i.e. a speaker switches between different codes (code-switching) and shifts between different styles of speaking (style-shifting).

idiom
/'ɪdɪəm/, `/'IdI@m/`, [N: idiom], [plural: -s]. Domain: lexicon. Hyperonyms: lexical unit. Hyponyms: pragmatic idiom, phrasal idiom. Def.: 1. Lexical unit larger than the word whose meaning is not derivable from the meanings of the composite words, which has restrictions on alternative word orders and possible component words, and may not be strictly grammatical. 2. A variety of a language (e.g. a dialect). E.g. 'Come to think of it, ...' in the meaning 'I just thought of another relevant point, namely ...'.

impostor
/ɪm'pɒstə/, `/Im'pQst@/`, [N: impostor], [plural: -s] . Domain: speaker recognition. Hyperonyms: applicant speaker. Hyponyms: intentional impostor, well-intentioned impostor. Synonyms: impersonator, usurper, usurpator.(Both terms are very rarely used.), non-registered speaker (in identification) . Cohyponym: registered speaker. Def.: In speaker identification: an applicant speaker who does not belong to the set of registered speakers. In speaker verification: a speaker whose real identity is different from his claimed identity. (Gibbon et al. 1997, p. 413)

infix
/'ɪnfɪks/, `/'InfIks/`, [N: infix], [plural: -es]. Domain: lexicon. Hyperonyms: affix. Cohyponym: prefix, suffix, interfix, superfix, circumfix. Meronym. sup.: word. Def.: Morph that is inserted into a stem. E.g. Latin 'iungere' vs. 'iugum'.

inflection
/ɪn'flekʃən/, `/In'flekS@n/`, [N: inflection], [pural: -s]. Domain: lexicon. Hyperonyms: morphological operation. Cohyponym: word formation. Meronym. sup.: morphology. Def.: Inflection is the branch of morphology which deals with markers of the relation of words to their contexts within sentences, on the basis of agreement (congruence), e.g. between subject and verb. (cf. also Gibbon et al. 1997, p. 214) E.g. The fish tastes good. The final 's' in 'tastes' indicates that the subject noun is '3rd Person Singular', while in The fish taste good, the lack of a final 's' indicates that the subject noun, which is 3rd person, is plural. .

inflectional affixation
/ɪn'flekʃənəl æfɪk'seɪʃən/, `/In'flekS@n@l {fIk'seIS@n/`, [N: [AJ: inflectional][N: affixation]], [plural: -s]. Domain: lexicon. Hyperonyms: affixation. Hyponyms: inflectional prefixation, inflectional suffixation. Cohyponym: derivational affixation. Def.: Morphological concatenation of a stem with a full set of inflectional affixes. (Gibbon et al. 1997, p. 215) E.g. English 'algorithm' + 's' = 'algorithms'; German 'ge' + 'segn' + 'et' + 'en' 'blessed' (plural participle or adjective). (Gibbon et al. 1997, p. 215).

input modality
/ˈɪnpʊt məˈdælɪti/, /'InpUt m@'d{lIti/, [N: [N: input][N: modality]], [plural: y/-ies]. Domain: multimodal systems. Hyperonyms: modality. Hyponyms: non-speech input modality. Cohyponym: output modality. Def.: A human sensory input channel used in communication. In human-machine communication, when input from a system the input modalities are generally acoustic (loudspeaker, headphones) or visual (monitor screen), but may be haptic or tactile (e.g. Braille tablets). The olfactory (smell) and gustatory (taste) senses are additionally used in human-human communication.

insertion
/ɪnˈsɜːʃən/, /In's3:S@n/, [N: insertion], [plural: -s]. Domain: speech recognition, language modelling. Synonyms: false alarm. Cohyponym: deletion, substitution. Def.: A response to a word that was not in the utterance presented to a speech recognition system (or false alarm).

Integrated Services Digital Network
/ˈɪntɪgreɪtɪd ˈsɜːvɪsɪz ˈdɪdʒɪtəl ˈnetwɜːk/, /'IntIgreItId 's3:vIsIz 'dIdZIt@l 'netw3:k/, [N: [AJ: Integrated][N: Services][AJ: Digital][N: Network]], [plural: none]. Hyperonyms: digital telephony. Synonyms: ISDN. Cohyponym: Global System for Mobile Communication, GSM; analogue telephony. Def.: A world-wide standard for digital telephony in fixed networks.

intensity
/ɪnˈtensɪti/, /In'tensIti/, [N: intensity], [plural: y/-ies]. Domain: physical characterisation. Hyperonyms: acoustic measure. Cohyponym: amplitude, fundamental frequency, F0. Def.: Intensity is power per unit area, or the way power is distributed in a space. Power itself is a measure of the rate at which energy is being expended ... Now it can be shown that intensity is proportional to the square of pressure. (Crystal 1988, p. 223)

intentional impostor
/ɪnˈtenʃənəl ɪmˈpɒstə/, /In'tenS@n@l Im'pQst@/, [N: [AJ: intentional][N: impostor]], [plural: -s]. Domain: speaker recognition. Hyperonyms: impostor. Hyponyms: acquainted impostor, unacquainted impostor. Cohyponym: well-intentioned impostor. Def.: An intentional impostor has the clear goal of being identified or verified though he is not registered (violation), or to be identified as somebody else (usurpation). (Gibbon et al. 1997, p. 422)

interaction
/ɪntərˈækʃən/, /Int@r'{kS@n/, [N: interaction], [plural: -s]. Hyperonyms: communication. Def.: Communication of information between two agents, in which (except for the special case of the initial turn) an agent's contribution at any given point can be construed as a response to the previous turn or turns. A signal or a stimulus coming from one agent provokes a change in the internal state(s) of, and a response(s) from, the other agent. Stimuli are understood in a broad sense, including multimodal stimuli (in which different media may be used). They may, for example, consist of a physical action (moving a pointer) or a linguistic act (uttering a sentence) or the coordination of both.

interactive dialogue system
/ɪntərˈæktɪv ˈdaɪəlɒg ˈsɪstəm/ , /Int@r'{ktIv 'daI@lQg 'sIst@m/ , [N: [AJ: interactive][N: dialogue][N: system]], [plural: -s] . Domain: interactive dialogue systems. Hyperonyms: dialogue system. Synonyms: Interactive Voice Response; spoken language dialogue system. Cohyponym: command system. Def.: A computer system capable of engaging in turn-by-turn communication with a human user. In the general case, communication between the person and the system could use any communication mode or medium (or several simultaneously). In this chapter, however, the term is usually used more restrictively to apply to systems whose primary mode of communication is spoken natural language. This interactive communication is based on the integration of a set of modules, each of which handles a complex task. The modules are linked to each other and interactions are controlled by a kernel module which has the overall task of managing the dialogue. (Gibbon et al. 1997, page 567 / 568)

interactive voice response
/ɪntər'æktɪv 'vɔɪs rɪ'spɒns/, /Int@r'{ktIv 'vOIs rI'spQns/, [N: [AJ: interactive][N: voice][N: response]], [plural: -s] . Domain: interactive dialogue systems, system design. Hyperonyms: dialogue system. Synonyms: IVR, interactive dialogue. Def.: Interactive Voice Response (IVR) is what the commercial world calls interactive dialogue. As such, its scope encompasses certain kinds of simple interaction which research scientists do not normally think of as dialogues. For example, a telephone caller calling a weather forecasting Audiotex service might be asked to say one of the words "today", "tomorrow" or "weekend". In the basis of what the system recognises, a canned weather forecast will be played. This is an example of IVR which is also widely known as Voice Response (VR), and a system which supports VR is usually known as a Voice Response Unit (VRU).

interfix
/'ɪntəfɪks/, /'Int@fIks/, [N: interfix], [plural: -es]. Domain: lexicon. Hyperonyms: affix. Cohyponym: prefix, suffix, infix, superfix, circumfix. Def.: A morph with connective function in word formation by compounding, inserted between the modifier and the head of a compound noun; in German 'Fugenelement'. However, the situation is more complex, because stem modification of the modifier, following inflectional rules for plural, may occur. E.g. German 'Sonntag' + infix 's' + 'Spaziergang' = 'Sonntagsspaziergang'; 'Mann' + stem-modification + interfix 'er' + 'Bekleidung' = 'Maennerbekleidung'.

interjection
/ɪntə'dʒekʃən/, /Int@'dZekS@n/, [N: interjection], [plural: -s]. Domain: lexicon. Hyperonyms: grammatical category. Cohyponym: conjunction, article, preposition, pronoun. Def.: A grammatical or function word belonging to closed class of words and word-like units whose distribution is independent of the rules of sentence grammar, and which have a controlling or expressive function in discourse. E.g. oh, wow, gee, uhuh, ouch.

interpolation
/ɪntəpə'leɪʃən/, /Int@p@'leIS@n/, [N: interpolation], [plural: -s]. Domain: language modelling. Hyperonyms: smoothing technique. Synonyms: linear interpolation, smoothing. Cohyponym: extrapolation. Meronym. sup.: language model smoothing. Def.: 1. Interpolation is the estimation of an unknown value of a function between two known values. 2. In the context of language model smoothing the relative frequencies of a specific model are interpolated with those of a more general model.

interpretative property
/ɪn'tɜːprɪtətɪv 'prɒpəti/, /In't3:prIt@tIv 'prQp@ti/, [N: [AJ: interpretative][N: property]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: property of a lexical sign. Cohyponym: compositional property, structural property. Def.: Interpretative properties of a sign are (a) surface properties (phonological and orthographic) and (b) meaning properties (semantic and pragmatic representation), which interpret the representation of the structure of a sign in terms of (a) the real world of acoustic signals (speech sounds) or visual signals (writing, gesture) and (b) the real world of meaning. (Gibbon et al. 1997, p. 194)

interpreted programming language
/ɪn'tɜːprɪtɪd 'prəʊgræmɪŋ 'læŋgwɪdʒ/, /In't3:prItId 'pr@Ugr{mIN 'l{NgwIdZ/, [N: [AJ: interpreted][AJ: programming][N: language]], [plural: -s]. Hyponyms: scripting language; Perl, Prolog, LISP, Java. Cohyponym: compiled programming language. Def.: A programming language which is not completely translated into the byte code of a processor before execution but which is parsed by the processor (the interpreter) during execution. Interpreted languages may be translated into an optimised format before interpretation, and are often provided additionally with a full compiler. Interpretation permits incremental development of programs without recompilation, and permits a simple form of rapid prototyping.

interpreting telephony
/ɪn'tɜːprɪtɪŋ tə'lefəni/, /In't3:prItIN t@'lef@ni/, [N: [AJ: interpreting][N: telephony]], [plural: none]. Domain: speech synthesis. Hyperonyms: speech output system. Def.: Speech translation via telephony. A spoken utterance in one language is decomposed into its linguistic message and its speaker specific properties. The linguistic message is converted to text and transmitted. At the receiver end the text is automatically translated into another language and then converted back to speech.

IPP
/'aɪ 'piː 'piː/, /'aI 'pi: 'pi:/, [N: IPP], [plural: none]. Domain: multimodal systems. Hyperonyms: multimodal WWW user interface. Def.: IPP is an example of an asynchronous system. Different applications of IPP can be executed at any time as well. IPP accepts as input: text, mouse deixis, and speech. The possible outputs are synthesised speech, text, graphs, and map displays. The choice of output modalities is decided based on the user's requests.

ISA hierarchy
/'ɪzə 'haɪərɑːki/, /'Iz@ 'haI@rA:ki/, [N: [AJ: ISA][N: hierarchy]], [plural: y/-ies]. Domain: terminology. Hyperonyms: hierarchy. Synonyms: taxonomy, logical concept hierarchy, generic concept hierarchy. Cohyponym: PARTOF hierarchy, meronomy, mereonomy, partitive hierarchy, ontological hierarchy. Def.: A hierarchy defined by the relation of generalisation and its inverse, specialisation.

ISA relation
/'ɪzə rɪ'leɪʃən/, /'Iz@ rI'leIS@n/, [N: [AJ: ISA][N: relation]], [plural: -s]. Domain: terminology. Hyperonyms: lexical relation. Cohyponym: PARTOF relation. Def.: The term is rather general, and covers relations which have been referred to in other formalisms and theoretical frameworks with terms such as: paradigmatic relation, classification, taxonomy, field, family, similarity, set partition, subset-set inclusion, element-set membership, generalisation, property, implication, inheritance. Typical ISA relations define, in phonology, the natural classes characterised by distinctive feature vectors or by distributional classes based on syllable or word positions; in morphology, affix and stem classes; in phrasal syntax, parts of speech and constituent categories; in semantics, synonym, antonym and hyponym sets, or semantic fields.

ISDN
/'aɪ 'es 'diː 'en/, /'aI 'es 'di: 'en/, [N: ISDN], [plural: none]. Hyperonyms: standard. Synonyms: Integrated Services Digital Network. Cohyponym: Global System for Mobile Communication (GSM). Def.: A world-wide standard for digital telephony in fixed networks.

isolated word speech recognition system
/'aɪsəleɪtɪd 'wɜːd 'spiːtʃ rekəg'nɪʃən 'sɪstəm/, /'aIs@leItId 'w3:d 'spi:tS rek@g'nIS@n 'sIst@m/, [N: [AJ: isolated][N: word][N: speech][N: recognition][N: system]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: speech recognition system. Cohyponym: connected word recognition system, continuous speech recognition system. Def.: An isolated word recognition system can only recognise speech units (words or fixed expressions) that are separated by (possibly tiny) pauses.

isolated word speech recognition
/'aɪsəleɪtɪd 'wɜːd 'spiːtʃ rekəg'nɪʃən/, /'aIs@leItId 'w3:d 'spi:tS rek@g'nIS@n/, [N:[AJ: isolated][N: word][N: speech][N: recognition]], [plural: -s]. Hyponyms: word spotting. Cohyponym: connected word speech recognition, continous speech recognition. Def.: The most elementary form of speech recognition, in which input words or fixed expressions are separated by (possibly tiny) pauses.

isolated word
/'aɪsəleɪtɪd 'wɜːd/, /'aIs@leItId 'w3:d/, [N: [AJ: isolated][N: word]], Domain: speech recognition. Hyperonyms: speech style. Cohyponym: connected word, continuous speech. Def.: A style of speech where the words (or small phrases) are uttered separately, with small pauses in between.

iterative design
/ˈɪtərətɪv dɪˈzaɪn/, /'It@r@tIv dI'zaIn/, [N: [AJ: iterative][N: design]], [plural: none]. Hyperonyms: experimental technique. Synonyms: rapid prototyping. Cohyponym: benchmark evaluation, user study, simulation study. Def.: Iterative design has been widely adopted in the field of human-computer interaction, especially for product development. It is suitable for the development of multimodal applications, since many detail implementation issues can be explored rather quickly. The iterative design cycle includes (re)design of the application, implementation, and (informal) user testing. Iterative design is highly desirable from the HCI point of view but is difficult to reconcile with the pipeline or cascaded process organisation in software development which is currently still predominant, for reasons of cost control mainly.

IVR
/ˈaɪ ˈviː ˈɑː/, /'aI 'vi: 'A:/, [N: IVR], [plural: -s] . Domain: interactive dialogue systems, system design. Hyperonyms: dialogue system. Synonyms: interactive voice response, interactive dialogue. Def.: Interactive Voice Response (IVR) is what the commercial world calls interactive dialogue. As such, its scope encompasses certain kinds of simple interaction which research scientists do not normally think of as dialogues. For example, a telephone caller calling a weather forecasting Audiotex service might be asked to say one of the words "today", "tomorrow" or "weekend". In the basis of what the system recognises, a canned weather forecast will be played. This is an example of IVR which is also widely known as Voice Response (VR), and a system which supports VR is usually known as a Voice Response Unit (VRU).

jacknife method
/ˈdʒæknaɪf ˈmeθəd/, /'dZ{knaIf 'meT@d/, [N: [N: jacknife][N: method]], [plural: -s]. Hyperonyms: testing method. Def.: A method of testing a speech recognition system, for example, in which part of the data is kept aside for testing and the rest used for training, and the part kept aside is rotated until all the data has been used for testing.

Java
/ˈdʒɑːvə/, /'dZA:v@/, [N: Java], [plural: none]. Hyperonyms: object-oriented programming language. Def.: An object-oriented programming language developed by Javasoft of SUN Microsystems. It has become the de facto standard programming language for applets, i.e. programs which are distributed over the WWW to run inside a WWW browser. Java features a large class library including classes for graphical display, and audio data access. The Java Speech API specification supports voice command recognition, dictation, and text-to-speech synthesis.

jitter
/ˈdʒɪtə/, /'dZIt@/, [N: jitter], [plural: -s]. Domain: physical characterisation. Hyperonyms: physical characteristic; frequency modulation. Cohyponym: wow. Def.: Jitter is a form of frequency modulation of a signal by noise. The term is used as a measure of the average perturbation of a speaker's fundamental frequency and of its magnitude.

judgment testing
/ˈdʒʌdʒmənt ˈtestɪŋ/ , /'dZVdZm@nt 'testIN/ , [N: [N: judgment][N: testing]] , [plural: none] . Domain: speech synthesis. Hyperonyms: testing procedure. Synonyms: opinion testing. Cohyponym: functional testing. Def.: Procedure whereby a group of listeners is asked to judge the performance of a speech output system along a number of rating scales.

J_ToBI
/ˈdʒeɪ ˈtəʊbi/, /'dZeI 't@Ubi/, [N: [N: J_ToBI]], [plural: none]. Domain: corpora. Hyperonyms: ToBI . Cohyponym: E_ToBI, GlaToBI, G_ToBI. Def.: J_ToBI is a variant of ToBI developed for the transcription of Standard (Tokyo) Japanese.

keyword detection
/ˈkiːwɜːd dɪˈtekʃən/, /'ki:w3:d dI'tekS@n/, [N: [N: keyword][N: detection]], [plural: -s]. Synonyms: word spotting. Def.: The detection of a given word or word sequence in a stream of speech or text.

knowledge-based speech recognition system
/nɒlɪdʒ'beɪst 'spiːtʃ rekəg'nɪʃən 'sɪstəm/, /nQlIdZ'beIst 'spi:tS rek@g'nIS@n 'sIst@m/, [N: [N: knowledge][AJ: based][N: speech][N: recognition][N: system]], [plural: -s]. Domain: speech recognition, corpora. Hyperonyms: speech recognition system. Cohyponym: stochastic speech recognition system. Def.: A knowledge-based speech recognition system specifies explicit acoustic-phonetic rules that are robust enough to allow recognition of linguistically meaningful units and that ignore irrelevant variation in these units. (Gibbon et al. 1997, p. 94)

labiodental consonant
/leɪbɪəʊ'dentəl 'kɒnsənənt/, /leIbI@U'dent@l 'kQns@n@nt/, [N: [AJ: labiodental][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, dental consonant, alveolar consonant, postalveolar consonant, retroflex consonant, palatal consonant, velar consonant, uvular consonant, pharyngeal consonant, glottal consonant. Def.: A labiodental consonant is a consonant classified on the basis of its place of articulation: a sound in which one lip is actively in contact with the teeth. (Crystal 1988, p. 172)

laboratory room
/lə'bɒrətri 'rʊm/, /l@'bQr@tri 'rUm/, [N: [N: laboratory][N: room]], [plural: -s]. Domain: physical characterisation. Hyperonyms: recording room. Cohyponym: soundproof booth, recording studio, anechoic chamber. Def.: Speech recordings in typical laboratory environments are sometimes made in a kind of workbench situation when no special recording facility is available. Recordings made in a laboratory environment are often used to test speech recognition systems, as lab speech recordings seem to reflect best natural speech recognition situations, without requiring too much effort concerning the recording setup. (Gibbon et al. 1997, p. 309)

laboratory testing
/lə'bɒrətri 'testɪŋ/, /l@'bQr@tri 'testIN/, [N: [N: laboratory][N: testing]], [plural: none] . Domain: speech synthesis. Hyperonyms: testing procedure. Cohyponym: field testing. Def.: Speech output test procedure entirely run in a laboratory, either abstracting from in vivo complications or trying to simulate real-life situations.

lamb
/'læm/, /'l{m/, [N: lamb], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker. Synonyms: vulnerable speaker. Cohyponym: resistant speaker. Def.: A speaker with a high mistrust rate. (Gibbon et al. 1997, p. 433)

language disorder
/'læŋgwɪdʒ dɪs'ɔːdə/, /'l{NgwIdZ dIs'O:d@/, [N: [N: language][N: disorder]], [plural: -s]. Domain: corpora. Hyperonyms: speech disorder. Synonyms: aphasia. Cohyponym: articulation disorder, resonance disorder, voice disorder, rhythm disorder. Def.: A language disorder is one which does not affect only the production of the speech message, but rather its content. (Gibbon et al. 1997, p. 115)

language model
/ˈlæŋgwɪdʒ ˈmɒdəl/, /'l{NgwIdZ 'mQd@l/, [N: [N: language][N: model]], [plural: -s] . Domain: language modelling. Hyperonyms: model; knowledge source (in a speech recognition system); formal representation (of the structure of a natural language). Hyponyms: uniform language model; finite state language model; stochastic language model, grammar based language model; bigram language model, trigram language model, n-gram language model. Cohyponym: acoustic-phonetic model. Meronym. sup.: automatic speech recognition system. Def.: A formal representation of the structure (usually the sentence structure) of a natural language. In most speech recognisers language models consist of an n-gram model, i.e. probabilities of the occurrence of words, words pairs, sequences of three words, etc. In NLP language models tend to consist of formal grammars, which are sometimes enriched by statistical information. A language model in speech recognition is used to improve the recognition accuracy. Its task is to capture the redundancy inherent to the word sequences to be recognised. This redundancy may result from both the task specific constraints and general linguistic constraints. (Gibbon et al. 1997, p. 845)

language variety
/ˈlæŋgwɪdʒ vəˈraɪəti/, /'l{NgwIdZ v@'raI@ti/, [N:[N: language][N: variety]], [plural: y/-ies]. Domain: . Hyponyms: dialect, idiolect, register, sociolect, sublanguage, vernacular. Def.: A language variety is one of a family of reasonably homogeneous situation-dependent codes used by a group of speakers and recognised by them to belong to a single language. Language varieties can be defined in terms of their location in a variety space with three main dimensions: regional variation (dialect), social variation (sociolect), functional variation (register, style, sublanguage, technical language). In general, systems are designed for speaker-dependent or speaker-independent performance within one language variety only.

language
/ˈlæŋgwɪdʒ/, /'l{NgwIdZ/, [N: language], [plural: -s]. Domain: . Hyponyms: formal language, natural language. Def.: 1. Language is the faculty underlying the human ability to communicate by means of intricate sound patterns. 2. A language is a system of signs defined by a syntax (a vocabulary and a set of combinatory rules), a semantics (a set of rules for assigning meanings to vocabulary elements and their combinations), and a pragmatics (a set of conventions for using the vocabulary elements and their combinations in specific situations).

laryngograph
/ləˈrɪŋgəgrɑːf/, /l@'rINg@grA:f/, [N: laryngograph], [plural: -s]. Domain: corpora, physical characterisation. Hyperonyms: device. Synonyms: electroglottograph. Def.: A laryngograph is a device for measuring the high frequency impedance across the glottis during phonation; the resulting signal (known as Lx) is closely related to the rate of articulation (in the articulatory domain), the fundamental frequency (in the acoustic domain), and the impression of pitch (in the auditory domain). In general, Lx is a closer and more noise-free approximation to phonetially relevant patterns than acoustically measured measurement of fundamental frequency. Laryngograph is a proprietary term.

larynx
/ˈlærɪŋks/, /'l{rINks/, [N: larynx], Domain: corpora, physical characterisation. Hyperonyms: articulator. Def.: Larynx is the part of the windpipe containing the vocal cords. (Crystal 1988, p. 174)

late integration
/ˈleɪt ɪntɪˈgreɪʃən/, /'leIt IntI'greIS@n/, [N:L [AJ: late][N: integration]], [plural: -s]. Domain: multimodal systems. Cohyponym: early integration. Def.: Integration of audio and visual information in HMMs where a first decision based on the separate signals is taken and then the final decision based on the combination of both results is made.

lateral
/'lætərəl/, **/'l{t@r@l/**, [N: lateral], [plural: -s]. Hyperonyms: consonant. Def.: Lateral is a term used in phonetic classification of consonant sounds on the basis of their manner of articulation: it refers to any sound where the air escapes around one or both sides of a closure made in the mouth. E.g. Example in English: [l].

learning effect
/'lɜːnɪŋ ɪ'fekt/, **/'l3:nIN I'fekt/**, [N: [N: learning][N: effect]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Def.: The effect that a first test for a naive subject is always harder than later tests, because the subject learns how to deal with the system, what kind of events to expect, etc.

leaving-one-out
/'liːvɪŋ 'wʌn 'aʊt/, **/'li:vIN 'wVn 'aUt/**, [N: [V: leaving][DET: one][PREP: out]], [plural: none]. Domain: language modelling. Hyperonyms: cross-validation. Def.: Leaving-one-out is a special kind of cross-validation where no additional test set is needed. Instead, it is generated from the training observations by leaving out one observation at a time.

left-to-right coarticulation
/'left tə 'raɪt kəʊɑːtɪkjʊ'leɪʃən/, **/'left t@ 'raIt k@UA:tIkjU'leIS@n/**, [N: [AJ: left][PREP: to][AJ: right][N: coarticulation]], [plural: none]. Hyperonyms: coarticulation. Synonyms: perseverative coarticulation, forward coarticulation. Cohyponym: backward coarticulation, anticipatory coarticulation, right-to-left coarticulation. Def.: In the string ...AB..., sound A influences sound B (or beyond). L>R coarticulation is thought to be largely due to lag in articulatory movement, induced by inertia. (Clark & Yallop, p. 87)

lemma
/'lemə/, **/'lem@/**, [N: lemma], [plural: lemmata]. Domain: lexicon. Cohyponym: abstract lemma, lexeme. Def.: 1. A lemma is a lexical access key, for which the canonical orthogrphy of a lexical entry is often used. 2. A headword under which variants of a lexical entry are listed. 3. An abstract lexical unit (abstract lemma).

lexical accent
/'leksɪkəl 'æksənt/, **/'leksIk@l '{ks@nt/**, [N: [AJ: lexical][N: accent]], [plural: -s]. Hyperonyms: accent. Synonyms: word accent. Cohyponym: phrase accent, sentence accent, syntactical accent, tonal accent. Def.: The emphasis which makes a particular (...) syllable [in a word] stand out (...). (Crystal 1988, p. 2)

lexical category
/'leksɪkəl 'kætəgəri/, **/'leksIk@l 'k{t@g@ri/**, [N: [AJ: lexical][N: category]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: part of speech. Cohyponym: grammatical category. Def.: Lexical categories are the parts of speech Noun, Adjective, Verb, Adverb, i.e. open classes which may be extended by morphological rules of word formation. (Gibbon et al. 1997, p. 218)

lexical database
/'leksɪkəl 'deɪtəbeɪs/ , **/'leksIk@l 'deIt@beIs/** , [N: [AJ: lexical][N: database]], [plural: -s]. Domain: lexicon. Hyperonyms: spoken language lexicon. Cohyponym: system lexicon. Def.: A lexical database is a set of more or less loosely related simpler databases (e.g. pronunciation table, index into a signal annotation file database, stochastic word model, linguistic lexical database with syntactic and semantic information). (Gibbon et al. 1997, p. 185)

lexical information model
/'leksɪkəl ɪnfə'meɪʃən 'mɒdəl/, **/'leksIk@l Inf@'meIS@n 'mQd@l/**, [N: [AJ: lexical][N: information][N: model]], [plural: -s]. Domain: lexicon. Hyperonyms: model. Def.: A model based on specific types of lexical information and lexical objects.

lexical item
/'leksɪkəl 'aɪtəm/, /'leksIk@l 'aIt@m/, [N:[AJ: lexical][N: item]], [plural: -s]. Domain: . Hyponyms: idiom, actual word, heterograph, homograph, homohym, homophone, potential word, synonym . Synonyms: lexical entry. Def.: A unit of language such as a word or an idiom which is inventarised in a lexicon.

lexical morph
/'leksɪkəl 'mɔːf/, /'leksIk@l 'mO:f/, [N: [AJ: lexical][N: morph]], [plural: -s]. Domain: lexicon. Hyperonyms: morph. Hyponyms: phonological lexical morph, grammatical lexical morph. Cohyponym: grammatical morph. Def.: Orthographic or phonological realisation of a lexical morpheme.

lexical morpheme
/'leksɪkəl 'mɔːfiːm/, /'leksIk@l 'mO:fi:m/, [N: [AJ: lexical][N: morpheme]], [plural: -s]. Domain: lexicon. Hyperonyms: morpheme. Cohyponym: grammatical morpheme. Def.: 1. A lexical morpheme is characterised by membership of a large, potentially open class, with meanings such as properties and roles of objects, states and events. (Gibbon et al. 1997, p. 214) 2. Indivisable word stem. (Gibbon et al. 1997, p. 196)

lexical object
/'leksɪkəl 'ɒbdʒekt/, /'leksIk@l 'QbdZekt/, [N: [AJ: lexical][N: object]], [plural: -s]. Domain: lexicon. Hyponyms: lexical sign class, archi-sign; lexicon, lexical database. Def.: 1. The basic object (such as a morpheme, a stem, a word or an idiom) described in a lexicon. 2. The lexical sign class or archi-sign in which similar lexical objects are grouped together, each characterised by subsets of the lexical information required to characterise specific lexical signs. (Gibbon et al. 1997, p. 192)

lexical sign class
/'leksɪkəl 'saɪn 'klɑːs/, /'leksIk@l 'saIn 'klA:s/, [N: [AJ: lexical][N: sign][N: class]], [plural: -es]. Domain: lexicon. Hyperonyms: lexical object. Synonyms: archi-sign. Def.: Group of similar lexical objects, each characterised by subsets of the lexical information required to characterise specific lexical signs. (Gibbon et al. 1997, p. 192)

lexical sign
/'leksɪkəl 'saɪn/, /'leksIk@l 'saIn/, [N: [AJ: lexical][N: sign]], [plural: -s]. Domain: lexicon. Hyperonyms: lexical object. Def.: A sign which is inventarised in a lexicon and whose properties are not (or at least not wholly) composed from those of constituent signs. A lexical sign has properties, often represented as attribute-value pairs, and known as lexical information. The microstructure of a lexicon is the structure associated with the types of lexical information assigned to a lexical sign. (Gibbon et al. 1997, p. 192)

lexical unit
/'leksɪkəl 'juːnɪt/, /'leksIk@l 'ju:nIt/, [N: [AJ: lexical][N: unit]], [plural: -s]. Domain: lexicon. Hyperonyms: speech unit. Hyponyms: word, idiom, discourse particle, hesitation, pragmatic idiom, functional unit. Synonyms: lexical item. Def.: A unit of language such as a word or an idiom which is inventarised in a lexicon.

lexicon architecture
/'leksɪkən 'ɑːkɪtektʃə/, /'leksIk@n 'A:kItektS@/, [N: [N: lexicon][N: architecture]], [plural: -s]. Domain: lexicon. Hyperonyms: structure; linguistic characterisation. Synonyms: lexicon macrostructure. Def.: The choice of basic objects and properties in the lexicon, and the macrostructure of the lexicon as a whole, such as a table of items, a trie (decision tree), an inheritance hierarchy, a semantic network, a database.

lexicon formalism
/'leksɪkən 'fɔːməlɪzəm/, /'leksIk@n 'fO:m@lIz@m/, [N: [N: lexicon][N: formalism]], [plural: -s]. Domain: lexicon. Hyperonyms: formal language. Cohyponym: lexicon theory, lexicon model, linguistic framework. Def.: A specially designed logic programming language such as DATR, or an algebraic formalism such as attribute-value matrices, or appropriate definitions in high level languages such as LISP or Prolog, with compiler concepts for translating these languages into conventional languages for efficient processing. (Gibbon et al. 1997, p. 192)

lexicon model
/'leksɪkən 'mɒdəl/, /'leksIk@n 'mQd@l/, [N: [N: lexicon][N: model]], [plural: -s]. Domain: lexicon. Hyperonyms: model; linguistic characterisation. Cohyponym: lexicon theory, lexicon formalism, linguistic framework. Def.: A lexicon model is the specification of the domain denoted by a lexicon theory, conceptually independent of the theory itself. A different definition is also common: the general structure of the objects and attribute-value structures in a formal lexicon. A lexicon model specifies the following kinds of information: - Types of lexical object and structure of lexical entries. - Types of lexical information associated with lexical objects in lexical entries. - Relations between lexical objects and structure of the lexicon as a whole lexicon architecture. (Gibbon et al. 1997, p. 193)

lexicon structure
/'leksɪkən 'strʌktʃə/, /'leksIk@n 'strVktS@/, [N: [N: lexicon][N: structure]], [plural: -s]. Domain: lexicon. Hyponyms: lexicon macrostructure, lexicon microstructure. Synonyms: lexicon architecture. Def.: The organisation of information in lexica in terms of macrostructure (the overall structure in which lexical entries are located) and microstructure (the internal structure of a lexical entry). (Gibbon et al. 1997, p. 221)

lexicon theory
/'leksɪkən 'θɪəri/, /'leksIk@n 'TI@ri/, [N: [N: lexicon][N: theory]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: theory; linguistic characterisation. Hyponyms: general lexicon theory, specific lexicon theory. Cohyponym: lexicon formalism, lexicon model, linguistic framework. Def.: A coherent and consistent set of expressions formulated in a well-defined formalism and interpreted with respect to a lexicon model. (Gibbon et al. 1997, p. 192)

lexicon
/'leksɪkən/, /'leksIk@n/, [N: lexicon], [plural: -s, lexica]. Domain: . Hyperonyms: lexical object, . Hyponyms: declarative lexicon, fully inflected form lexicon, morph lexicon, morpheme lexicon, procedural lexicon, pronunciation lexicon, prosodic lexicon, stem lexicon, . Cohyponym: grammar. Def.: A collection of lexical items such as words or idioms organised within a lexical macrostructure (a list, a tree structure, etc.), each of which has a regular microstructure to which types of lexical information are assigned.

linear discriminant analysis
/'lɪnɪə dɪs'krɪmɪnənt ə'næləsɪs/, /'lInI@ dIs'krImIn@nt @'n{l@sIs/, [N: [AJ: linear][AJ: discriminant][N: analysis]], [plural: linear discriminant analyses]. Hyperonyms: statistical technique. Def.: A statistical technique that under certain assumptions finds a linear transformation that will best separate a set of classes when the classification decision is the identification of the closest class centroid measured using Euclidean distances.

linear interpolation
/'lɪnɪər ɪntɜːpə'leɪʃən/, /'lInI@r Int3:p@'leIS@n/, [N: [AJ: linear][N: interpolation]], [plural: -s]. Domain: language modelling. Hyperonyms: interpolation, smoothing technique. Cohyponym: linear discounting, absolute discounting. Def.: 1. Interpolation of unknown values of a linear function y=c+mx between two known values. 2. A technique in the context of language model smoothing by which the relative frequencies of a specific model are interpolated with those of a more general model.

linear predictive coding
/ˈlɪnɪə prɪˈdɪktɪv ˈkəʊdɪŋ/, **/’lInI@ prI’dIktIv ’k@UdIN/**, [N: [AJ: linear][AJ: predictive][N: coding]], [plural: none]. Domain: corpora, speech synthesis. Hyperonyms: signal processing technique. Synonyms: LPC. Def.: A signal processing technique used in speech coding and in speech analysis (for speech recognition, for example). The technique assumes that the speech signal is generated by an autoregressive process, that is, by an all-pole filter equivalent to a series-resonant circuit. A given sample is represented as (predicted by) a weighted sum of past samples, the weights being the coefficients. The order of prediction is the number of coefficients; an order of 8-10 is common.

Lingua Franca
/ˈlɪŋgwə ˈfræŋkə/, **/’lINgw@ ’fr{Nk@/**, [N:[N: Lingua][N: Franca]], [plural: none]. Cohyponym: vernacular. Def.: A common language used by speakers of different language for practical professional communication purposes; nowadays frequently, but not necessarily, English.

linguistic interface
/lɪŋˈgwɪstɪk ˈɪntəfeɪs/, **/lIN’gwIstIk ’Int@feIs/**, [N: [AJ: linguistic][N: interface]], [plural: -s] . Domain: speech synthesis. Cohyponym: acoustic interface. Meronym. sup.: text-to-speech system. Meronym. sub.: text preprocessing, grapheme-phoneme conversion, word stress assignment, sentence accent assignment, boundary position assignment, intonation pattern assignment. Def.: First part of a text-to-speech system, which parses the input text and transforms spelling into an abstract phonological code (which in turn is converted to sound by the acoustic interface), adding onformation about timing and pitch contours.

linguistic representation
/lɪŋˈgwɪstɪk reprɪzenˈteɪʃən/, **/lIN’gwIstIk reprIzen’teIS@n/**, [N: [AJ: linguistic][N: representation], [plural: -s]. Domain: multimodal systems. Hyperonyms: output modality representation. Cohyponym: analogue representation, iconic representation, arbitrary representation, static-dynamic representation. Def.: 1. A representation of a language sign in terms of phonological, morphological, syntactic, semantic, and/or pragmatic information. 2. In multimodal systems, a representation based on a high level of abstraction and not able to give relevant details to distinguish specific entities as would an analogue representation. The string ‘my book’ distinguishes a particular book from other books but it does not give any specific information, for example, title, author, size, and collection.

linguistics
/lɪŋgˈwɪstɪks/, **/lINg’wIstIks/**, [N: linguistics], [plural: none]. Hyponyms: discourse analysis, morphology, morphotactics, phonetics,phonology, pragmatics, psycholinguistics,semantics,sociolinguistics, syntax . Def.: 1. The scientific study of language. 2. The component of a system which handles linguistic objects such as spellings, grammar, lexicon.

lip smack
/ˈlɪp ˈsmæk/, **/’lIp ’sm{k/**, [N: [N: lip][N: smack]], [plural: -s]. Domain: corpora, system design. Hyperonyms: noise, non-linguistic phenomenon. Def.: A popping noise made by the lips, often resulting from an implosion (due to suction caused by inhalation) just before an utterance.

lipreading
/ˈlɪpriːdɪŋ/, **/’lIpri:dIN/**, [N: lipreading], [plural: none]. Domain: multimodal systems. Meronym. sup.: recognition process. Def.: Recognising spoken language from lip movement.

list
/ˈlɪst/, /’lIst/, [N: list], [plural: -s]. Domain: language modelling; lexicon. Def.: 1. A flat or linear data structure, generally implemented recursively with pointer pairs, the first of which points to the first element in the list, the second of which points to the rest of the list. 2. In language modelling, a list implementation works as follows: for each word we have a pointer into the actual list, and for each observed trigram, an entry of the list consists of two parts, namely the index pair and the trigram count. For efficiency, this organisation should be combined with the counts for the bigram models. (Gibbon et al. 1997, p. 256)

logical concept hierarchy
/ˈlɒdʒɪkəl ˈkɒnsept ˈhaɪərɑːki/, /’lQdZIk@l ’kQnsept ’haI@rA:ki/, [N: [AJ: logical][N: concept][N: hierarchy]], [plural: y/-ies]. Domain: terminology. Hyperonyms: hierarchy. Synonyms: generic concept hierarchy, taxonomy, ISA hierarchy. Cohyponym: meronomy, mereonomy, PARTOF hierarchy. Def.: Hierarchy of concepts holding an ISA relation, i.e. a hierarchy defined by the relation of generalisation and its inverse, specialisation.

logographic orthography
/lɒgəˈgræfɪk ɔːˈθɒgrəfi/, /lQg@’gr{fIk O:’TQgr@fi/, [N: [AJ: logographic][N: orthography]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: orthography. Cohyponym: alphabetic orthography, syllabic orthography. Def.: In logographic orthography, characters are closely related to simplex words. (Gibbon et al. 1997, p. 188) E.g. Chinese, arabic numerals..

Lombard effect
/ˈlɒmbɑːd ɪˈfekt/, /’lQmbA:d I’fekt/, [N: [N: Lombard][N: effect]], [plural: -s]. Hyperonyms: speech style. Def.: The effect that humans speak at a higher level (use more vocal effort) in conditions of higher environmental noise, resulting in measurable changes in many parameters of voice production, inluding timing, pitch and voice quality. The influence of the physical environment on speech production via acoustic feedback. (Gibbon et al. 1997, p. 83))

look-ahead model
/ˈlʊk əˈhed ˈmɒdəl/, /’lUk @’hed ’mQd@l/, [N: [V: look][AV: ahead][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: coarticulation model. Hyponyms: target-based model, feature-based model, goal-based model. Cohyponym: time-locked model, hybrid model, expansion model. Def.: In the look-ahead model, the influence of a vowel on segments does not start from a given time but rather from the last preceding vowel (in the case of forward coarticulation) or the following vowel (in the case of backward coarticulation). Three parameters influence sequences of consonants: targets, features, and goals. There are three corresponding variants of look-ahead models: the target-based model, the feature-based model, the goal-based model.

low-pass filter
/ˈləʊpɑːs ˈfɪltə/, /’l@UpA:s ’fIlt@/, [N: [AJ: low][V: pass][N: filter]], [plural: -s]. Domain: physical characterisation. Hyperonyms: filter. Cohyponym: high-pass filter, band-pass filter, band-stop filter, notch filter, all-pass filter. Def.: A low-pass filter removes or reduces the amplitude of high frequencies in a signal, i.e. it attenuates frequencies above a given threshold.

LPC
/ˈel ˈpiː ˈsiː/, /’el ’pi: ’si:/, [N: LPC], [plural: none]. Domain: corpora, speech synthesis. Hyperonyms: signal processing technique. Synonyms: linear predictive coding. Def.: A signal processing technique used in speech coding and in speech analysis (for speech recognition, for example). The technique assumes that the speech signal is generated by an autoregressive process, that is, by an all-pole filter equivalent to a series-resonant circuit.

macrotemporal fusion
/mɑːkrəʊˈtempərəl ˈfjuːʒən/, /mA:kr@U’temp@r@l ’fju:Z@n/, [N: [AJ: macrotemporal][N: fusion]], [plural: none]. Domain: multimodal systems. Hyperonyms: fusion. Cohyponym: microtemporal fusion, contextual fusion. Def.: Macrotemporal fusion combines sequential information units in temporal proximity when the information units are complementary.

magnitude estimation
/ˈmægnɪtjuːd estɪˈmeɪʃən/, /’m{gnItju:d estI’meIS@n/, [N: [N: magnitude][N: estimation]], [plural: -s] . Hyperonyms: rating method. Def.: Rating method where the subject is presented with an (auditory) stimulus and is asked to express the perceived strength/quality of the relevant attribute (e.g. intelligibility) numerically (“type in a value”) or graphically (“draw a line on the computer screen”).

MAUS
/ˈmaʊs/, /’maUs/, [N: MAUS], [plural: none]. Hyperonyms: segmentation tool, labelling tool. Def.: An automatic segmentation and labelling tool for speech verification. Its primary feature is a generator of pronunciation variants for a given utterance; these variants are stored as a hypothesis graph. A standard Viterbi alignment then finds the best path through the graph.

maximum approximation
/ˈmæksɪməm əprɒksɪˈmeɪʃən/, /’m{ksIm@m @prQksI’meIS@n/, [N: [N: maximum][N: approximation]], [plural: -s]. Domain: language modelling. Hyperonyms: alignment algorithm. Synonyms: Viterbi decoding, Viterbi approximation, Viterbi alignment. Meronym. sup.: HMM recogniser. Def.: A popular alignment algorithm that finds the best path through a probability graph. Maximum approximation is usually applied to HMM output.

MEDITOR
/ˈmedɪtə/, /’medIt@/, [N: MEDITOR], [plural: none]. Domain: multimodal systems. Def.: A text editor for blind people, able to perform the main actions as a normal text editor. The system uses four input devices: a speech recognition system, a braille keyboard, a normal keyboard, and a mouse. A braille display, a speech synthesis module and a screen are the output devices of the system.

medium
/ˈmiːdɪəm/, /’mi:dI@m/, [N: medium], [plural: media]. Domain: multimodal systems. Hyperonyms: physical device. Def.: Physical device to capture input from or present feedback to a human communication partner. E.g. microphone, keyboard, mouse, camera, text/image/video display, loudspeaker.

melting pot
/ˈmeltɪŋ ˈpɒt/, /’meltIN ’pQt/, [N: [AJ: melting][N: pot]], [plural: -s]. Domain: multimodal systems. Def.: A melting pot encapsulates types of structural parts of a multimodal event. The content of a structural part is a piece of time-stamped information. Melting pots are built from elementary input events by different fusion mechanisms: microtemporal, macrotemporal, and contextual fusion.

menu dialogue system
/ˈmenjuː ˈdaɪəlɒg ˈsɪstəm/ , /’menju: ’daI@lQg ’sIst@m/ , [N: [N: menu][N: dialogue][N: system]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: dialogue system. Def.: Human-system interaction is reduced to a question-answer user interface. The dialogue model is merged into the task model from which it cannot be distinguished. Dialogues of this kind are often represented by branching tree structures. (Gibbon et al. 1997, p. 570)

mereonomic relation
/merɪəˈnɒmɪk rɪˈleɪʃən/, /merI@'nQmIk rI'leIS@n/, [N: [AJ: mereonomic][N: relation]], [plural: -s]. Domain: terminology. Hyperonyms: semantic relation. Synonyms: meronomic relation, meronymic relation, PARTOF relation. Cohyponym: taxonomic relation, taxonymic relation, ISA relation. Def.: Fundamental syntactic or combinatorial PARTOF relation. Like ISA, the term is also rather general, and a wide range of different relations are covered by it in different approaches to linguistics in general and lexicography in particular: syntagmatic relations, mereological (merological) / mereonomic (meronomic) relations, part-whole relations, part-part relations, (immediate) constituency / domination, command relations (e.g. c-command), dependency relations, government relations, argument structure, thematic role structure, subcategorisation frames, case frames, valency, anaphoric binding relations, categorial functor-argument application, concatenation, linear ordering, prosodic (autosegmental) association and precedence relations, child-child (sister) relations, parent-child (mother-daughter) relations.

mereonomy
/merɪˈnɒnəmi/, /merI'Qn@mi/, [N: mereonomy], [plural: y/-ies]. Domain: terminology. Hyperonyms: hierarchy. Synonyms: meronomy, PARTOF hierarchy, partitive hierarchy, ontological hierarchy. Cohyponym: taxonomy, ISA hierarchy, logical concept hierarchy, generic concept hierarchy. Def.: A hierarchy defined by the relation of parts to wholes, and parts to parts.

meronomic relation
/merəˈnɒmɪk rɪˈleɪʃən/, /mer@'nQmIk rI'leIS@n/, [N: [AJ: meronomic][N: relation]], [plural: -s]. Domain: terminology. Hyperonyms: semantic relation. Synonyms: mereonomic relation, meronymic relation, PARTOF relation. Cohyponym: taxonomic relation, taxonymic relation, ISA relation. Def.: Fundamental syntactic or combinatorial PARTOF relation. Like ISA, the term is also rather general, and a wide range of different relations are covered by it in different approaches to linguistics in general and lexicography in particular: syntagmatic relations, mereological (merological) / mereonomic (meronomic) relations, part-whole relations, part-part relations, (immediate) constituency / domination, command relations (e.g. c-command), dependency relations, government relations, argument structure, thematic role structure, subcategorisation frames, case frames, valency, anaphoric binding relations, categorial functor-argument application, concatenation, linear ordering, prosodic (autosegmental) association and precedence relations, child-child (sister) relations, parent-child (mother-daughter) relations.

meronomy
/məˈrɒnəmi/, /m@'rQn@mi/, [N: meronomy], [plural: y/-ies]. Domain: terminology. Hyperonyms: hierarchy. Synonyms: mereonomy, PARTOF hierarchy, partitive hierarchy, ontological hierarchy. Cohyponym: taxonomy, ISA hierarchy, logical concept hierarchy, generic concept hierarchy. Def.: A hierarchy defined by the relation of parts to wholes, and parts to parts.

meronymic relation
/merəˈnɪmɪk rɪˈleɪʃən/, /mer@'nImIk rI'leIS@n/, [N: [AJ: meronymic][N: relation]], [plural: -s]. Domain: terminology. Hyperonyms: semantic relation. Synonyms: meronomic relation, mereonomic relation, PARTOF relation. Cohyponym: taxonomic relation, taxonymic relation, ISA relation. Def.: Fundamental syntactic or combinatorial relation. Like ISA, the term is also rather general, and a wide range of different relations are covered by it in different approaches to linguistics in general and lexicography in particular: syntagmatic relations, mereological (merological) / mereonomic (meronomic) relations, part-whole relations, part-part relations, (immediate) constituency / domination, command relations (e.g. c-command), dependency relations, government relations, argument structure, thematic role structure, subcategorisation frames, case frames, valency, anaphoric binding relations, categorial functor-argument application, concatenation, linear ordering, prosodic (autosegmental) association and precedence relations, child-child (sister) relations, parent-child (mother-daughter) relations.

metaphoric gesture
/metə'fɒrɪk 'dʒestʃə/, /met@'fQrIk 'dZestS@/, [N: [AJ: metaphoric][N: gesture]], [plural: -s]. Domain: Spoken Language Technolgy: multimodal systems. Hyperonyms: gesture. Cohyponym: deictic gesture, iconic gesture, symbolic gesture. Def.: Metaphoric gestures involve the manipulation of some abstract object or tool.

metrical phonology
/'metrɪkəl fə'nɒlədʒi/, /'metrIk@l f@'nQl@dZi/, [N: [AJ: metrical][N: phonology]], [plural: none]. Hyperonyms: phonological theory. Cohyponym: autosegmental phonology, generative phonology, phonemic phonology. Def.: A theory of prosodic phonology which analyses sequences of phonological units into bbinary branching trees; related to dependency phonology.

microphone
/'maɪkrəfəʊn/, /'maIkr@f@Un/, [N: microphone], [plural: -s]. Domain: physical characterisation. Hyperonyms: device. Hyponyms: unidirectional microphone, bidirectional microphone, omnidirectional microphone, ultradirectional microphone, pressure zone microphone, headset microphone; headmounted microphone, table-top microphone, handheld microphone, room microphone; dynamic microphone, condenser microphone. Def.: A transduction device which converts variations in air pressure into variations into electrical signals.

microtemporal fusion
/maɪkrəʊ'tempərəl 'fjuːʒən/, /maIkr@U'temp@r@l 'fju:Z@n/, [N: [AJ: microtemporal][N: fusion]], [plural: none]. Domain: multimodal systems. Hyperonyms: fusion. Cohyponym: macrotemporal fusion, contextual fusion. Def.: Microtemporal fusion combines information units that are produced simultaneously or very close in time.

MIME
/'maɪm/, /'maIm/, [N: MIME], [plural: none]. Hyperonyms: document descriptor. Synonyms: Multi-purpose Internet Mail Extension. Def.: A document descriptor that is associated with a document to identify its type and format.

misclassification
/mɪsklæsɪfɪ'keɪʃən/, /mIskl{sIfI'keIS@n/, [N: misclassification], [plural: -s]. Domain: speaker recognition. Hyperonyms: identity assignment. Def.: Erroneous identity assignment to a registered speaker in speaker identification.

mistaken speaker
/mɪs'teɪkən 'spiːkə/, /mIs'teIk@n 'spi:k@/, [N: [AJ: mistaken][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: registered speaker. Def.: The registered speaker owning the identity assigned erroneously to another registered speaker by a speaker identification system. (Gibbon et al. 1997, p. 414)

mixed-initiative dialogue
/'mɪkst ɪ'nɪʃətɪv 'daɪəlɒg/, /'mIkst I'nIS@tIv 'daI@lQg/, [N: [AJ: mixed][N: initiative][N: dialogue]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: dialogue. Def.: A type of human-machine dialogue control in which the initiative to bring up topics or information items can change between a system and a human. Specifically, the system is set up to process all relevant information offered by the human, irrespective of the phase the dialogue is in and irrespective of the precise prompt given by the system.

modality
/məʊ'dælətɪ/, /m@U'd{l@ti/, [N: modality], [plural: y/-ies]. Domain: multimodal systems. Hyperonyms: speech style. Def.: The way a communicating agent conveys information to a communication partner (human or machine). E.g. intonation, gaze, hand gestures, body gestures, facial expressions.

modeless operation
/ˈməʊdləs ɒpəˈreɪʃən/, /'m@Udl@s Qp@'reIS@n/, [N: [AJ: modeless][N: operation]], [plural: -s]. Hyperonyms: property of an automatic dictation system. Def.: A property of automatic dictation systems in which it is unnecessary for the user to make an explicit distinction between speech to be transcribed as text and spoken commands for editing, correction, formatting, etc.

monitoring
/ˈmɒnɪtərɪŋ/, /'mQnIt@rIN/, [N: monitoring], [plural: none]. Domain: corpora, speech synthesis. Hyponyms: phoneme monitoring, syllable monitoring, word monitoring. Synonyms: on-line monitoring. Cohyponym: validation. Def.: Monitoring is the task of controlling and modifying technical and phonetic characteristics of signals on-line, e.g. during the course of speaking or of a recording. (Gibbon et al. 1997, p. 129)

monomodal input event
/mɒnəʊˈməʊdəl ˈɪnpʊt ɪˈvent/ , /mQn@U'm@Ud@l 'InpUt I'vent/ , [N: [AJ: monomodal][N: input][N: event]], [plural: -s]. Domain: multimodal systems. Hyperonyms: input event. Synonyms: unimodal input event. Cohyponym: multimodal input event. Def.: Set of user input events that belong together and are intended to convey a chunk of information, consisting of at least two parts in one modality.

mood identification
/ˈmuːd aɪdentɪfɪˈkeɪʃən/, /'mu:d aIdentIfI'keIS@n/, [N: [N: mood][N: identification]], [plural: none]. Domain: speaker recognition. Hyperonyms: speaker classification task. Cohyponym: sex identification, age identification, health state identification, accent identification, speaker cluster selection. Def.: A task that consists in determining whether a speaker is angry, sad, stressed, calm, happy, relaxed, etc. (Gibbon et al. 1997, p. 408)

morph
/ˈmɔːf/, /'mO:f/, [N: morph], [plural: -s]. Domain: lexicon. Hyperonyms: morphological entity. Hyponyms: phonological morph, orthographic morph; affix, root; free morph, bound morph.. Cohyponym: morpheme. Def.: In traditional linguistics: the orthographic or phonological form (realisation) of a morpheme. (Gibbon et al. 1997, p. 215) E.g. In English: 's' /s/, /z/ and 'es' /Iz/ are morphs of the 3rd P. Sg. Present Tense morpheme: 'eats' /i:t/ + /s/;'sings' /sIN/ + /z/; 'catches' /k{tS/ + /Iz/.

morph lexicon
/ˈmɔːf ˈleksɪkən/, /'mO:f 'leksIk@n/, [N: [N: morph][N: lexicon]], [plural: morph lexica, -s]. Domain: lexicon. Hyperonyms: lexicon. Cohyponym: morpheme lexicon, stem lexicon, fully inflected form lexicon. Def.: Lexicon based on the morphological parts of words, coupled with lexical rules for defining the composition of words from these parts. (Gibbon et al. 1997, p. 199)

morpheme lexicon
/ˈmɔːfiːm ˈleksɪkən/, /'mO:fi:m 'leksIk@n/, [N: [N: morpheme][N: lexicon]], [plural: morpheme lexica, -s]. Domain: lexicon. Hyperonyms: lexicon. Cohyponym: morph lexicon, stem lexicon, fully inflected form lexicon. Def.: An inventory of the morphemes of a language.

morpheme
/ˈmɔːfiːm/, /'mO:fi:m/, [N: morpheme], [plural: -s]. Domain: lexicon. Hyperonyms: morphological unit, subword unit. Hyponyms: lexical morpheme, grammatical morpheme. Meronym. sup.: word. Def.: 1. In traditional terminology: A morpheme is the minimal meaning-bearing unit of a language. (Gibbon et al. 1997, p. 214) 2. In a sign-based model: A morpheme is the smallest abstract sign-structured component of a word, and its assigned representations of its meaning, distribution and surface properties. (Gibbon et al. 1997, p. 214)

morphing
/ˈmɔːfɪŋ/, /'mO:fIN/, [N: morphing], [plural: none]. Domain: multimodal systems. Hyperonyms: interpolation technique. Def.: Technique to interpolate between synthetic objects or images. Real video footage of a person can be used to generate videos of the same person saying arbitrary text/utterances. A set of phonemes is labelled automatically manually from training data, as well as from the new audio track one desires to animate. The system selects the closest mouth video image and stitches it into the background image using a morphing technique. Head direction and orientation have to be adapted accordingly.

morphographemic alternation
/mɔːfəʊgrəˈfiːmɪk ɒltəˈneɪʃən/, /mO:f@Ugr@'fi:mIk Qlt@'neIS@n/, [N: [AJ: morphographemic][N: alternation]], [plural: -s]. Domain: lexicon. Def.: Morphographemic alternations are the differences between spellings of parts of composite words and spellings of corresponding parts of simplex words. E.g. English 'knife' - 'knives'.

morphological decomposition
/mɔːfəʊlˈɒdʒɪkəl diːkɒmpəˈzɪʃən/, /mO:f@Ul'QdZIk@l di:kQmp@'zIS@n/, [N: [AJ: morphological][N: decomposition]], [plural: -s]. Domain: lexicon. Hyperonyms: analysis of words. Def.: Analysis of orthographic or phonological words into morphemes, i.e. elements belonging to the finite set of smallest subword parts with an identifiable meaning. Morphological decomposition is necessary when the language/spelling allows words to be strung together without intervening spaces or hyphens.

morphological word
/mɔːfəˈlɒdʒɪkəl wɜːd/ , /mO:f@'lQdZIk@l w3:d/ , [N: [AJ: morphological][N: word]], [plural: -s]. Domain: lexicon. Hyperonyms: word. Cohyponym: orthographic word, phonetic word, phonological word, syntactic word, prosodic word, graphemic word. Def.: Word based on the indivisibility and fixed internal structure of words. (Gibbon et al. 1997, p. 196)

morphology
/mɔːˈfɒlədʒi/, /mO:'fQl@dZi/, [N: morphology], [plural: y/-ies]. Domain: lexicon. Cohyponym: syntax, phonetics, phonology, pragmatics, semantics. Meronym. sup.: linguistics. Meronym. sub.: word formation, inflection. Def.: 1. The branch of linguistics which deals with the internal structure of words. Morphology is the definition of the composition of words as a function of the meaning, syntactic function, and phonological or orthographic form of their parts. (Gibbon et al. 1997, p. 214) 2. Morphology is concerned with generalisations about words as lexical signs, in respect of surface form, meaning, distribution and composition. More generally, morphological information is information about semantically relevant word structure. (Gibbon et al. 1997, p. 212)

morphophoneme
/mɔːfəʊˈfəʊniːm/, /mO:f@U'f@Uni:m/, [N: morphophoneme], [plural: -s]. Domain: lexicon. Hyperonyms: phonological entity. Synonyms: archiphoneme (is a near-synonym). Def.: Morphophonemes stand for classes of morphologically and phonologically related phoneme alternants, such as those which occur in final devoicing in German: cf. the p-b alternation class with singular and plural in /zi:p - zi:be/ 'Sieb' - 'Siebe'. (Gibbon et al. 1997, p. 207)

morphophonemic transcription
/mɔːfəʊfəˈniːmɪk træn'skrɪpʃən/, /mO:f@Uf@'ni:mIk tr{n'skrIpS@n/, [N: [AJ: morphophonemic][N: transcription]], [plural: -s]. Hyperonyms: phonemic transcription. Def.: A morphophonemic transcription provides a simplification of phonological information with respect to the phonological level; the simplifications utilise knowledge about the morphological structure of words, and permit the use of morphophonemes, which stand for classes of morphologically and phonologically related phonemes. Citations of morphophonemic representations are often delimited with brace brackets {...}. (Gibbon et al. 1997, p. 207) E.g. The phonemic representation of German 'Weg /ve:k/ ('way') - 'Wege /ve:g@/ ('ways') corresponds to a morphophonemic representation {ve:G} - {ve:G+@}. The morphophoneme {G} stands for the phoneme set {/k/, /g/}. (Gibbon et al. 1997, p. 207).

morphophonemics
/mɔːfəʊfə'niːmɪks/, /mO:f@Uf@'ni:mIks/, [N: morphophonemics], [plural: always plural]. Synonyms: morphophonology. Meronym. sup.: morphology, phonology. Def.: A branch of linguistics referring to the analysis and classification of the morphological and phonological factors which interact to affect the appearance of morphemes, or, correspondingly, the grammatical factors which affect the appearance of phonemes. (Crystal 1988, pp. 200-201)

morphophonological alternation
/mɔːfəʊfɒnə'lɒdʒɪkəl ɔltə'neɪʃən/, /mO:f@UfQn@'lQdZIk@l Qlt@'neIS@n/, [N: [AJ: morphophonological][N: alternation]], [plural: -s]. Domain: lexicon. Def.: Morphophonological alternations are the differences between pronunciations of parts of composite words and pronunciations of corresponding parts of simplex words. (Gibbon et al. 1997, p. 216) E.g. /f/ - /v/ in 'knife' - 'knives' /naIf/ - /naIvz/.

morphophonological rule
/mɔːfəʊfɒnə'lɒdʒɪkəl 'ruːl/ , /mO:f@UfQn@'lQdZIk@l 'ru:l/ , [N: [AJ: morphophonological][N: rule]], [plural: -s]. Domain: lexicon. Hyperonyms: rule. Def.: A morphophonological rule describes morphophonological alternations (Gibbon et al. 1997, p. 216)

morphotactics
/mɔːfəʊ'tæktɪks/, /mO:f@U't{ktIks/, [N: morphotactics], [plural: always plural]. Domain: lexicon. Cohyponym: word syntax. Meronym. sup.: morphology. Def.: Morphotactics is the definition of the phonological and orthographic forms of words as a function of the forms of their parts. (Gibbon et al. 1997, p. 214)

motion blur
/'məʊʃən 'blɜː/, /'m@US@n 'bl3:/, [N: [N: motion][N: blur]], [plural: -s]. Domain: multimodal systems. Def.: Parameter values driving a facial model are blurred with their neighbourhood parameters (corresponding to the precedent and successive frames) using a Gaussian filter.

Motion Picture Experts Group
/'məʊʃən 'pɪktʃə 'ekspɜːts 'gruːp/, /'m@US@n 'pIktS@ 'eksp3:ts 'gru:p/, [N: [N: Motion][N: Picture][N: Experts][N: Group]], [plural: none]. Hyperonyms: committee, ISO standard. Hyponyms: MPEG-1, MPEG-2, MPEG-3, MPEG-4. Synonyms: MPEG. Def.: A committee that proposed a family of standards for multi-media file formats. MPEG is now an ISO standard.

MPEG
/'empeg/, /'empeg/, [N: MPEG], [plural: none]. Hyperonyms: committee, ISO standard. Hyponyms: MPEG-1, MPEG-2, MPEG-3, MPEG-4. Synonyms: Motion Picture Experts Group. Def.: A committee called the Motion Picture Experts Group (MPEG) has proposed a family of standards for multi-media file formats. MPEG is now an ISO standard.

MPEG-1
/'empeg 'wʌn/, /'empeg 'wVn/, [N: MPEG-1], [plural: none]. Hyperonyms: MPEG, compression scheme. Cohyponym: MPEG-2, MPEG-3, MPEG-4. Def.: Media: audio, video; Description: video recorder or standard TV quality data; Data rate: $<$ 4 Mb/s.

MPEG-2
/'empeg 'tuː/, /'empeg 'tu:/, [N: MPEG-2], [plural: none]. Hyperonyms: MPEG, compression scheme. Cohyponym: MPEG-1, MPEG-3, MPEG-4. Def.: Media: audio, video; Description: high definition TV (HDTV) quality data; Data rate: 2-15 Mb/s.

MPEG-3
/'empeg 'θriː/, /'empeg 'Tri:/, [N: MPEG-3], [plural: none]. Hyperonyms: MPEG, compression scheme. Cohyponym: MPEG-1, MPEG-2, MPEG-4. Def.: MPEG-3 is defined specifically for audio data. It is a lossy compression scheme that results in very low data rates (approx. 10% of the data rate of audio CD-ROMs) at little or no perceivable loss of quality. Media: audio; Description: low data rate, high quality audio; Data rate: 8-320 Kb/s.

MPEG-4
/'empeg 'fɔː/, /'empeg 'f0:/, [N: MPEG-4], [plural: none]. Hyperonyms: MPEG, compression scheme. Cohyponym: MPEG-1, MPEG-2, MPEG-3. Def.: Media: audio, video; Description: low quality, very low data rate for videoconferencing via telephone or ISDN lines; Data rate: 8-64 Kb/s.

MS-MIN
/'em 'es 'mɪn/, /'em 'es 'mIn/, [N: MS-MIN], [plural: -s]. Domain: multimodal systems. Hyperonyms: frame-merging architecture. Synonyms: multi-state mutual information network. Def.: Each grouped sequence of input events is assigned a score based on their mutual information. A dynamic programming algorithm (similar to Viterbi search or Dynamic Time Warping used in speech recognisers) determines the best sequence of input event interpretations that fit the whole multimodal input event.

multi-layer perceptron
/'mʌlti 'leɪə pə'septrɒn/, /'mVlti 'leI@ p@'septrQn/, [N: [AJ: multi][N: layer][N: perceptron]], [plural: -s]. Hyperonyms: neural network. Def.: A particular kind of neural network with multiple layers of elements whose output is a non-linear function of their inputs.

Multi-purpose Internet Mail Extension
/'mʌlti 'pɜːpəs 'ɪntənet 'meɪl ɪks'tenʃən/, /'mVlti 'p3:p@s 'Int@net 'meIl Iks'tenS@n/, [N: [N: Multi-Purpose][N: Internet][N: Mail][N: Extension]], [plural: none]. Hyperonyms: document descriptor. Synonyms: MIME. Def.: A document descriptor that is associated with a document to identify its type and format.

multi-state mutual information network
/'mʌlti 'steɪt 'mjuːtjʊəl ɪnfə'meɪʃən 'netwɜːk/, /'mVlti 'steIt 'mju:tjU@l Inf@'meIS@n 'netw3:k/, [N: [AJ: multi-state][AJ: mutual][N: information][N: network]], [plural: -s]. Domain: multimodal systems. Hyperonyms: frame-merging architecture. Synonyms: MS-MIN. Def.: Each grouped sequence of input events is assigned a score based on their mutual information. A dynamic programming algorithm (similar to Viterbi search or Dynamic Time Warping used in speech recognisers) determines the best sequence of input event interpretations that fit the whole multimodal input event.

multi-stroke gesture
/'mʌlti 'strəʊk 'dʒestʃə/, /'mVlti 'str@Uk 'dZestS@/, [N: [AJ: multi][N: stroke][N: gesture]], [plural: -s]. Domain: multimodal systems. Hyperonyms: gesture. Cohyponym: single-stroke gesture. Def.: Gesture consisting of more than one stroke, i.e. the smallest meaningful unit of gesture input consists of multiple strokes.

Multi-Vendor Integration Protocol
/'mʌlti 'vendər ɪntɪ'greɪʃən 'prɒtəkɒl/, /'mVlti 'vend@r IntI'greIS@n 'prQt@kQl/, [N: [AJ: Multi][N: Vendor][N: Integration][N: Protocol]], [plural: -s]. Domain: system design. Hyperonyms: bus. Synonyms: MVIP. Def.: MVIP is a multiplexed digital telephony highway for use within one computer chassis. It provides standard connection for digital telephone traffic between individual circuit boards. It supports telephone circuit-switching under direct computer control, using digital switch elements distributed amongst circuit boards in a standard computer. MVIP software standards allow system integrators to combine MVIP-compatible products from different vendors. The communication technologies that are supported include call management, voice store and forward, speech recognition, text-to-speech, Fax, data communications, and digital circuit-switching. The objective of an MVIP bus is to carry telephone traffic. It allows the interface to the telephone network to be separated from voice processing resources so that the telephone interface may be obtained from one vendor while the voice processing resources are obtained from others. A single MVIP bus has a capacity of 256 full-duplex telephone channels.

multi-word unit
/'mʌlti 'wɜːd 'juːnɪt/, /'mVlti 'w3:d 'ju:nIt/, [N: [N: multi-word][N: unit]], [plural: -s]. Domain: dialogue representation. Hyperonyms: lexical unit. Synonyms: multi-word, idiom. Cohyponym: word. Def.: A lexical unit consisting of a fixed combination of more than one word. E.g. 'I see', 'I'm sorry', 'thank you', 'sort of'..

multimedia system
/mʌlti'miːdɪə 'sɪstəm/, /mVlti'mi:dI@ 'sIst@m/, [N: [N: multimedia][N: system]], [plural: -s]. Domain: multimodal systems. Cohyponym: multimodal system. Def.: Multimedia systems offer more than one device for user input to the system, and for system feedback to the user. Such devices include microphone, speaker, keyboard, mouse, touch screen, camera. In contrast to multimodal systems, multimedia systems do not generate abstract concepts automatically (which are typically encoded manually as meta-information instead), and they do not transform the information.

multimodal input event
/mʌlti'məʊdəl 'ɪnpʊt ɪ'vent/, /mVlti'm@Ud@l 'InpUt I'vent/, [N: [AJ: multimodal][N: input][N: event]], [plural: -s]. Domain: multimodal systems. Cohyponym: unimodal input event, monomodal input event. Def.: Set of user input events that belong together and are intended to convey a chunk of information, consisting of at least two parts in different modalities.

multimodal interface
/mʌlti'məʊdəl 'ɪntəfeɪs/, /mVlti'm@Ud@l 'Int@feIs/, [N: [AJ: multimodal][N: interface]], [plural: -s]. Domain: multimodal systems. Hyperonyms: multimodal system. Cohyponym: multimodal speech system. Def.: An interface that combines speech input or output with other input and output modalities.

multimodal speech system
/mʌlti'məʊdəl 'spiːtʃ 'sɪstəm/, /mVlti'm@Ud@l 'spi:tS 'sIst@m/, [N: [AJ: multimodal][N: speech][N: system]], [plural: -s]. Domain: multimodal systems. Hyperonyms: multimodal system. Cohyponym: multimodal interface. Def.: Multimodal speech systems attempt to achieve ease of communication by integrating automatic speech recognition with other non-verbal cues, and by integrating non-verbal cues with speech synthesis to improve on the output side of a multimodal application (e.g., in talking heads).

multimodal system
/mʌlti'məʊdəl 'sɪstəm/, /mVlti'm@Ud@l 'sIst@m/, [N: [AJ: multimodal][N: system]], [plural: -s]. Domain: multimodal systems. Hyponyms: multimodal interface, multimodal speech system. Cohyponym: multimedia system. Def.: Multimodal systems represent and manipulate information from different human communication channels at multiple levels of abstraction. Multimodal systems can automatically extract meaning from multimodal, raw input data, and conversely they produce perceivable information from symbolic abstract representations.

multimodality
/mʌltiməʊ'dælɪti/, /mVltim@U'd{lIti/, [N: multimodlity], [plural: none]. Domain: multimodal systems. Def.: The cooperation between several modalities in order to improve the (human–computer) interaction.

muscle-based model
/'mʌsəl 'beɪst 'mɒdəl/, /'mVs@l 'beIst 'mQd@l/, [N: [N: muscle][AJ: based][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: physically-based model. Cohyponym: structural model. Def.: This method integrates anatomical features (e.g. skull, skin, muscle) and properties of the face (elasticity of the skin and muscle contraction). The spring-mass model simulates skin and muscle behaviour. Each muscle is characterised by a vector that represents a direction, a magnitude, and a zone of influence.

MVIP
/'em 'viː 'aɪ 'piː/, /'em 'vi: 'aI 'pi:/, [N: MVIP], [plural: -s]. Domain: system design. Hyperonyms: bus. Synonyms: Multi-Vendor Integration Protocol. Def.: MVIP is a multiplexed digital telephony highway for use within one computer chassis. It provides standard connection for digital telephone traffic between individual circuit boards. It supports telephone circuit-switching under direct computer control, using digital switch elements distributed amongst circuit boards in a standard computer. MVIP software standards allow system integrators to combine MVIP-compatible products from different vendors. The communication technologies that are supported include call management, voice store and forward, speech recognition, text-to-speech, Fax, data communications, and digital circuit-switching. The objective of an MVIP bus is to carry telephone traffic. It allows the interface to the telephone network to be separated from voice processing resources so that the telephone interface may be obtained from one vendor while the voice processing resources are obtained from others. A single MVIP bus has a capacity of 256 full-duplex telephone channels.

n-gram grammar
/'engræm 'græmə/, /'engr{m 'gr{m@/, [N:[N: n-gram][N: grammar]], [plural: -s]. Domain: . Cohyponym: null grammar, unigram grammar, bigram grammar, trigram grammar. Def.: A probabilistic grammar based on transition probabilities of words, predicting the probability of a word in a given context from the product of the a priori probability of the word and the probability of its (n-1) predecessors. The transition probabilities of words are calculated from their distribution in a corpus. Analogously, the probability of a word in a bigram or trigram grammar is calculated using the probabilities of the preceding word or the preceding two words.

n-gram language model
/'engræm 'læŋgwɪdʒ 'mɒdəl/, /'engr{m 'l{NgwIdZ 'mQd@l/, [N: [N: n-gram][N: language][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: stochastic language model. Synonyms: n-gram language model. Cohyponym: bigram language model, trigram language model. Def.: An n-gram language model is a stochastic language model that is based on a n-gram grammar, i.e. on conditional probabilities depending only on the (n-1) immediate predecessor words. (Gibbon et al. 1997, p. 243)

narrow phonetic transcription
/'nærəʊ fə'netɪk træn'skrɪpʃən/, /'n{r@U f@'netIk tr{n'skrIpS@n/, [N: [AJ: narrow][AJ: phonetic][N: transcription]], [plural: -s]. Domain: corpora, lexicon. Hyperonyms: phonetic transcription. Cohyponym: broad phonetic transcription. Def.: Phonetic transcription which is relatively detailed and represents more than the minimum number of phonetic features which are needed to represent phonemes. (Crystal 1988, p. 313)

nasal
/'neɪzəl/, /'neIz@l/, [N: nasal], [plural: -s]. Hyperonyms: consonant; manner of articulation. Cohyponym: plosive, trill, tap, flap, fricative, lateral fricative, approximant, lateral approximant. Def.: A nasal consonant or vowel is classified on the basis of manner of articulation: a sound produced while the soft palate is lowered to allow resonance of the nasal cavity and escape of air through the nose. (Crystal 1988, p. 203) E.g. [m, n, N].

Natural Language Processing
/'nætʃərəl 'læŋgwɪdʒ 'prəʊsesɪŋ/, /'n{tS@r@l 'l{NgwIdZ 'pr@UsesIN/, [N: [AJ: Natural][N: Language][N: Processing]], [plural: none]. Synonyms: NLP. Cohyponym: SLP, spoken language processing. Def.: The parsing of written language into semantic concepts.

natural language
/'nætʃərəl 'læŋgwɪdʒ/, /'n{tS@r@l 'l{NgwIdZ/, [N: [AJ: natural][N: language]], [plural: -s]. Hyperonyms: language. Cohyponym: formal language. Def.: The forms of language used by humans for communication, as opposed to artificial or formal languages such as logics, algebras, programming languages. The controlled language used between people and invented systems can be also termed "natural" if it is what users spontaneously produce in response to the situation.

near end echo
/nɪər'end 'ekəʊ/, /nI@r'end 'ek@U/, [N: [AJ: near][N: end][N: echo]], [plural: -es]. Hyperonyms: echo. Cohyponym: far end echo. Def.: The electrical echo of the output of a telephone system generated at the point where the system is connected to the network. Near end echo is always an electrical phenomenon.

neural network based approach
/'njuːrəl 'netwɜːk 'beɪst ə'prəʊtʃ/, /'nju:r@l 'netw3:k 'beIst @'pr@UtS/, [N: [AJ: neural][N: network][AJ:based][N: approach]], [plural: -es]. Domain: multimodal systems. Hyperonyms: template matching. Cohyponym: Principal Component Analysis, PCA, geometric template matching, deformable template matching, optical flow technique. Def.: Another variant of template matching are neural-network based approaches to, for instance, speech or face recognition, for example applying multi-layer perceptrons or Kohonen self-organising maps.

neural network
/'njuːrəl 'netwɜːk/, /'nju:r@l 'netw3:k/, [N: [AJ: neural][N: network]], [plural: -s]. Hyponyms: multi-layer perceptrion, back-propagation network, Kohonen map. Def.: An artificial neural network is a pattern recognition system modelled on the human nervous system, in which input values are mapped via a network of weighted functions (nodes) into output classes.

newsgroup
/'njuːzgruːp/, /'nju:zgru:p/, [N: newsgroup], [plural: -s]. Def.: Moderated or unmoderated Internet information service; a newsgroup is focused on a specific subject, e.g. Spoken Language Processing (SLP). Users subscribe to a newsgroup to read and post mailings.

NIST-SPHERE header
/'nɪst 'sfiːə 'hedə/, /'nIst 'sfi:@ 'hed@/, [N: [N: NIST-SPHERE][N: header]], [plural: -s]. Hyperonyms: audio signal header. Def.: Standard header for audio signals proposed by the National Institute of Standards and Technology (NIST), consisting of a simple ASCII formatted text information describing the signal following the header; the header options can be user-specified.

NLP
/'en 'el 'piː/, /'en 'el 'pi:/, [N: NLP], [plural: none]. Synonyms: Natural Language Processing. Cohyponym: SLP, Spoken Language Processing. Def.: A theoretical and applied discipline, involving computational linguistics and artificial intelligence, which is concerned with natural language understanding (parsing and interpretation) and natural language generation (production of sentences and texts).

noise-cancelling microphone
/nɔɪz 'kænsəlɪŋ 'maɪkrəfəʊn/, /nOIz 'k{ns@lIN 'maIkr@f@Un/, [N: [N: noise][V: cancelling][N: microphone]], [plural: -s]. Hyperonyms: microphone. Synonyms: close-talking microphone. Def.: A noise-cancelling microphone is more sensitive to nearby sound sources (i.e. the mouth) than to distant sound sources (e.g. noise sources).

non-adaptive language model
/nɒnə'dæptɪv 'læŋgwɪdʒ 'mɒdəl/, /nQn@'d{ptIv 'l{NgwIdZ 'mQd@l/, [N: [AJ: non-adaptive][N: language][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: language model. Cohyponym: adaptive language model. Def.: A non-adaptive language model does not depend on the test data, but remains unchanged as trained on the original training data. (Gibbon et al. 1997, p. 257)

non-parametric test
/nɒnpærə'metrɪk test/, /nQnp{r@'metrIk test/, [N: [AJ: non-parametric][N: test]], [plural: -s]. Domain: assessment methodologies. Hyperonyms: statistical test. Synonyms: distribution-free test. Cohyponym: parametric test. Def.: Statistical test employed in simple hypothesis testing: non-parametric tests are used when discrete, rather than continuous, measures are obtained.

non-registered speaker
/nɒn'redʒɪstəd 'spiːkə/, /nQn'redZIst@d 'spi:k@/, [N: [AJ: non-registered][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker. Synonyms: impostor. Cohyponym: registered speaker. Def.: A speaker who does not belong to the list of registered users for a given speaker recognition system.

notation
/nəʊ'teɪʃən/, /n@U'teIS@n/, [N: notation], [plural: -s]. Hyponyms: transcription. Def.: An inventory of symbols used to represent the vocabulary of a formalism.

notch filter
/'nɒtʃ 'fɪltə/ , /'nQtS 'fIlt@/ , [N: [N: notch][N: filter]], [plural: -s]. Domain: physical characterisation. Hyperonyms: filter. Cohyponym: low-pass filter, high-pass filter, band-pass filter, band-stop filter, all-pass filter. Def.: A notch filter passes all frequencies except those between two threshold frequencies (the notch). It can be understood as a cascade of a low pass filter and a high pass filter, where the cut-off frequency of the low-pass filter is lower than the cut-off frequency of the high-pass filter. A notch filter is used to filter out narrow band noise, usually in the form of simple tones such as whistles.

noun
/'naʊn/, /'naUn/, [N: noun], [plural: -s]. Domain: lexicon. Hyperonyms: lexical category. Cohyponym: verb, adverb, adjective. Def.: One of the four main lexical categories (parts of speech), typically occurring as subject of a sentence, modified by adjectives and articles, with characteristic internal structure, and in many languages inflecting for case, number and gender.

Nyquist rate
/'njuːkwɪst 'reɪt/, /'nju:kwIst 'reIt/, [N: [N: Nyquist][N: rate]], [plural: -s]. Hyperonyms: sampling rate. Def.: The Nyquist rate is half the sampling rate of a time-sampled signal. It corresponds to the highest frequency that can normally be determined in the sampled signal.

object
/'ɒbdʒekt/, /'QbdZekt/, [N: object], [plural: -s]. Domain: terminology. Def.: An object of the real world.

object-oriented programming language
/'ɒbdʒekt ɒrɪ'entɪd 'prəʊgræmɪŋ 'læŋgwɪdʒ/, /'QbdZekt QrI'entId 'pr@Ugr{mIN 'l{NgwIdZ/, [N: [AJ: object-oriented][AJ: programming][N: language]], [plural: none]. Hyperonyms: programming language. Hyponyms: Java, C++, python. Def.: An object-oriented programming language uses a programming model in which the main data structure is a set of hierarchically related objects, each of which inherits properties (data structures and algorithms) from the object immediately above it in the hierarchy, which it shares with other similar objects which inherit from the same object, in addition to having its own properties.

off-line comprehension test
/'ɒflaɪn kɒmprɪ'henʃən 'test/, /'QflaIn kQmprI'henS@n 'test/, [N: [PREP: off][N: line][N: comprehension][N: test]], [plural: -s]. Domain: speech synthesis. Hyperonyms: comprehension test, off-line test. Cohyponym: on-line comprehension test. Def.: Comprehension test in which content questions have to be answered in an open or closed response mode. (Gibbon et al. 1997, p. 496)

off-line identification test
/'ɒflaɪn aɪdentɪfɪ'keɪʃən 'test/, **/'QflaIn aIdentIfI'keIS@n 'test/**, [N: [PREP: off][N: line][N: identification][N: test]], [plural: -s]. Domain: speech synthesis. Hyperonyms: identification test, off-line test. Cohyponym: on-line identification test. Def.: Identification test where subjects are asked to transcribe the separate elements (sounds, words) making up the test items, either in the form that subjects are forced to select the appropriate response from a limited number of pregiven categories (closed set), or in the form that the only restriction are the constraints imposed by the language. Transcription can be in normal spelling or in some unambiguous notation. (Gibbon et al. 1997, p. 495)

off-line testing
/'ɒflaɪn 'testɪŋ/, **/'QflaIn 'testIN/**, [N: [PREP: off][N: line][N: testing]], [plural: none]. Domain: speech synthesis. Hyperonyms: testing procedure. Hyponyms: off-line identification test, off-line comprehension test. Cohyponym: on-line test. Def.: Procedure in which subjects are given some time to reflect before responding to a (spoken) stimulus.

omnidirectional microphone
/ɒmnidaɪ'rekʃənəl 'maɪkrəfəʊn/, **/QmnidaI'rekS@n@l 'maIkr@f@Un/**, [N: [AJ: omnidirectional][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: microphone. Cohyponym: ultradirectional microphone, bidirectional microphone, unidirectional microphone, pressure zone microphone, headset microphone. Def.: The omnidirectional microphone is sensitive to sound without regard to the direction of the incidence. Thus it will pick up the wanted sound produced by the speaker as well as unwanted background noise. This feature makes an omnidirectional microphone a bad choice when unwanted noise sources are to be expected. On the other hand, it is the most simple type of microphone from the viewpoint of microphone design. As a matter of fact, omnidirectional microphones are the most natural microphones available since the least design compromises have to be made. Thus, omnidirectional microphones are the best choice for high-quality speech recordings as long as the ambient noise floor can be kept low. (Gibbon et al. 1997, p. 303)

on-line comprehension test
/'ɒnlaɪn kɒmprɪ'henʃən 'test/, **/'QnlaIn kQmprI'henS@n 'test/**, [N: [PREP: on][N: line][N: comprehension][N: test]], [plural: -s]. Domain: speech synthesis. Hyperonyms: comprehension test, on-line test. Cohyponym: off-line comprehension test. Def.: On-line comprehension tests require the subject to indicate whether a statement is true or not. (Gibbon et al. 1997, p. 496)

on-line identification test
/'ɒnlaɪn aɪdentɪfɪ'keɪʃən 'test/, **/'QnlaIn aIdentIfI'keIS@n 'test/**, [N: [PREP: on][N: line][N: identification][N: test]], [plural: -s]. Domain: speech synthesis. Hyperonyms: identification test, on-line test. Cohyponym: off-line identification test. Def.: On-line identification tests require the subject to decide whether the stimulus does or does not exist as a word in the language. (Gibbon et al. 1997, p. 496)

on-line testing
/'ɒnlaɪn 'testɪŋ/ , **/'QnlaIn 'testIN/** , [N: [PREP: on][N: line][N: testing]] , [plural: none] . Domain: speech synthesis. Hyperonyms: testing procedure. Hyponyms: on-line comprehension test, on-line identification test. Cohyponym: off-line testing. Def.: Procedure that requires an immediate response from the subjects, tapping the perception process before it is finished.

onset
/'ɒnset/, **/'Qnset/**, [N: onset], [plural: -s]. Synonyms: margin; slope; trough. Cohyponym: nucleus; crest; peak . Meronym. sup.: syllable. Def.: A term used in phonetics and phonology to refer to the opening segment of a linguistic unit (e.g. a syllable, a tone unit) ... (Crystal, p. 212) Consonants generally occur before or behind, as onset or coda, to a nucleus that contains the vowel or vowel-like feature in a syllable.

ontological hierarchy
/ɒntə'lɒdʒɪkəl 'haɪərɑːki/, /Qnt@'lQdZIk@l 'haI@rA:ki/, [N: [AJ: ontological][N: hierarchy]], [plural: y/-ies]. Domain: terminology. Hyperonyms: hierarchy. Synonyms: partitive hierarchy, meronomy, mereonomy, PARTOF hierarchy. Cohyponym: taxonomy, ISA hierarchy, logical concept hierarchy, generic concept hierarchy. Def.: Hierarchy of concepts in a PARTOF relation, i.e. a mereonomy.

OOV word
/'əʊ 'əʊ 'viː 'wɜːd/, /'@U '@U 'vi: 'w3:d/, [N: [N: OOV][N: word]], [plural: -s]. Domain: speech recognition, language modelling. Synonyms: out-of-vocabulary word. Def.: Word not listed in the lexicon of a spoken language system.

OOV-rejection
/'əʊ 'əʊ 'viː rɪ'dʒekʃən/, /'@U '@U 'vi: rI'dZekS@n/, [N: [N: OOV][N: rejection]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: performance measure. Cohyponym: recognition accuracy, error recovery, response time, situational awareness. Def.: Rejection rate for out-of-vocabulary words.

opinion testing
/ə'pɪnɪən 'testɪŋ/ , /@'pInI@n 'testIN/ , [N: [N: opinion][N: testing]] , [plural: none] . Domain: speech synthesis. Hyperonyms: testing procedure. Synonyms: judgment testing. Cohyponym: functional testing. Def.: Procedure whereby a group of listeners is asked to judge the performance of a speech output system along a number of rating scales.

optical flow technique
/'ɒptɪkəl 'fləʊ tek'niːk/, /'QptIk@l 'fl@U tek'ni:k/, [N: [AJ: optical][N: flow][N: technique]], [plural: -s]. Domain: multimodal systems. Hyperonyms: template matching. Cohyponym: Principal Component Analysis, PCA, geometric template matching, deformable template matching, neural network based approach. Def.: The optical flow technique allows the detection of motion (from a sequence of images) rather than facial feature displacements, by computing the difference in image intensity between two consecutive frames. It works at the pixel level. The computation is done pixel per pixel. This technique may be used to extract muscle contraction. Windows are placed around muscle locations. The velocity of each muscle contraction is computed.

oral dialogue
/'ɔːrəl 'daɪəlɒg/, /'O:r@l 'daI@lQg/, [N: [AJ: oral][N: dialogue]] , [plural: -s] . Hyperonyms: dialogue. Def.: See spoken language dialogue. This term is quite widely used, though it is less favoured by native speakers of English than by those who have learned it as a second-language.

organic speech disorder
/ɔː'gænɪk 'spiːtʃ dɪs'ɔːdə/, /O:'g{nIk 'spi:tS dIs'O:d@/, [N: [AJ: organic][N: speech][N: disorder]], [plural: -s]. Domain: corpora. Hyperonyms: speech disorder. Cohyponym: functional speech disorder. Def.: Speech disorders where there is a clear organic (anatomical, physiological, neurological) cause. (Gibbon et al. 1997, p. 114)

orthographic break
/ɔːθə'græfɪk 'breɪk/, /O:T@'gr{fIk 'breIk/, [N: [AJ: orthographic][N: break]], [plural: -s]. Domain: lexicon. Synonyms: orthographic syllable. Def.: The orthographic break is in general only needed for splitting words at line-breaks and does not correspond closely to either syllable or morph boundaries but combines phonological, morphological, orthographic or other criteria. Conventions differ from language to language. (Gibbon et al. 1997, p. 213)

orthographic lexical morph
/ɔːθə'græfɪk 'leksɪkəl 'mɔːf/, /O:T@'gr{fIk 'leksIk@l 'mO:f/, [N: [AJ: orthographic][AJ: lexical][N: morph]], [plural: -s]. Domain: lexicon. Hyperonyms: lexical morph, orthographic morph. Cohyponym: phonological lexical morph. Def.: The orthographic realisation of a lexical morpheme. (Gibbon et al. 1997, p. 199)

orthographic morph
/ɔːθəˈgræfɪk ˈmɔːf/, /O:T@'gr{fIk 'mO:f/, [N: [AJ: orthographic][N: morph]], [plural: -s]. Domain: lexicon. Hyperonyms: morph. Hyponyms: orthographic lexical morph. Cohyponym: phonological morph. Def.: Orthographic morphs are morphs consisting of graphemes (either single letters or fixed combinations of letters). (Gibbon et al. 1997, p. 215) E.g. 'catches' = catch + es: 'es' is an orthographic morph of the English 3rd P.Sg. Pres. Tense morpheme..

orthographic noise
/ɔːθəˈgræfɪk ˈnɔɪz/, /O:T@'gr{fIk 'nOIz/, [N: [AJ: orthographic][N: noise]], [plural: none]. Domain: lexicon. Def.: Intrusive artefacts in text-to-speech or speech recognition systems due to the use of orthography in interface specifications; orthographic noise is due to homography, homophony, and other irregularities in grapheme-phoneme relationships at character and word level. (Gibbon et al. 1997, p. 203)

orthographic transcription
/ɔːθəˈgræfɪk trænˈskrɪpʃən/, /O:T@'gr{fIk tr{n'skrIpS@n/, [N: [AJ: orthographic][N: transcription]], [plural: -s]. Hyperonyms: transcription . Cohyponym: phonemic transcription, phonetic transcription. Def.: Transcription of a speech signal in standard orthography.

orthographic word
/ɔːθəˈgræfɪk ˈwɜːd/, /O:T@'gr{fIk 'w3:d/, [N: [AJ: orthographic][N: word]], [plural: -s]. Domain: lexicon. Hyperonyms: word. Cohyponym: graphemic word, phonetic word, phonological word, morphological word, syntactic word, prosodic word. Def.: A word defined in terms of its conformity to the orthographic structure and punctuation of a language, for instance as delimited by spaces or punctuation marks. (Gibbon et al. 1997, p. 196)

orthography
/ɔːˈθɒgrəfi/, /O:'TQgr@fi/, [N: orthography], [plural: y/-ies]. Domain: lexicon, graphemics. Hyperonyms: surface representation. Hyponyms: alphabetic orthography, syllabic orthography, logographic orthography; Roman orthography, Cyrillic orthography, Greek orthography; standard orthography, reformed orthography. Synonyms: spelling. Cohyponym: pronunciation, phonology, phonetics. Meronym. sup.: writing. Meronym. sub.: letters, characters. Def.: The two-dimensional visual representation of word forms by standardised graphical characters (Gibbon et al. 1997, p. 188)

OSQL
/ˈəʊ ˈes ˈkjuː ˈel/, /'@U 'es 'kju: 'el/, [N: OSQL], [plural: none]. Hyperonyms: query language. Cohyponym: SQL, Standard Query Language. Def.: An object-oriented high-level database query language for which a draft standard has been published by the ODMG.

out-of-vocabulary word
/ˈaʊt əv vəˈkæbjʊləri ˈwɜːd/, /'aUt @v v@'k{bjUl@ri 'w3:d/, [N: [PREP: out][PREP: of][N: vocabulary][N: word]], [plural: -s]. Domain: speech recognition, language modelling . Synonyms: OOV word. Def.: Word not listed in the lexicon of a spoken language system.

output device
/ˈaʊtpʊt dɪˈvaɪs/, /'aUtpUt dI'vaIs/, [N: [N: output][N: device]], [plural: -s]. Domain: multimodal systems. Hyperonyms: technical device. Hyponyms: visual device, acoustic device, haptic device. Def.: Modality used by a multimodal system to communicate with the user.

output medium
/ˈaʊtpʊt ˈmiːdɪəm/, /'aUtpUt 'mi:dI@m/, [N: [N: output][N: medium]], [plural: um/-a]. Domain: multimodal systems. Hyperonyms: modality. Hyponyms: non-speech output modality. Cohyponym: input modality. Def.: An output channel from machine to human in human-machine communication. Output such as synthesised speech, synthesised faces, talking heads (combination of speech and face synthesis), synthetic agents, force feedback, traditional multimedia output (text, graphics, video, sound).

output modality
/ˈaʊtpʊt məˈdælɪti/, /'aUtpUt m@'d{lIti/, [N: [N: output][N: modality]], [plural: -s]. Domain: multimodal systems. Hyperonyms: modality. Hyponyms: non-speech output modality. Cohyponym: input modality. Def.: A human speech or gestural output channel used in communication, generally acoustic or visual.

paired comparison
/ˈpeəd kəmˈpærɪsən/, /'pe@d k@m'p{rIs@n/, [N: [AJ: paired][N: comparison]], [plural: -s]. Domain: speech synthesis, assessment methodologies. Hyperonyms: testing procedure. Def.: A psychophysical testing procedure that is used when subjects are required to judge between two stimuli. In Language Engineering this might be judging which of two recogniser outputs has more or less intelligibility.

palatal consonant
/ˈpælətəl ˈkɒnsənənt/, /'p{l@t@l 'kQns@n@nt/, [N: [AJ: palatal][N: consonant]], [plural: -s]. Hyperonyms: consonant. Synonyms: palatal. Cohyponym: bilabial consonant, labiodental consonant, dental consonant, alveolar consonant, postalveolar consonant, retroflex consonant, velar consonant, uvular consonant, pharyngeal consonant, glottal consonant. Def.: Palatal consonant is a term used in the phonetic classification of speech sounds on the basis of their place of articulation: it refers to a sound made when the front of the tongue is in contact with or approaches the hard palate. (Crystal 1988, p. 219)

paradigmatic relation
/pærədɪgˈmætɪk rɪˈleɪʃən/, /p{r@dIg'm{tIk rI'leIS@n/, [N: [AJ: paradigmatic][N: relation]] , [plural: -s]. Domain: lexicon. Hyperonyms: linguistic characterisation. Cohyponym: syntagmatic relation. Def.: Paradigmatic relations are classificatory, disjunctive, element-class, subclass-superclass relations.

PARADISE
/ˈpærədaɪs/, /'p{r@daIs/, [N: PARADISE], [plural: none]. Domain: multimodal systems. Hyperonyms: evaluation framework. Def.: Framework for evaluating dialogue systems from a user point of view. It assumes that the ultimate measure of success for a dialogue system is user satisfaction. Since many different factors influence user satisfaction, dependent on the application, PARADISE proposes to use statistical methods of determining the most significant predictors of cumulative user satisfaction for a specific application, out of a large set of potentially useful variables.

paralinguistic feature
pærəlɪŋˈgwɪstɪk ˈfiːtʃə/, p{r@lIN'gwIstIk 'fi:tS@/, [N: [AJ: paralinguistic][N: feature]], [plural: -s]. Domain: dialogue representation. Def.: Concomitant properties of voice such as laughter, tempo, loudness, and so on that occur during speech. We exclude features that do not accompany speech but rather occur in isolation.

parametric model
/pærəˈmetrɪk ˈmɒdəl/, /p{r@'metrIk 'mQd@l/, [N: [AJ: parametric][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: synthetic model. Def.: A facial model is created and animated through a set of parameters. Generally, parameters can be divided into two groups: conformation and expression parameters. The former refer to parameters acting on the facial topology (including position and size of the nose and eyes, global size of the head). The latter specify facial expressions such as brow action, mouth movement, and blink.

parametric synthesis
/pærəˈmetrɪk ˈsɪnθəsɪs/, /p{r@'metrIk 'sInT@sIs/, [N: [AJ: parametric][N: synthesis]], [plural: parametric syntheses]. Domain: speech synthesis. Hyperonyms: speech synthesis. Cohyponym: concatenative synthesis. Def.: Speech synthesis by modelling the (human) articulatory or vocal tract. Parameters control the shape of the vocal tract and hence determine the speech output.

parametric test
/pærəˈmetrɪk test/, /p{r@'metrIk test/, [N: [AJ: parametric][N: test]], [plural: -s]. Domain: assessment methodologies. Hyperonyms: testing procedure. Cohyponym: non-parametric test, distribution-free test. Def.: Statistical test employed in simple hypothesis testing: parametric tests are used when continuous measures are available.

parsing
/ˈpɑːzɪŋ/, /'pA:zIN/, [N: parsing], [plural: none]. Domain: speech recognition, speech synthesis. Hyponyms: morphological parsing, syntactic parsing, prosodic parsing; sentence parsing. Def.: The labelling of the parts of speech or grammatical elements of sentences, e.g. subject, predicate, past tense, noun, verb, either by a human analyst or by means of a parser (a programme based on a parsing algorithm). (Crystal 1988, p. 221)

part of speech
/ˈpɑːt əv ˈspiːtʃ/, /'pA:t @v 'spi:tS/, [N: [N: part][PREP: of][N: speech]], [plural: parts of speech]. Domain: lexicon. Hyponyms: grammatical category, lexical category. Synonyms: syntactic category, POS. Def.: The traditional term for a grammatical class of words. (Crystal 1988, p. 222) E.g. noun, adjective, article, pronoun, verb, adverb, preposition, conjunction, interjection, proper noun, common noun, intransitive verb, transitive verb, ditransitive verb, prepositional verb .

partial action frame
/ˈpɑːʃəl ˈækʃən ˈfreɪm/, /'pA:S@l '{kS@n 'freIm/, [N: [AJ: partial][N: action][N: frame]], [plural: -s]. Domain: multimodal systems. Def.: Input from each modality is interpreted separately and then parsed and transformed into semantic frames containing slots that specify command parameters (parameter slots). The information in these (partial) action frames may be incomplete or ambiguous if not all elements of the command were expressed in a single modality. A domain-independent frame-merging algorithm combines partial frames into complete frames.

partial synonym
/ˈpɑːʃəl ˈsɪnənɪm/, /'pA:S@l 'sIn@nIm/, [N: [AJ: partial][N: synonym]], [plural: -s]. Domain: lexicon. Hyperonyms: synonym. Cohyponym: full synonym. Def.: Two words are partial synonyms if they have at least one meaning in common and if either has additional readings not shared by the other. (Gibbon et al. 1997, p. 850) E.g. 'Manual' and 'handbook' are partial synonyms. 'Manual' is also, among other things, a term for a traditional organ keyboard..

partitive hierarchy
/ˈpɑːtɪtɪv ˈhaɪərɑːki/, /'pA:tItIv 'haI@rA:ki/, [N: [AJ: partitive][N: hierarchy]], [plural: y/-ies]. Domain: terminology. Hyperonyms: hierarchy. Synonyms: ontological hierarchy, meronomy, mereonomy, PARTOF hierarchy. Cohyponym: taxonomy, ISA hierarchy. Def.: Hierarchy of concepts in a PARTOF relation, i.e. a mereonomy.

PARTOF hierarchy
/ˈpɑːtəv ˈhaɪərɑːki/, /'pA:t@v 'haI@rA:ki/, [N: [AJ: PARTOF][N: hierarchy]], [plural: y/-ies]. Domain: terminology. Hyperonyms: hierarchy. Synonyms: meronomy, mereonomy, partitive hierarchy, ontological hierarchy. Cohyponym: taxonomy, ISA hierarchy, logical concept hierarchy, generic concept hierarchy. Def.: A hierarchy defined by the relation of parts to wholes, and parts to parts.

PARTOF relation
/ˈpɑːtəv rɪˈleɪʃən/, /'pA:t@v rI'leIS@n/, [N: [AJ: PARTOF][N: relation]], [plural: -s]. Domain: terminology. Hyperonyms: lexical relation. Synonyms: meronomic relation, mereonomic relation, meronymic relation. Cohyponym: ISA relation, taxonomic relation, taxonymic relation. Def.: Fundamental syntactic or combinatorial relation. Like ISA, the term is also rather general, and a wide range of different relations are covered by it in different approaches to linguistics in general and lexicography in particular: syntagmatic relations, mereological (merological) / mereonomic (meronomic) relations, part-whole relations, part-part relations, (immediate) constituency / domination, command relations (e.g. c-command), dependency relations, government relations, argument structure, thematic role structure, subcategorisation frames, case frames, valency, anaphoric binding relations, categorial functor-argument application, concatenation, linear ordering, prosodic (autosegmental) association and precedence relations, child-child (sister) relations, parent-child (mother-daughter) relations.

passive vocabulary size
/ˈpæsɪv vəˈkæbjʊləri ˈsaɪz/, /'p{sIv v@'k{bjUl@ri 'saIz/, [N: [AJ: passive][N: vocabulary][N: size]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: vocabulary size. Cohyponym: active vocabulary size; extension vocabulary size, exception vocabulary size, user vocabulary size. Def.: The number of words the system has in store to be loaded into the active vocabulary.

PCA
/ˈpiː ˈsiː ˈeɪ/, /'pi: 'si: 'eI/, [N: PCA], [plural: none]. Domain: multimodal systems. Hyperonyms: template matching. Synonyms: Principal Component Analysis. Cohyponym: geometric template matching, deformable template matching, optical flow technique, neural network based approach. Def.: The simplest version of template matching. As applied to face recognition, a test image is classified based on its (Euclidean) distance to templates generated from the faces in the training set (database). The Kurhunen-Loeve procedure and eigenfaces are based on this simple template matching method. Eigenfaces correspond to characteristic feature images and can be viewed as the principal components of a test image with respect to characteristic features obtained from the database of faces. This technique has been applied to recognise lip shapes.

PDF
/ˈpiː ˈdiː ˈef/, /'pi: 'di: 'ef/, [N: PDF], [plural: none]. Hyperonyms: formal language. Synonyms: Portable Document Format. Def.: PDF is a proprietary format (Adobe) for describing the page layout of documents combined with the ability to perform text searches in the document, dynamic linking of documents, multi-media content, and input via forms, e.g. for interactive documents. Generally, PDF files are much smaller than PostScript, and they may be edited. PDF has become the most widespread format for online manuals and document collections on CD-ROM, e.g. conference proceedings. Unlike PostScript, it is not coded in ASCII format but in 8-bit code.

peak
/ˈpiːk/, /'pi:k/, [N: peak], [plural: -s]. Synonyms: nucleus; crest. Cohyponym: onset; coda; margin; slope; trough. Meronym. sup.: syllable. Def.: A peak contains the vowel or vowel-like features in a syllable.

PEB
/ˈpiː ˈiː ˈbiː/, /'pi: 'i: 'bi:/, [N: PEB], [plural: -s]. Domain: system design. Hyperonyms: bus. Synonyms: Pulse coded modulation Expansion Bus. Def.: PEB is seen as an internal switching matrix capable of routing any time slot to an adequate audio port of the speech recogniser.

performance evaluation
/pəˈfɔːməns ɪvæljuˈeɪʃən/ , /p@'fO:m@ns Iv{lju'eIS@n/ , [N: [N: performance][N: evaluation]], [plural: -s]. Domain: speech synthesis, multimodal systems. Hyperonyms: black box approach. Cohyponym: diagnostic evaluation, adequacy evaluation. Def.: 1. Performance evaluation of a system as a whole, typically used to compare systems developed by different manufacturers, or to establish the improvement of one system relative to an earlier edition (comparative testing). Black box evaluations consider the overall performance of a system without reference to any internal components or behaviours. Evaluations of this kind address large questions such as "How good is it as an integrated system?" rather than detailed questions of the "What is its word recognition rate?" variety. 2. Performance evaluations measure system performance in specific areas. Performance evaluation is only meaningful if a well-defined baseline performance exists, typically a previous version of the system, or a different technology that supports the same functionality. Performance evaluation is typically used by system developers and program managers.

performance
/pəˈfɔːməns/, /p@'fO:m@ns/, [N: performance], [plural: none]. Domain: interactive dialogue systems. Cohyponym: competence. Def.: Term denoting the production and perception of individual speech events based on competence, i.e. on the intuitive knowledge of the 'ideal' speaker/hearer. (cf. also Bussmann, p. 567) It is generally held that there is a dislocation between competence and performance such that there is not a straightforward mapping from one to the other.

performance-driven face synthesis
/pəˈfɔːməns drɪvən ˈfeɪs ˈsɪnθəsɪs/, /p@'fO:m@ns drIv@n 'feIs 'sInT@sIs/, [N: [N: performance][AJ: driven][N: face][N: synthesis]], [plural: none]. Domain: multimodal systems. Hyperonyms: face synthesis. Cohyponym: audio-driven face synthesis, puppeteer control face synthesis, text-to-visual-speech face synthesis. Def.: A person's movements are tracked and converted into parameters controlling the facial models. Some techniques track reflective spots attached artificially on the person's face, others track the actor's facial features directly. A mapping is constructed from the extracted data and the facial model parameters. This method works well if the facial features or reflective spots are always visible. Using head-mounted cameras eliminates such a constraint since then the reflective spots are always visible, but the display is even more obtrusive. Performance-driven face synthesis is well suited for reproducing one's actions. However, the facial model only knows how to mimic one's behaviour. No new animation can be done without having to first record the actor performing the actions, which can be a disadvantage for some applications (e.g. conversational systems). This technique is not easily adaptable to lip shape computation during speech when precise control of the lip movement is required. But replaying concatenated articulation sequences is less difficult and might be more appropriate in some applications (e.g. games).

period
/ˈpɪərɪəd/, /'pI@rI@d/, [N: period], [plural: -s]. Domain: physical characterisation. Def.: The temporal interval between consecutive points of the same phase in a sinusoidal signal; the period measured in (milli)seconds is the inverse of frequency measured in (kilo)hertz.

perl
/ˈpɜːl/, /'p3:l/, [N: perl], [plural: none]. Hyperonyms: interpreted programming language. Cohyponym: python. Def.: Perl is an interpreted programming language designed for rapid programming of scripts; its main features are powerful text manipulation operations such as regular expressions, associative arrays, and ease of system access, e.g. for file and directory access and manipulation.

perplexity
/pɜːˈpleksɪti/, /p3:'pleksIti/, [N: perplexity], [plural: y/-ies]. Domain: language modelling. Def.: The (corpus) perplexity is a quantitative measure of the redundancy (or difficulty) of a recognition task for a given text corpus and a given language model. It measures how well the word sequences can be predicted by the language model.

perseverative coarticulation
/pəˈsevərətɪv kəʊɑːtɪkjuːleɪʃən/, **/p@'sev@r@tIv k@UA:tIkju:leIS@n/**, [N: [AJ: perseverative][N: coarticulation]], [plural: none]. Domain: multimodal systems. Hyperonyms: coarticulation. Synonyms: forward coarticulation, left-to-right coarticulation. Cohyponym: backward coarticulation, anticipatory coarticulation, right-to-left coarticulation. Def.: In the string ...AB..., sound A influences sound B (or beyond). L > R coarticulation is thought to be largely due to lag in articulatory movement, induced by inertia. (Clark & Yallop, p. 87)

personal-password speaker recognition system
/ˈpɜːsənəl ˈpɑːswɜːd ˈspiːkə rekəgˈnɪʃən ˈsɪstəm/, **/'p3:s@n@l 'pA:sw3:d 'spi:k@ rek@g'nIS@n 'sIst@m/**, [N: [AJ: personal][N: password][N: speaker][N: recognition][N: system]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker recognition system. Cohyponym: common-password speaker recognition system. Def.: A text-dependent speaker recognition system for which each registered speaker has his own voice password.

pharyngeal consonant
/fəˈrɪnˈdʒɪəl ˈkɒnsənənt/, **/f@'rIn'dZI@l 'kQns@n@nt/**, [N: [AJ: pharyngeal][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, labiodental consonant, dental consonant, alveolar consonant, postalveolar consonant, retroflex consonant, palatal consonant, velar consonant, uvular consonant, glottal consonant. Def.: Pharyngeal consonant is a term used in the phonetic classification of consonant sounds on the basis of their place of articulation: it refers to a sound made in the pharynx. (Crystal 1988, p. 226)

pharynx
/ˈfærɪŋks/, **/'f{rINks/**, [N: pharynx], [pural: -es]. Hyperonyms: articulator. Meronym. sup.: vocal tract. Def.: Pharynx is the tubular cavity which constitutes the throat above the larynx. (Crystal 1988, p. 226)

phase
/ˈfeɪz/, **/'feIz/**, [N: phase], [plural: -s]. Domain: physical characterisation. Hyperonyms: acoustic measure. Def.: The time displacement between two sinusoidal waveforms of the same frequency. (Clark & Yallop, p. 213)

phone
/ˈfəʊn/, **/'f@Un/**, [N: phone], [plural: -s]. Meronym. sup.: word. Def.: 1. A segment of a spoken utterance which is assigned to a single phoneme; in this role it is called an allophone of that phoneme. 2. Informally, in speech technology, a subword unit of speech that represents a particular sound.

phoneme monitoring
/ˈfəʊniːm ˈmɒnɪtərɪŋ/ , **/'f@Uni:m 'mQnIt@rIN/** , [N: [N: phoneme][N: monitoring]], [plural: none]. Domain: speech synthesis. Hyperonyms: monitoring. Cohyponym: word monitoring, syllable monitoring. Def.: Testing the intelligibility of individual sounds. (Gibbon et al. 1997, p. 490)

phoneme
/ˈfəʊniːm/, **/'f@Uni:m/**, [N: phoneme], [plural: -s]. Hyperonyms: phonological unit. Def.: 1. In traditional phonology: the smallest distinctive unit of sound in a language, where 'distinctive' means 'which distinguishes one word from another'; phoneme contrasts are often tested with minimal pairs, i.e. words differing only in one phoneme at the same position, e.g. /mi:l/ - /pi:l/ 'meal' - 'peal'. 2. In structuralist phonology: A set of phonetically similar sounds in complementary distribution (allophones). 3. In Prague and generative phonology: a bundle of distinctive features. The phoneme is the reference unit for alphabetic orthographies. Because of theoretical and practical problems with the notion of phoneme (e.g. coarticulation between adjacent phonemes, highly restricted distributional patterns within syllables, the importance of hierarchically larger units) phonological developments in the last quarter of the 20th century have not taken the phoneme to be a basic unit of linguistic description but rather an epiphenomenon. (Gibbon et al., p. 212)

phonemic transcription
/fə'niːmɪk træn'skrɪpʃən/, **/f@'ni:mIk tr{n'skrIpS@n/**, [N: [AJ: phonemic][N: transcription]], [plural: -s]. Hyperonyms: transcription. Synonyms: broad phonetic transcription. Cohyponym: phonetic transcription, orthographic transcription. Def.: A transcription in which only phonemes, i.e. minimal contrastive units in the sound system of a language, are used, and non-minimal phonetic information is excluded. Citations of phonemic transcriptions are conventionally enclosed in oblique bars (slashes), i.e. /.../. E.g. /pIn/, /pen/, /p{n/.

phonetic transcription
/fəʊ'netɪk træn'skrɪpʃən/, **/f@U'netIk tr{n'skrIpS@n/**, [N: [AJ: phonetic][N: transcription]], [plural: -s]. Hyperonyms: transcription. Hyponyms: broad phonetic transcription, narrow phonetic transcription. Cohyponym: phonemic transcription. Def.: A phonetic transcription gives details of pronunciation beyond the phonemically minimal features. The relation between the phonemic and the phonetic level can be described by general rules mapping phonemes to their detailed realisations (allophones) in specific contexts. Citations of phonetic forms are standardly delimited by square brackets [...]. (Gibbon et al. 1997, p. 209) E.g. German [?apf@l] 'apple'.

phonetically balanced sentence
/fə'netɪkəli 'bælənst 'sentəns/, **/f@'netIk@li 'b{l@nst 'sent@ns/**, [N: [AV: phonetically][AJ: balanced][N: sentence]], [plural: -s]. Domain: corpora, assessment methodologies, speaker recognition, system design. Hyperonyms: sentence. Cohyponym: phonetically rich sentence. Def.: Sentence containing phonemes according to their frequency of occurrence in a given language.

phonetically rich sentence
/fə'netɪkəli rɪtʃ 'sentəns/, **/f@'netIk@li rItS 'sent@ns/**, [N: [AV: phonetically][AJ: rich][N: sentence]] , [plural: -s]. Domain: corpora. Hyperonyms: sentence. Cohyponym: phonetically balanced sentence. Def.: Sentence containing approximately uniform phoneme frequency distributions.

phonetics
/fə'netɪks/, **/f@'netIks/**, [N: phonetics], [plural: always plural]. Meronym. sup.: linguistics. Meronym. sub.: articulatory phonetics, acoustic phonetics, auditory phonetics. Def.: The science of the sounds of human languages, their production, transmission and perception, whose methodology encompasses both the expert judgment and transcription of sound properties by a phonetician and the quantitative physical measurement and statistical evaluation of the events involved in articulation, transmission and perception. (Crystal 1988, p. 229)

phonological lexical morph
/fɒnə'lɒdʒɪkəl 'leksɪkəl 'mɔːf/, **/fQn@'lQdZIk@l 'leksIk@l 'mO:f/**, [N: [AJ: phonological][AJ: lexical][N: morph]], [plural: -s]. Domain: lexicon. Hyperonyms: lexical morph, phonological morph. Cohyponym: orthographic lexical morph. Def.: The phonological realisation of a lexical morpheme. (Gibbon et al. 1997, p. 199)

phonological morph
/fɒnə'lɒdʒɪkəl 'mɔːf/, **/fQn@'lQdZIk@l 'mO:f/**, [N: [AJ: phonological][N: morph]], [plural: -s]. Domain: lexicon. Hyperonyms: morph. Cohyponym: orthographic morph. Def.: In traditional phonology: phonological morphs are morphs consisting of phoneme sequences with a prosodic pattern (e.g. word stress). (Gibbon et al. 1997, p.215) E.g. recognition /,rek@g'nIS@n/.

phonological unit
/fɒnə'lɒdʒɪkəl 'juːnɪt/, **/fQn@'lQdZIk@l 'ju:nIt/**, [N:[AJ: phonological][N: unit]], [plural: -s]. Hyponyms: phoneme syllable. Def.: A unit of phonological description such as the distinctive feature, the phoneme, the syllable.

phonological word
/fɒnə'lɒdʒɪkəl 'wɜːd/, /fQn@'lQdZIk@l 'w3:d/, [N: [AJ: phonological][N: word]], [plural: -s]. Domain: lexicon. Hyperonyms: word. Cohyponym: orthographic word, morphological word, syntactic word, prosodic word. Def.: A word defined in terms of its conformity to the phonotactic structure of a language. (Gibbon et al. 1997, p. 196)

phonology
/fə'nɒlədʒi/, /f@'nQl@dZi/, [N: phonology], [plural: none]. Synonyms: functional phonetics (Crystal 1988). Cohyponym: phonetics, morphology, syntax, semantics, pragmatics. Meronym. sup.: linguistics. Def.: 1. Phonology is the subdivision of linguistics concerned with the descriptiion of the sound systems of languages. (Crystal 1988) 2. Phonology is the study of the units of pronunciation of a language. (Gibbon et al. 1997, p. 188)

phonostylistics
/fɒnəʊstaɪ'lɪstɪks/, /fQn@UstaI'lIstIks/, [N: phonostylistics], [plural: always plural]. Def.: Phonostylistics investigates pronunciation variants which correlate with different speech styles, such as lento or allegro speech in formal or informal contexts. (Gibbon et al. 1997, p. 191)

phonotypic transcription
/fɒnəʊ'tɪpɪk træn'skrɪpʃən/, /fQn@U'tIpIk tr{n'skrIpS@n/, [N: [AJ: phonotypic][N: transcription]], [plural: -s]. Hyperonyms: phonetic transcription. Def.: A phonotypic transcription is a specific version of the phonetic level of transcription, defined as a mapping from the phonemic level using regular phonological rules of assimilation, deletion, epenthesis. (Gibbon et al. 1997, p. 209)

phrasal idiom
/'freɪzəl 'ɪdɪəm/, /'freIz@l 'IdI@m/, [N: [AJ: phrasal][N: idiom]], [plural: -s]. Domain: lexicon. Hyperonyms: idiom. Def.: Lexical unit that is larger than the word. (Gibbon et al. 1997, p. 196)

phrase structure grammar
/'freɪz 'strʌktʃə 'græmə/, /'freIz 'strVktS@ 'gr{m@/, [N: [N: phrase][N: structure][N: grammar]], [plural: -s]. Domain: language modelling. Hyperonyms: grammar. Synonyms: context free grammar.. Def.: Phrase structure grammars contain rules which are capable not only of genertaing strings of linguistic elements, but also of providing a constituent analysis of the strings. (Crystal 1988, p. 233) Phrase structures are formalised by means of context-free grammars.

physically-based model
/'fɪzɪkəli 'beɪst 'mɒdəl/, /'fIzIk@li 'beIst 'mQd@l/, [N: [AV: physically][AJ: based][N: model]], [plural: -s]. Domain: multimodal systems . Hyperonyms: synthetic model, animation control technique. Hyponyms: structural model, muscle-based model. Cohyponym: parametric model, procedural model, free form deformation. Def.: Skin properties and muscle actions are simulated using an elastic spring mesh and forces.

pitch tracker
/'pɪtʃ 'trækə/, /'pItS 'tr{k@/, [N: [N: pitch][N: tracker]], [plural: -s]. Def.: An acoustic measuring device (hardware or software) used to measure, record and display the fundamental frequency contour of voiced speech.

pitch
/'pɪtʃ/, /'pItS/, [N: pitch], [plural: -es]. Hyperonyms: perceptual property. Def.: A perceptual property of the speech signal correlating strongly with fundamental frequency and glottal phonation rate.

playback technique
/'pleɪbæk tek'niːk/, /'pleIb{k tek'ni:k/, [N: [N: playback][N: technique]], [plural: -s]. Domain: speech synthesis, consumer off-the-shelf products. Hyperonyms: speech synthesis technique. Cohyponym: concatenation technique, production model. Def.: Synthesis by playback of pre-recorded words or phrases ('canned speech'). Generally, this provides good voice quality but low flexibility. There is no way of adopting the intonation or the voice properties; this must be implemented by pre-recording all possible voices and intonations. The vocabulary is limited by the recordings made. Sometimes a string of digits is merged into a standard carrier sentence, which provides some flexibility.

plosive
/'pləʊsɪv/, /'pl@UsIv/, [N: plosive], [plural: -s]. Hyperonyms: consonant; manner of articulation. Cohyponym: nasal, trill, tap, flap, fricative, lateral fricative, approximant, lateral approximant. Def.: Plosive is a term used in phonetic classification of consonant sounds on the basis of their manner of articulation: it refers to a sound made when a complete closure in the vocal tract is suddenly released. (Crystal 1988, p. 235) E.g. p, b, t, d, g, k].

pointing
/'pɔɪntɪŋ/, /'p0IntIN/, [N: pointing], [plural: none]. Domain: multimodal systems. Hyperonyms: gesture. Cohyponym: 2D gesture, 3D gesture. Def.: Pointing refers to the spatio-temporal relation between an indexical gesture and the object it signifies; in man-machine communication by means of a pointing device such as a mouse or finger/pen input on a touchpad or touch-sensitive screen.

poor impostor
/'pʊə ɪm'pɒstə/, /'pU@ Im'pQst@/, [N: [AJ: poor][N: impostor]], [plural: -s]. Domain: speaker recognition. Hyperonyms: impostor. Synonyms: badger. Cohyponym: skilled impostor. Def.: Impostor with a low success rate in claiming an identity averaged over each claimed identity. (Gibbon et al. 1997, p. 441)

population
/pɒpjʊ'leɪʃən/, /pQpjU'leIS@n/, [N: population], [plural: -s]. Domain: corpora, assessment methodologies, speaker recognition. Cohyponym: sample. Def.: The complete set of similar objects of which a subset (the sample) are to be subjected to statistical analysis.

Portable Document Format
/'pɔːtəbəl 'dɒkjʊmənt 'fɔːmət/, /'p0:t@b@l 'dQkjUm@nt 'f0:m@t/, [N: [AJ: Portable][N: Document][N: Format]], [plural: none]. Hyperonyms: formal language. Synonyms: PDF. Def.: PDF is a proprietary format (Adobe) for describing the page layout of documents combined with the ability to perform text searches in the document, dynamic linking of documents, multi-media content, and input via forms, e.g. for interactive documents. Generally, PDF files are much smaller than PostScript, and they may be edited. PDF has become the most widespread format for online manuals and document collections on CD-ROM, e.g. conference proceedings. Unlike PostScript, it is not coded in ASCII format but in 8-bit code.

POS
/'piː 'əʊ 'es/, /'pi: '@U 'es/, [N: POS], [plural: -es]. Domain: lexicon. Hyponyms: grammatical category, lexical category. Synonyms: part of speech; syntactic category. Def.: Part of speech (pars orationis) is the traditional term for a grammatical word class such as noun or verb. The term is the etymological source of the word parsing, i.e. the analysis of a sentence into its parts. E.g. noun, adjective, article, pronoun, verb, adverb, preposition, conjunction, interjection, proper noun, common noun, intransitive verb, transitive verb, ditransitive verb, prepositional verb .

position tracker
/pə'zɪʃən 'trækə/, /p@'zIS@n 'tr{k@/, [N: [N: position][N: tracker]], [plural: -s]. Domain: multimodal systems. Hyperonyms: device. Hyponyms: data glove. Def.: Device mounted on human body parts to capture their location.

postalveolar consonant
/pəʊstælvi'əʊlə 'kɒnsənənt/, /p@Ust{lvi'@Ul@ 'kQns@n@nt/, [N: [AJ: postalveolar][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, labiodental consonant, dental consonant, alveolar consonant, retroflex consonant, palatal consonant, velar consonant, uvular consonant, pharyngeal consonant, glottal consonant. Def.: Postalveolar consonant is a term used in the phonetic classification of consonant sounds on the basis of their place of articulation: it refers to a sound made by the front of the tongue in contact against the roof of the mouth a little behind the alveolar ridge. (Crystal 1988, p. 238)

posterior probability
/pɒs'tɪərɪə prɒbə'bɪlɪti/, /pQs'tI@rI@ prQb@'bIlIti/, [N: [AJ: posterior][N: probability]], [plural: y/-ies]. Domain: language modelling. Hyperonyms: probability. Synonyms: a posteriori probability. Cohyponym: prior probability, a priori probability. Def.: The probability that some observed event belongs to some previously established category, given all information previously established (for instance by statistical training) about with this event.

PostScript
/'pəʊstskrɪpt/, /'p@UstskrIpt/, [N: PostScript], [plural: none]. Hyperonyms: formal language. Def.: PostScript is a proprietary language for describing the page layout of documents. It is platform independent and has become the de facto standard language for laser printers. Word processors, graphics applications, etc. create PostScript files which are then transferred to a printer. PostScript features a font inclusion mechanism so that a document can be printed on any suitable printer. PostScript was developed by Adobe Corporation. The current version is PostScript level 3. PostScript files can be viewed with the popular freeware software Ghostview, but in general they cannot be edited except by an expert PostScript programmar once they have been created. Postscript is encoded in 7-bit ASCII notation, a factor in its popularity as an easily transmittable file format.

potential word
/pəʊ'tenʃəl 'wɜːd/, /p@U'tenS@l 'w3:d/, [N: [AJ: potential][N: word]], [plural: -s]. Domain: lexicon. Hyperonyms: word, lexical item. Cohyponym: actual word. Def.: A word form which is not lexicalised with a specific meaning but which in principle can be constructed on the basis of the phonotactic and morphotactic regularities of a language for the purpose of creating new terms or ad hoc words. (Gibbon et al. 1997, p. 195)

pragmatic idiom
/præg'mætɪk 'ɪdɪəm/, /pr{g'm{tIk 'IdI@m/, [N: [AJ: pragmatic][N: idiom]], [plural: -s]. Domain: lexicon. Hyperonyms: idiom. Def.: An idiom with a specific function in structuring or controlling discourse such as a greeting, a farewell, an apology.

pragmatics
/præg'mætɪks/, /pr{g'm{tIks/, [N: pragmatics], [plural: always plural]. Cohyponym: semantics, syntax. Meronym. sup.: linguistics. Def.: The study of language from the point of view of the users, especially of the choices they make, the constraints they encounter in using language in social interaction, and the effects their use of language has on the other participants in an act of communication. (Crystal 1988, p. 240)

pre-emphasis
/priː'emfəsɪs/, /pri:'emf@sIs/, [N: [N: pre-emphasis]], [plural: pre-emphases]. Def.: A filtering process applied to a speech waveform or spectrum in order to make the average power spectrum flatter than it would otherwise be by increasing the energy of higher frequencies relative to lower frequencies.

predictive model
/prɪ'dɪktɪv 'mɒdəl/, **/prI'dIktIv 'mQd@l/**, [N: [AJ: predictive][N: model]], [plural: -s]. Hyperonyms: evaluation method. Cohyponym: experimental technique, expert evaluation. Def.: Predictive models predict user behaviour and performance variables based on a theory or an empirical model. They are useful since they allow the evaluation of multimodal interfaces at the design stage. Thus, a design can be improved before implementation. On the other hand, specifying data to a predictive model may be as time consuming as the implementation. In addition, model prediction may be wrong.

prefix
/'priːfɪks/, **/'pri:fIks/**, [N: prefix], [plural: -es]. Domain: lexicon. Hyperonyms: affix. Cohyponym: suffix, circumfix. Meronym. sup.: word. Def.: A prefix is an affix attached to the beginning of a stem. E.g. stem 'select' + prefix 'pre' = 'preselect'.

PREMO
/'priːməʊ/, **/'pri:m@U/**, [N: PREMO], [plural: none]. Domain: multimodal systems. Hyperonyms: standard. Synonyms: PResentation Environments for Multimedia Objects. Def.: Standard that defines a middleware framework encompassing the management of distributed media resources, such as video, audio (both captured and synthetic), and in principle is extensible to new modalities such as haptic output and speech or gestural input. It also provides an object-oriented programming infra-structure to support the development of such applications. PREMO also serves as a reference model. The PREMO environment allows existing media devices to inter-operate, and be interfaced to an application. While the ISO MPEG specification describes the details of a video format, PREMO concentrates on how an MPEG coder/decoder can be used together with other media processing entities like a graphics renderer.

PResentation Environments for Multimedia Objects
/prezən'teɪʃən ɪn'vaɪrənmənts fə mʌlti'miːdɪə 'ɒbdʒekts/, **/prez@n'teIS@n In'vaIr@nm@nts f@ mVlti'mi:dI@ 'QbdZekts/**, [N: [N: Presentation][N: Environments][PREP: for][AJ: Multimedia][N: Objects]], [plural: none]. Domain: multimodal systems. Hyperonyms: standard. Synonyms: PREMO. Def.: Standard that defines a middleware framework encompassing the management of distributed media resources, such as video, audio (both captured and synthetic), and in principle is extensible to new modalities such as haptic output and speech or gestural input. It also provides an object-oriented programming infra-structure to support the development of such applications. PREMO also serves as a reference model. The PREMO environment allows existing media devices to inter-operate, and be interfaced to an application. While the ISO MPEG specification describes the details of a video format, PREMO concentrates on how an MPEG coder/decoder can be used together with other media processing entities like a graphics renderer.

pressure zone microphone
/'preʃə 'zəʊn 'maɪkrəfəʊn/, **/'preS@ 'z@Un 'maIkr@f@Un/**, [N: [N: pressure][N: zone][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: microphone. Cohyponym: omnidirectional microphone, unidirectional microphone,bidirectional microphone, ultradirectional microphone, headset microphone. Def.: A pressure zone microphone basically consists of an omnidirectional microphone mounted close to or into a boundary surface. The distance to the surface is significantly shorter than the wavelength given by the highest frequency to be picked up. Thus, the incident and the reflected sound will always interfere constructively, i.e. there are no comb filter distortions with this type of microphone. (Gibbon et al. 1997, p. 306)

Principal Component Analysis
/ˈprɪnsɪpəl kəmˈpəʊnənt əˈnæləsɪs/, /'prInsIp@l k@m'p@Un@nt @'n{l@sIs/, [N: [AJ: Principal][N: Component][N: Analysis]], [plural: none]. Domain: multimodal systems. Hyperonyms: template matching. Synonyms: PCA. Cohyponym: geometric template matching, deformable template matching, optical flow technique, neural network based approach. Def.: The simplest version of template matching. A test image is classified based on its (Euclidean) distance to templates generated from the faces in the training set (database). The Kurhunen-Loeve procedure and eigenfaces are based on this simple template matching method. Eigenfaces correspond to characteristic feature images and can be viewed as the principal components of a test image with respect to characteristic features obtained from the database of faces. This technique has been applied to recognise lip shapes.

prior probability
/ˈpraɪə prɒbəˈbɪlɪti/, /'praI@ prQb@'bIlIti/, [N: [AJ: prior][N: probability]], [plural: y/-ies]. Domain: language modelling. Hyperonyms: probability. Synonyms: a priori probability, priors. Cohyponym: posterior probability. Def.: The probability of observing some phenomenon estimated from previously collected training data, but independent of future observations.

procedural model
/prəˈsiːdʒərəl ˈmɒdəl/, /pr@'si:dZ@r@l 'mQd@l/, [N:[AJ: procedural][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: animation control technique, synthetic model. Cohyponym: parametric model, free form deformation. Def.: This method is not based on biological studies. Rather, the idea is to simulate the action of a muscle by a few parameters. Muscles are simulated by specialised procedures. These procedures are considered as Abstract Muscle Actions (AMAs) and can have up to 24 parameters. They work on specific facial regions that correspond to one muscle. They compute the displacement occurring under muscle contraction.

production model
/prəˈdʌkʃən ˈmɒdəl/, /pr@'dVkS@n 'mQd@l/, [N: [N: production][N: model]], [plural: -s]. Domain: speech synthesis, consumer off-the-shelf products. Hyperonyms: speech synthesis model. Cohyponym: concatenation technique, playback technique. Def.: A physical model of the vocal folds and the vocal tract used to produce sounds that resemble speech. These models are often LPC based (specifying sounds by formants in position and width). The voice quality is not as good as in the playback or concatenation technique, but since this technique comprises pure synthesis, every parameter is controlled. Thus the change of voice characteristics, pitch, intonation, and stress can be exploited by the system. The vocabulary is limited by the pronunciation rules.

pronunciation lexicon
/prənʌnsiˈeɪʃən ˈleksɪkən/ , /pr@nVnsi'eIS@n 'leksIk@n/ , [N: [N: pronunciation][N: lexicon]], [plural: pronunciation lexica ,-s]. Domain: lexicon. Hyperonyms: lexicon. Synonyms: pronunciation dictionary, pronunciation table. Meronym. sup.: acoustic-phonetic model. Def.: A table containing pairs of orthographic and phonemic representations of words, sometimes with variant orthographies and pronunciations, and either in book form or as a database.

pronunciation table
/prənʌnsiˈeɪʃən ˈteɪbəl/, /pr@nVnsi'eIS@n 'teIb@l/, [N: [N: pronunciation][N: table]], [plural: -s]. Domain: lexicon. Hyperonyms: database type. Synonyms: pronunciation dictionary, pronunciation lexicon. Def.: A pronunciation table defines the relation between orthographic and phonemic representations of words. Often they are defined as functions which assign pronunciations (frequently a set of variant pronunciations) to orthographic representations (Gibbon et al. 1997, p. 226)

prosodic feature
/prəˈzɒdɪk ˈfiːtʃə/, /pr@'zQdIk 'fi:tS@/, [N: [AJ: prosodic][N: feature]], [plural: -s]. Hyponyms: pitch, loudness, tempo, rhythm. Def.: A term used in suprasegmental phonetics and phonology to refer collectively to variations in pitch, loudness, tempo and rhythm. (Crystal 1988, p. 249)

prosodic lexicon
/prə'zɒdɪk 'leksɪkən/, **/pr@'zQdIk 'leksIk@n/**, [N: [AJ: prosodic][N: lexicon]], [plural: prosodic lexica, -s]. Domain: lexicon. Hyperonyms: lexicon. Def.: A lexicon of prosodic units such as terminal pitch contours (rises, falls, fall-rises), with their meanings.

prosodic parsing
/prə'zɒdɪk 'pɑːzɪŋ/, **/pr@'zQdIk 'pA:zIN/**, [N: [AJ: prosodic][N: parsing]], [plural: none]. Domain: speech recognition, speech synthesis. Hyperonyms: parsing. Cohyponym: morphological parsing, syntactic parsing. Def.: 1. In speech recognition: Prosodic parsing is the analysis of the speech signal in respect of the fundamental frequency (F0) trajectory in relation to words, sentences and dialogue units. 2. In speech synthesis: Prosodic parsing is the analysis of sentence structure for the generation of intonation patterns in speech synthesis. (Gibbon et al. 1997, p. 210)

prosodic transcription
/prə'zɒdɪk træn'skrɪpʃən/, **/pr@'zQdIk tr{n'skrIpS@n/**, [N: [AJ: prosodic][N: transcription]], [plural: -s]. Hyperonyms: transcription. Def.: A transcription of the prosodic features of an utterance such as stress/accent, intonation, boundaries, using a prosodic transcription system or alphabet such as ToBI or SAMPROSA.

prosodic word
/prə'sɒdɪk 'wɜːd/, **/pr@'sQdIk 'w3:d/**, [N: [AJ: prosodic][N: word]], [plural: -s]. Domain: lexicon. Hyperonyms: word. Cohyponym: orthographic word, phonological word, morphological word, syntactic word. Def.: Word based on its conformity to the accentuation and the rhythm patterning of the language. (Gibbon et al. 1997, p. 196)

prosody
/'prɒsədi/, **/'prQs@di/**, [N: prosody], [plural: –]. Synonyms: non-segmental phonology. Meronym. sup.: phonology. Def.: Prosody covers all properties of pronunciation which are not directly concerned with defining consonants and vowels. Prosody in this sense covers, for example, syllable structure and phonological word phonotactics, as well as the more traditional categories of intonation, accent, and phrasing. In the lexicon, prosodic information is in general restricted to the prosodic properties of words, such as stress position (e.g. in English, Dutch, and German words), or tonal accent words (e.g. in Swedish), or to rhythmically relevant units such as the syllable and the foot. (Gibbon et al. 1997, p. 210-211)

pseudo-impostor bundle
/'sjuːdəʊ ɪm'pɒstə 'bʌndəl/, **/'sju:d@U Im'pQst@ 'bVnd@l/**, [N: [AJ: pseudo][N: impostor][N: bundle]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker group. Def.: The group of speakers who have been used to build the impostor model of a given registered speaker. (Gibbon et al. 1997, p. 421)

pseudo-impostor
/'sjuːdəʊ ɪm'pɒstə/, **/'sju:d@U Im'pQst@/**, [N: [AJ: pseudo][N: impostor]], [plural: -s]. Domain: speaker recognition. Hyperonyms: impostor. Synonyms: background speaker. Def.: Speaker used to model the impostor during the registration phase. (Gibbon et al. 1997, p. 421)

psycholinguistics
/saɪkəʊlɪŋ'gwɪstɪks/, **/saIk@UlIN'gwIstIks/**, [N: psycholinguistics], [plural: always plural]. Meronym. sup.: linguistics. Def.: The study of the mental processes underlying the planning, production, perception and comprehension of speech. (Crystal 1988, p. 251)

Pulse coded modulation Expansion Bus
/p'ʌls 'kəʊdɪd mɒdjʊ'leɪʃən ɪks'pænʃən 'bʌs/, **/p'Vls 'k@UdId mQdjU'leIS@n Iks'p{nS@n 'bVs/**, [N: [N: Pulse][AJ: coded][N: modulation][N: Expansion] [N: Bus]], [plural: -es]. Domain: system design. Hyperonyms: bus. Synonyms: PEB. Def.: PEB is seen as an internal switching matrix capable of routing any time slot to an adequate audio port of the speech recogniser.

puppeteer control face synthesis
/pʌpə'tɪə kən'trəʊl 'feɪs 'sɪnθəsɪs/, **/pVp@'tI@ k@n'tr@Ul 'feIs 'sInT@sIs/**, [N: [N: puppeteer][N. control][N: face][N: synthesis]], [plural: none]. Domain: multimodal systems. Hyperonyms: face synthesis. Cohyponym: audio-driven face synthesis, performance-driven face synthesis, text-to-visual-speech face synthesis. Def.: A puppeteer moves input devices such as a data glove or joystick, or uses a keyboard to drive a facial model. Each input device control is associated with a facial parameter. For example, a key or a hand shape corresponds to a particular facial expression: raising eyebrows or opening the mouth. As the puppeteer moves the hand or presses different keys, the facial model moves accordingly. This technique is often used for real-time applications and movies.

puppeteer control
/pʌpə'tɪə kən'trəʊl/, **/pVp@'tI@ k@n'tr@Ul/**, [N: [N: puppeteer][N: control]], [plural: none]. Domain: multimodal systems. Def.: A specific technical device, such as a data glove, controls the parameters of the facial model.

python
/'paɪθən/, **/'paIT@n/**, [N: python], [plural: none]. Hyperonyms: object-oriented programming language. Cohyponym: perl. Def.: An object-oriented programming language designed to overcome the limited data modelling capabilities of perl. One of its distinguishing features is the built-in interface to many windowing environments. Python is freely available for mosts platforms.

qualitative evaluation
/'kwɒlɪtətɪv ɪvælju'eɪʃən/, **/'kwQlIt@tIv Iv{ljU'eIS@n/**, [N: [AJ: qualitative][N: evaluation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: evaluation. Cohyponym: quantitative evaluation. Def.: Qualitative evaluation tests the intelligibility of the system. The amount of intelligibility a synthetic model adds during speech recognition tests is compared to the amount of intelligibility a human speaker adds during the same tests. The test is performed in different audiovisual situations: audio alone (degraded or normal audio), visual alone (of the synthetic actor and of the human subject), and audio-visual combined (of the synthetic actor and of the human subject). Benoit and his team included also the following conditions: lip alone of the synthetic face, jaw and lip alone of the synthetic face, subject's lips. The audio stimuli can be degraded by adding noise. For each setting a confusion matrix is established. The comparison over these matrices gives the overall intelligibility of each phonemic item in each setting.

quantisation
/kwɒntaɪ'zeɪʃən/, **/kwQntaI'zeIS@n/**, [N: quantisation], [plural: -s]. Meronym. sub.: Pulse Code Modulation. Def.: Digital encoding of amplitude.

quantitative evaluation
/'kwɒntɪtətɪv ɪvælju'eɪʃən/, **/'kwQntIt@tIv Iv{ljU'eIS@n/**, [N: [AJ: quantitative][N: evaluation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: evaluation. Cohyponym: qualitative evaluation. Def.: Quantitative evaluation compares computed values with real values. For example, values of lip height and lip width parameters of a synthetic face can be compared with the same values obtained from the analysis of a human subject. Image analysis or FACS can be used to analyse and compare muscle contraction from real and synthetic images. The weighting of different parameters and the definition of equalness in real and synthesised parameters is still a problematic open issue (e.g. lip width could be more important than upper lip raiser).

QuickTime
/'kwɪktaɪm/, /'kwIktaIm/, [N: QuickTime], [plural: none]. Hyperonyms: file format. Def.: A meta file format for multi-media data and a toolbox for accessing this data. QuickTime was developed by Apple Computer, and is available for both Windows and Macintosh operating systems (for other operating systems, a subset of the QuickTime functionality is accessible). The basic metaphor underlying QuickTime is that of a multi-track recording, where each track may contain text, graphics, audio, or video data in a large variety of formats, including streaming audio and video, and MPEG data. The tracks are synchronised, and may be switched on or off for playback, e.g. to play movies in different languages. The current version is QuickTime 3.0, and simple players and plug-ins for web browsers can be downloaded free of charge. QuickTime is supported by most multi-media editing tools, and a system development kit may be licensed from Apple.

RAID
/'reɪd/, /'reId/, [N: RAID], [plural: none]. Synonyms: Redundant Array of Inexpensive Disks. Def.: In a RAID array, several hard disks are combined in such a way that failure or removal of a disk does not interrupt the operation of the array as a whole. This is possible by distributing data over the individual hard disks, and by data duplication. Several RAID levels have been specified. They differ in the degree of redundancy and safety.

ram
/'ræm/, /'r{m/, [N: ram], [plural: -s]. Domain: speaker recognition. Hyperonyms: registered speaker. Synonyms: resistant speaker. Cohyponym: vulnerable speaker. Def.: A registered speaker with a low mistrust rate. (Gibbon et al. 1997, p. 433)

rapid prototyping
/'ræpɪd 'prətətaɪpɪŋ/, /'r{pId 'pr@t@taIpIN/, [N: [AJ: rapid][N: prototyping]], [plural: none]. Hyperonyms: experimental technique. Synonyms: iterative design. Cohyponym: benchmark evaluation, user study, simulation study. Def.: Rapid prototyping has been widely adopted in the field of human-computer interaction, especially for product development. It is suitable for the development of multimodal applications, since many detail implementation issues can be explored rather quickly. The iterative design cycle includes (re)design of the application, implementation, and (informal) user testing. Iterative design is highly desirable from the HCI point of view but is difficult to reconcile with the pipeline or cascaded process organisation in software development which is currently still predominant, for reasons of cost control mainly.

read speech
/'red 'spiːtʃ/, /'red 'spi:tS/, [N: [AJ: read][N: speech]], [plural: none]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: speaking style. Cohyponym: spontaneous speech, dictation speech. Def.: This speaking style is that of a radio or television news reader, somebody giving a talk to a literature society, or someone who gives a poor presentation at a conference. Although in real life this speaking style hardly occurs, many dictation systems are trained on this type of material. The style is characterised by well formulated sentences, very few hesitations and a more or less predictable intonation.

receiver operating characteristic curve
/rɪ'siːvər 'ɒpəreɪtɪŋ kærəktə'rɪstɪk 'kɜːv/, /rI'si:v@r 'Qp@reItIN k{r@kt@'rIstIk 'k3:v/, [N: [N: receiver][V: operating][AJ: characteristic][N: curve]], [plural: -s]. Domain: speaker recognition. Synonyms: ROC curve. Def.: A curve that plots the tradeoff between false alarms and false rejections in, for example, a speaker verification or a wordspotting system.

recognition accuracy
/rekəg'nɪʃən 'ækjʊrəsi/, /rek@g'nIS@n '{kjUr@si/, [N: [N: recognition][N: accuracy]], [plural: y/-ies]. Domain: speech synthesis, speech recognition, consumer off-the-shelf products. Hyperonyms: performance measure. Synonyms: accuracy. Cohyponym: OOV-rejection, error recovery, response time, situational awareness. Def.: The accuracy for a word recognition system is defined as the number of correctly recognised words divided by the number of words in the test.

recognition component
/rekəg'nɪʃən kəm'pəʊnənt/ , **/rek@g'nIS@n k@m'p@Un@nt/** , [N: [N: recognition][N: component]], [plural: -s]. Domain: speech recognition. Cohyponym: search component. Meronym. sup.: spoken language recognition system. Def.: In the recognition component, intervals of the speech signal are mapped by probabilistic systems such as Hidden Markov Models, Neural Networks, Dynamic Programming algorithms, Fuzzy Logic knowledge bases, to word hypotheses; the resulting mapping is organised as a word lattice or word graph, i.e. a set of word hypotheses, each assigned in principle to a temporal interval in the speech signal. (Gibbon et al. 1997, p. 190)

recording studio
/rɪ'kɔːdɪŋ 'stjuːdɪəʊ/, **/rI'kO:dIN 'stju:dI@U/**, [N: [N: recording][N: studio]], [plural: -s]. Domain: physical characterisation. Hyperonyms: recording room. Cohyponym: laboratory room, soundproof booth, anechoic chamber. Def.: Speech recordings may be made in a professional recording studio. The advantage of this type of recording environment is that it is widely available and that the recording location may be rented only for the recording sessions. The major disadvantage of using a recording studio is that the recording conditions and especially the acoustic conditions are not standardised in any way. Moreover, it will generally not be possible to design the acoustic environment of the recording room according to the needs of speech recordings. (Gibbon et al. 1997, p. 309/310)

redundancy
/rɪ'dʌndənsi/, **/rI'dVnd@nsi/**, [N: redundancy], [plural: none]. Domain: multimodal systems. Hyperonyms: cooperation type. Cohyponym: complementarity, equivalence, specialisation, concurrency, transfer. Def.: The same chunk of information is transmitted using more than one modality. E.g. A costumer saying "I want the second item on the right", simultaneously pointing in that direction..

Redundant Array of Inexpensive Disks
/'ræpɪd ə'reɪ əv ɪnɪk'spensɪv 'dɪsks/, **/'r{pId @'reI @v InIk'spensIv 'dIsks/**, [N: [AJ: Redundant][N: Array][PREP: of][AJ: Inexpensive][N: Disks]], [plural: always plural]. Synonyms: RAID. Def.: In a RAID array, several hard disks are combined in such a way that failure or removal of a disk does not interrupt the operation of the array as a whole. This is possible by distributing data over the individual hard disks, and by data duplication. Several RAID levels have been specified. They differ in the degree of redundancy and safety.

register
/'redʒɪstə/, **/'redZIst@/**, [N: register], [plural: -s]. Hyperonyms: language variety. Meronym. sup.: natural language. Def.: An occupation-oriented functional language variety such as the language of .

registered speaker
/'redʒɪstəd 'spiːkə/ , **/'redZIst@d 'spi:k@/** , [N: [AJ: registered][N: speaker]], [plural: -s] . Domain: speaker recognition. Hyperonyms: speaker. Hyponyms: dependable speaker, mistaken speaker, violated speaker, casual registered speaker. Synonyms: reference speaker, valid speaker, authorised speaker, subscriber, client. Cohyponym: impostor, non-registered speaker. Def.: A speaker who belongs to the list of registered users for a given speaker recognition system (usually a speaker who is entitled to use the facilities, the access of which is restricted by the system). (Gibbon et al. 1997, p. 413)

regular grammar
/ˈkɒntekst ˈfriː ˈgræmə/, /'kQntekst 'fri: 'gr{m@/, [N: [N: context][AJ: free][N: grammar]], [plural: -s]. Domain: language modelling. Hyperonyms: grammar, context-free grammar. Hyponyms: deterministic regular grammar, nondeterministic regular grammar. Cohyponym: finite-state automaton. Def.: A regular grammar is a set of rules which defines linear structures over strings of symbols from a vocabulary V. It is defined formally as a quadrulple <N,T,S,R>, where N is a finite set of nonterminal symbols in V, T is a finite set of terminal symbols in V, S is a start symbol (defining the root of the tree structures) in N, and R is a set of rules either of the form A -> gB or the form A -> Bg (but not mixed), where A is an element of N, g is a non-zero string of symbols from T. Regular grammars are also known as Type 3 grammars in the Chomsky hierarchy of formal grammars. Hidden Markov Models are stochastic (probabilistic) regular grammars in which terminal symbols and transitions are annotated with application probabilities on the basis of corpus analyses. Regular grammars are weakly equivalent to, and processed by, finite state automata. (Crystal 1988, p. 71)

rejection
/rɪˈdʒekʃən/, /rI'dZekS@n/, [N: rejection], [plural: -s]. Domain: speaker recognition. Hyperonyms: decision outcome (of a speaker recognition system). Hyponyms: false rejection. Cohyponym: acceptance. Def.: 1. Decision outcome which consists in refusing to assign a registered identity (or class) in the context of open-set speaker identification or classification, or which consists in responding negatively to a speaker (class) verification trial. 2. The decision of an automatic speech recognition (ASR) device that the input (or part of the input) cannot be mapped onto one or more words in the vocabulary with sufficient confidence. This results in the failure to recognise the input.

resistant speaker
/rɪˈzɪstənt ˈspiːkə/, /rI'zIst@nt 'spi:k@/, [N: [AJ: resistant][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: registered speaker. Synonyms: ram. Def.: A registered speaker with a low mistrust rate. (Gibbon et al. 1997, p. 433)

resonance disorder
/ˈrezənəns dɪsˈɔːdə/, /'rez@n@ns dIs'O:d@/, [N: [N: resonance][N: disorder]], [plural: -s]. Domain: corpora. Hyperonyms: speech disorder. Cohyponym: articulation disorder, voice disorder, language disorder, rhythm disorder. Def.: A resonance disorder involves lesions of the oral, nasal, or laryngeal cavities. Apart from functional causes, resonance disorders can result from surgical removal of the tonsils, a cleft palate, or nose polyps. (Gibbon et al. 1997, p. 114)

restricted language
/rɪˈstrɪktɪd ˈlængwɪdʒ/, /rI'strIktId 'l{ngwIdZ/, [N: [AJ: restricted][N: language]], [plural: -s]. Hyperonyms: natural language. Def.: A variety of natural language which is restricted by externally imposed rules of use. These rules typically limit the vocabulary and the range of acceptable syntactic constructions. Restricted languages tend to be used in contexts where rapid, effective communication of a small set of basic facts is paramount, for example, in air traffic control. Because of the tightly constrained nature of restricted languages, they are seen by many to be good candidates for modelling interactive dialogue systems. However, this advantage must be weighed against the safety-critical function of many such languages in real use.

retroflex consonant
/ˈretrəʊfleks ˈkɒnsənənt/, /'retr@Ufleks 'kQns@n@nt/, [N: [AJ: retroflex][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, labiodental consonant, dental consonant, alveolar consonant, postalveolar consonant, palatal consonant, velar consonant, uvular consonant, pharyngeal consonant, glottal consonant. Def.: Retroflex consonant is a term used in the phonetic classification of consonant sounds on the basis of their place of articulation: it refers to a sound made when the tip of the tongue is curled back in direction of the front part of the hard palate. (Crystal 1988, p. 265)

right-to-left coarticulation
/ˈraɪt tə ˈleft kəʊɑːtɪkjʊˈleɪʃən/, /'raIt t@ 'left k@UA:tIkjU'leIS@n/, [N: [AJ: right][PREP: to][AJ: left][N: coarticulation]], [plural: none]. Domain: multimodal systems. Hyperonyms: coarticulation. Synonyms: anticipatory coarticulation, backward coarticulation. Cohyponym: forward coarticulation, perseverative coarticulation, left-to-right coarticulation. Def.: In the string ...CD..., sound D influences sound C (or earlier sounds). L < R coarticulation is thought to be due to deliberate high-level organisation of the neuromuscular commands for the relevant sounds. This high-level planning is complicated by the differences in innervation latencies among the various articulatory muscle systems. (Clark & Yallop, p. 87) E.g. [S] in [Su:] 'shoe' is rounded, anticipating the lip rounding of the vowel..

ROC curve
/ɑːrəʊˈsiː kɜːv/, /A:r@U'si: k3:v/, [N: ROC curve], [plural: -s]. Domain: speaker recognition. Synonyms: receiver operating characteristic curve. Def.: A curve that plots the tradeoff between false alarms and false rejections in, for example, a speaker verification or a wordspotting system.

root
/ˈruːt/, /'ru:t/, [N: root], [plural: -s]. Domain: lexicon. Hyperonyms: morph. Synonyms: basis. Def.: Roots are the morphs which realise lexical morphemes and inflectable grammatical morphemes, and function as the smallest type of stem in derivation and compounding. (Gibbon et al. 1997, p. 215)

sample
/ˈsæmpəl/, /'s{mp@l/, [N: sample], [plural: -s]. Meronym. sup.: population. Def.: Typically, a measure cannot be taken on all units of a population. In these cases, a sample is taken. Provided precautions are taken, this sample may be used to study the variable of concern in the population.

sampling rate
/ˈsæmplɪŋ ˈreɪt/, /'s{mplIN 'reIt/, [N:[N: sampling][N: rate]], [plural: -s]. Hyponyms: Nyquist rate. Def.: The number of speech samples taken per unit time. The maximum frequency (F(max)) encoded is directly determined by the sampling rate. If T is the time in seconds between successive samples, then the maximum frequency is the reciprocal of 2T: F(max) = 1/2 T. (Clark & Yallop 1995, p. 258)

sampling
/ˈsæmplɪŋ/, /'s{mplIN/, [N: sampling], [plural: none]. Def.: The digital encoding of the instantaneous values of the amplitude at regular discrete intervals of time along the speech time domain waveform is known as the process of sampling. (Clark & Yallop 1995, p. 258)

scale normalisation
/ˈskeɪl nɔːməlaɪˈzeɪʃən/, /'skeIl nO:m@laI'zeIS@n/, [N: [N: scale][N: normalisation]], [plural: none]. Domain: multimodal systems. Def.: 1. Technique used to make quantities comparable by referring them to a range of standardised values. 2. Technique used in face recognition to ensure that the face to be recognised and the face stored in the database (the two are then compared) are of the same size. Scale normalisation can be achieved by locating both eyes in the image and by applying rotation, translation and scaling to align them with reference faces.

scenario
/səˈnɑːrɪəʊ/, /s@'nA:rI@U/, [N: scenario], [plural: -s]. Domain: dialogue representation. Def.: The various practical conditions and attendant circumstances which affect the collection of dialogue data. Such conditions are important to keep track of, since they might have an effect (foreseen or unforeseen) on the value of the corpus as a basis for further research and development.

schwa
/'ʃwɑː/, /'SwA:/, [N: schwa], [plural: -s]. Hyperonyms: vowel. Def.: Schwa is the usual name for the neutral (mid central unrounded) vowel /@/. (Crystal 1988) The schwa is characterised by a relaxed position of the tongue, and evenly spaced formants at about 500, 1000, 1500 Hz. E.g. In English heard at the beginning of words, e.g. ago, amaze, or in the middle of words, e.g. afterwards. (Crystal 1988) .

SCR
/'es 'siː 'ɑː/, /'es 'si: 'A:/, [N: SCR], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: ratio. Synonyms: system correction rate. Def.: The percentage of all system turns which are correction turns.

script language
/'skrɪpt 'læŋgwɪdʒ/, /'skrIpt 'l{NgwIdZ/, [N: [N: script][N: language]], [plural: -s]. Hyperonyms: programming language. Hyponyms: JavaScript, ECMAScript. Synonyms: scripting language. Cohyponym: embedded programming language. Def.: A script language or scripting language is an interpreted programming language that is primarily used for quick programming by system administrators and inter- or intra-application communication. Script languages may run on their own and call other applications or even access functions inside these applications. The most well-known examples are the UNIX shell scripting languages bash, awk, sed, perl, and batch programming languages for more elementary operating systems.

SDL
/spesɪfɪ'keɪʃən 'ænd dɪ'skrɪpʃən 'læŋgwɪdʒ/, /spesIfI'keIS@n '{nd dI'skrIpS@n 'l{NgwIdZ/, [N: [N: Specification][C: and][N: description][N: language]], [plural: -s]. Domain: interactive dialogue systems. Synonyms: specification and description language. Def.: A graphical language for describing state transition diagrams for event-driven systems. It was standardised by CCITT. (Gibbon et al. 1997, p. 573)

search component
/'sɜːtʃ kəm'pəʊnənt/, /'s3:tS k@m'p@Un@nt/, [N: [N: search][N: component]], [plural: -s]. Domain: speech recognition, language modelling. Cohyponym: recognition component. Meronym. sup.: spoken language recognition system. Def.: The search component enhances the information from the speech signal with top-down information from a language model in order to narrow down the lexical search space. (Gibbon et al. 1997, p. 190)

search engine
/'sɜːtʃ 'endʒɪn/, /'s3:tS 'endZIn/, [N: [N: search][N: engine]], [plural: -s]. Hyperonyms: software. Synonyms: search robot. Meronym. sub.: search robot. Def.: Software that downloads and traverses linked world-wide web pages and indexes the information it finds on these pages in a database, permitting fast full text search of the WWW.

search robot
/'sɜːtʃ 'rəʊbɒt/, /'s3:tS 'r@UbQt/, [N: [N: search][N: robot]], [plural: -s]. Hyperonyms: software. Meronym. sup.: search engine. Def.: The part of a search engine which traverses web sites in search of pages for indexing.

search
/'sɜːtʃ/, /'s3:tS/, [N: search], [plural: -es]. Domain: language modelling. Hyponyms: single best sentence search, word graph search . Def.: The mapping of a key or search category to a value or set of values in a search space.

segment
/'segmənt/, /'segm@nt/, [N: segment], [plural: -s]. Domain: speech synthesis, corpora. Hyperonyms: speech unit. Hyponyms: consonant, vowel. Def.: Segments are temporal intervals in speech sounds; usually the term refers to the consonants and vowels of in the sound system of a language.

segmental quality
/seg'mentəl 'kwɒlɪti/, /seg'ment@l 'kwQlIti/, [N: [AJ: segmental][N: quality]], [plural: y/-ies]. Domain: speech synthesis. Cohyponym: prosodic quality. Def.: The quality of the phoneme-synthesis factors in a speech synthesiser.

segmentation
/segmen'teɪʃən/, /segmen'teIS@n/, [N: segmentation], [plural: -s]. Domain: corpora, speech synthesis, assessment methodologies. Hyponyms: manual segmentation, automatic segmentation, semi-automatic segmentation, morphological segmentation, phonological segmentation. Def.: The procedure of isolating minimal distinctive temporal phonetic segments (phones). (Gibbon et al. 1997, p. 206)

semantic fusion
/sə'mæntɪk 'fjuːʒən/, /s@'m{ntIk 'fju:Z@n/, [N: [AJ: semantic][N: fusion]], [plural: -s]. Domain: multimodal systems. Hyperonyms: fusion. Synonyms: semantic-level fusion. Cohyponym: signal-level fusion. Def.: Semantic fusion performs the combination of multimodal input at the meaning level. Semantic fusion of multimodal input proceeds in two steps. First, input events in different modalities are combined in a low-level interpretation module by grouping input events in different modalities to multimodal input events. Next, the multimodal input event is passed on to the high-level interpretation module to derive the meaning of multimodal input events by extracting and combining the information chunks. Thus, the high-level interpretation module determines what type of action the user wants to trigger, and what its parameters are. This parametrised action is then passed to the application's dialogue manager that can initiate the execution of the intended action.

semantic-level fusion
/sə'mæntɪk 'levəl 'fjuːʒən/, /s@'m{ntIk 'lev@l 'fju:Z@n/, [N: [AJ: semantic][N: level][N: fusion]], [plural: -s]. Domain: multimodal systems. Hyperonyms: fusion. Synonyms: semantic fusion. Cohyponym: signal-level fusion. Def.: Semantic fusion performs the combination of multimodal input at the meaning level. Semantic fusion of multimodal input proceeds in two steps. First, input events in different modalities are combined in a low-level interpretation module by grouping input events in different modalities to multimodal input events. Next, the multimodal input event is passed on to the high-level interpretation module to derive the meaning of multimodal input events by extracting and combining the information chunks. Thus, the high-level interpretation module determines what type of action the user wants to trigger, and what its parameters are. This parametrised action is then passed to the application's dialogue manager that can initiate the execution of the intended action.

semantics
/sə'mæntɪks/, /s@'m{ntIks/, [N: semantics], [plural: –]. Cohyponym: phonology, phonetics, morphology, syntax, pragmantics. Meronym. sup.: linguistics. Def.: Semantics is the study of the meaning in language. (Crystal 1988, p. 273)

semivowel
/'semivaʊəl/, /'semivaU@l/, [N: semivowel], [plural: -s]. Hyperonyms: consonant. Def.: A sound functioning as a consonant but lacking the phonetic characteristics normally associated with consonants (such as friction or closure); instead, its quality is phonetically that of a vowel, though, occurring as it does at the margins of a syllable, its duration is much less than that typical of vowels. (Crystal 1988, p. 276)

sentence accent
/'sentəns 'æksənt/, /'sent@ns '{ks@nt/, [N: [N: sentence][N: accent]], [plural: -s]. Domain: corpora, speech synthesis. Hyperonyms: accent. Synonyms: contrastive accent (Crystal 1988, p. 2). Cohyponym: word accent. Def.: Sentence accent is the emphasis which makes a particular word or syllable stand out [in a sentence]. (Crystal 1988, p. 2)

sentence error rate
/'sentəns 'erə 'reɪt/, /'sent@ns 'er@ 'reIt/, [N: [N: sentence][N: error][N: rate]], [plural: -s]. Domain: speech recognition. Hyperonyms: error rate. Cohyponym: word error rate. Def.: Proportion of utterances/sentences that contain at least one recognition error.

sentence syntax
/'sentəns 'sɪntæks/, /'sent@ns 'sInt{ks/, [N: [N: sentence][N: syntax]], [plural: none]. Domain: lexicon. Hyperonyms: syntax. Def.: Sentence syntax defines the structure of a (generally unlimited) set of sentences.

sex identification
/'seks aɪdentɪfɪ'keɪʃən/, /'seks aIdentIfI'keIS@n/, [N: [N: sex][N: identification]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker classification task. Cohyponym: age identification, health state identification, mood identification, accent identification, speaker cluster selection. Def.: If the goal is to decide whether a given speech utterance was uttered by a male speaker or a female speaker, this particular problem of speaker classification can be referred to as sex identification. (Gibbon et al. 1997, p. 408)

SGML
/'es 'dʒiː 'em 'el/, /'es 'dZi: 'em 'el/, [N: SGML], [plural: none]. Hyperonyms: formal language. Hyponyms: eXtended Markup Language (XML). Synonyms: Standard Generalized Markup Language. Def.: An ISO standard (ISO 8879:1986) for the description of text by its structure.

sheep
/'ʃiːp/, /'Si:p/, [N: sheep], [plural: –]. Domain: speaker recognition. Hyperonyms: registered speaker. Synonyms: dependable speaker. Cohyponym: goat, unreliable speaker. Def.: A registered speaker with a low misclassification rate. (Gibbon et al. 1997, p. 432)

shimmer
/'ʃɪmə/, /'SIm@/, [N: shimmer], [plural: -s]. Domain: physical characterisation. Def.: Shimmer is a measure of the average perturbation of someone's fundamental frequency and of its magnitude.

sign language
/'saɪn 'læŋgwɪdʒ/, /'saIn 'l{NgwIdZ/, [N: [N: sign][N: language]], [plural: -s]. Domain: multimodal systems. Meronym. sub.: handshape, hands placement, hands orientation, hands movement, facial expression, body gestures. Def.: Exclusively visual representation of words by (sequences of) hand and face gestures. Also the environment, e.g. people present in the scene, are relevant. Sign language is used by hearing-impaired people to communicate.

signal detection theory
/'sɪgnəl dɪ'tekʃən 'θiːəri/, /'sIgn@l dI'tekS@n 'Ti:@ri/, [N: [N: signal][N: detection][N: theory]], [plural: y/-ies]. Domain: assessment methodologies. Hyperonyms: theory. Def.: A model that may be used for studying speech recogniser performance. The basic idea behind signal detection theory is that errors convey information concerning how the system is operating (in this respect, it is an advance on simple error measures).

signal-level fusion
/'sɪgnəl 'levəl 'fjuːʒən/, /'sIgn@l 'lev@l 'fju:Z@n/, [N: [N: signal][N: level][N: fusion]], [plural: -s]. Domain: multimodal systems. Hyperonyms: fusion. Synonyms: lexical fusion. Cohyponym: semantic-level fusion, semantic fusion. Def.: Signal-level fusion performs the combination of multimodal input at the level of the input signal. Signal-level fusion has to date been tried for audio-visual speech recognition, combining speech as audio signals and lip movements as visual signals. Other types of signal-level fusion have been explored in the robotics field (e.g. combining image data with other sensor input, such as laser ranger finders, or infrared sensors.

signal-to-noise ratio
/ˈsɪgnəl tə ˈnɔɪz ˈreɪʃəʊ/, **/’sIgn@l t@ ’nOIz ’reIS@U/**, [N: [N: signal][PREP: to][N: noise][N: ratio]] , [plural: -s]. Domain: physical characterisation. Hyperonyms: ratio. Def.: The ratio of information-carrying signals such as speech to background noise, expressed in dB.

Signalyze
/ˈsɪgnəlaɪz/, **/’sIgn@laIz/**, [N: Signalyze], [plural: none]. Hyperonyms: software. Def.: Data analysis, display, segmentation and labelling software for speech signal processing on the Macintosh.

simplex word
/ˈsɪmpleks ˈwɜːd/, **/’sImpleks ’w3:d/**, [N: [AJ: simplex][N: word]], [plural: -s]. Domain: lexicon. Hyperonyms: word. Cohyponym: composite word, compound, complex word. Meronym. sup.: word formation. Def.: A word that is not derived or composed of other words. (Bussmann, p. 686) E.g. simplex word ’blue’ + simplex word ’berry’ = complex word ’blueberry’.

simulation study
/sɪmjʊˈleɪʃən ˈstʌdi/, **/sImjU’leIS@n ’stVdi/**, [N: [N: simulation][N: study]], [plural: y/-ies]. Hyperonyms: experimental technique. Cohyponym: benchmark evaluation, user study, rapid prototyping, iterative design. Def.: System performance that is not yet feasible can be simulated, and thus systems and issues in human-computer interaction can be examined without having to first implement a system. The Wizard-of-Oz technique is widely accepted for simulation studies.

single-stroke gesture
/ˈsɪŋgəl ˈstrəʊk ˈdʒestʃə/, **/’sINg@l ’str@Uk ’dZestS@/**, [N: [AJ: single][N: stroke][N: gesture]], [plural: -s]. Domain: multimodal systems. Hyperonyms: gesture. Cohyponym: multi-stroke gesture. Def.: Gesture consisting of one stroke, i.e. the smallest meaningful unit of gesture input is one stroke.

singleton event
/ˈsɪŋgəltən ɪˈvent/, **/’sINg@lt@n I’vent/**, [N: [N: singleton][N: event]], [plural: -s]. Domain: language modelling. Hyperonyms: event. Cohyponym: doubleton event, unseen event. Def.: Event that was observed exactly once. (Gibbon et al. 1997, p. 249)

situational awareness
/sɪtjʊˈeɪʃənəl əˈweənəs/, **/sItjU’eIS@n@l @’we@n@s/**, [N: [AJ: situational][N: awareness]], [plural: none]. Domain: speech synthesis, speech recognition, consumer off-the-shelf products. Hyperonyms: performance measure. Cohyponym: recognition accuracy, OOV-rejection, error recovery, response time. Def.: Users that give commands to a system have a certain expectation of what they can say. This might depend on the internal state of the system (‘active vocabulary’), but if the user is not aware of that state, for whatever reason, it is said that he has lost his situational awareness. This measure is difficult to quantify, it is a mostly subjective impression of both the test subject and the experiment leader. Situational awareness can be expressed as the number of times a test subject uttered a command at a time that was not allowed.

skilled impostor
/ˈskɪld ɪmˈpɒstə/, **/’skIld Im’pQst@/**, [N: [AJ: skilled][N: impostor]], [plural: -s]. Domain: speaker recognition. Hyperonyms: impostor. Synonyms: wolf. Cohyponym: poor impostor. Def.: Impostor with a high success rate in claiming an identity averaged over each claimed identity. (Gibbon et al. 1997, p. 441)

SLP
/esel'piː/, /esel'piː/, [N: SLP], [plural: none]. Domain: processing. Hyperonyms: language processing. Hyponyms: automatic speech recognition, automatic speech synthesis, speaker recognition. Synonyms: speech processing. Cohyponym: natural language processing, text processing, word processing. Def.: Spoken language processing is an area of research and development in the field of human language technologies concerned with input and output systems which process natural human speech.

smoothing
/'smuːðɪŋ/, /'smu:DIN/, [N: smoothing], [plural: none]. Domain: language modelling. Hyperonyms: language modelling . Hyponyms: linear discounting, linear interpolation, absolute discounting. Meronym. sup.: stochastic language modelling. Def.: 1. Low pass filtering of signals. 2. Smoothing is a method that is needed in the context of stochastic language modelling to counteract the effect of sparse training data. The goal of smoothing is to guarantee that all probabilities are different from zero.

snake
/'sneɪk/, /'sneIk/, [N: snake], [plural: -s]. Domain: multimodal systems. Synonyms: active contour. Def.: Deformable contour defined by a set of nodes connected by springs. Snakes are first located on the face. Contours are tracked by applying an image force field that is computed from the gradient of the intensity image. Muscle contraction is estimated from contour deformations. The import of visual information to recognise audio signals is around 7 percent.

sociolect
/'səʊsɪəʊlekt/, /'s@UsI@Ulekt/, [N: sociolect], [plural: -s]. Hyperonyms: language variety. Meronym. sup.: natural language. Def.: Sociolect is a term used by sociolinguists to refer to a linguistic variety defined on social grounds. (Crystal 1988)

sociolinguistics
/'səʊsɪəʊlɪŋ'gwɪstɪks/, /'s@UsI@UlIN'gwIstIks/, [N: sociolinguistics], [plural: always plural]. Meronym. sup.: linguistics. Def.: A branch of linguistics which studies all aspects of the relationship between language and society. Sociolinguists study such matters as the linguistic identity of social groups, social attitudes to language, standard and non-standard forms of language, the patterns and needs of national language use, social varieties and levels of language, the social basis of multilingualism, and so on. (Crystal 1988, p. 282)

software-only recogniser
/'sɒftweər 'əʊnli 'rekəgnaɪzə/, /'sQftwe@r '@Unli 'rek@gnaIz@/, [N: [N: software][AJ: only][N: recogniser]], [plural: -s]. Def.: A speech recogniser capable of operating on a standard personal computer with multimedia sound input, without needing additional processing hardware.

Sound Exchange
/'saʊnd ɪks'tʃeɪndʒ/, /'saUnd Iks'tSeIndZ/, [N: [N: Sound][N: Exchange]], [plural: none]. Hyperonyms: software. Synonyms: SOX. Def.: A versatile tool for converting between various audio formats. It can read and write various types of audio files, and optionally applies some special effects (e.g. echo, channel averaging, or rate conversion).

soundproof booth
/ˈsaʊndpruːf ˈbuːð/, /'saUndpru:f 'bu:D/, [N: [AJ: soundproof][N: booth]], [plural: -s]. Domain: physical characterisation. Hyperonyms: recording room. Cohyponym: laboratory room, recording studio, anechoic chamber. Def.: A sound-insulated and acoustically treated booth or small chamber is often used in clinical audiometry or in psycho-acoustic experiments. The advantage of this kind of equipment is that it is comparably inexpensive and may easily be standardised. The kind of environment this equipment provides is, however, not recommended for high quality speech recordings for scientific purposes, since small rooms exhibit strong eigenmodes at relatively high frequencies which may lie well within the speech frequency region. Due to the small dimensions of the booth the acoustic treatment of the inner surface will generally not suffice to provide enough absorption for the resonances to disappear. As a consequence, speech recordings produced in this environment will exhibit strong linear distortions, i.e. sound colouration. (Gibbon et al. 1997, p. 309)

SOX
/ˈsɒks/, /'sQks/, [N: SOX], [plural: none]. Hyperonyms: software. Synonyms: Sound Exchange. Def.: A versatile tool for converting between various audio formats. It can read and write various types of audio files, and optionally applies some special effects (e.g. echo, channel averaging, or rate conversion). However, it does not necessarily perform optimal rate conversions.

speaker adaptive system
/ˈspiːkər əˈdæptɪv ˈsɪstəm/, /'spi:k@r @'d{ptIv 'sIst@m/, [N: [N: speaker][AJ: adaptive][N: system]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: speech recognition system. Cohyponym: speaker dependent system, speaker-dependent speech recognition system, speaker independent system, speaker-independent speech recognition system. Def.: A speaker adaptive system starts out as a speaker independent system, but gradually changes its speech models such that the system adapts to a specific user. Performance (after adaptations) is typically that of speaker dependent systems.

speaker alignment
/ˈspiːkər əˈlaɪnmənt/, /'spi:k@r @'laInm@nt/, [N: [N: speaker][N: alignment]], [plural: -s]. Domain: speaker recognition. Hyperonyms: evaluation method, recognition task. Cohyponym: speaker matching, speaker labelling, speaker change detection. Def.: The identity and order of speakers taking part in a conversation are known and the goal is to localise when each of their interventions begins and ends. (Gibbon et al. 1997, p. 411)

speaker change detection
/ˈspiːkə ˈtʃeɪndʒ dɪˈtekʃən/, /'spi:k@ 'tSeIndZ dI'tekS@n/, [N: [N: speaker][N: change][N: detection]], [plural: -s]. Domain: speaker recognition. Hyperonyms: evaluation method, recognition task. Cohyponym: speaker matching, speaker labelling, speaker alignment. Def.: The goal is to detect a change of speaker along a speech stream. (Gibbon et al. 1997, p. 411)

speaker characterisation
/ˈspiːkə kærəktəraɪˈzeɪʃən/, /'spi:k@ k{r@kt@raI'zeIS@n/, [N: [N: speaker][N: characterisation]], [plural: -s]. Def.: Each speaker has some associated properties, such as sex, age, dialect, profession, etc. Some control over these properties can be obtained by selecting the test speakers or specific material in the database.

speaker class identification
/ˈspiːkə ˈklɑːs aɪdentɪfɪˈkeɪʃən/, /'spi:k@ 'klA:s aIdentIfI'keIS@n/, [N: [N: speaker][N: class][N: identification]], [plural: -s]. Domain: speaker recognition. Hyperonyms: evaluation method; speaker recognition. Cohyponym: speaker class verification. Def.: Any decision-making process that uses some features of the speech signal to determine the class to which the speaker of a given utterance belongs.

speaker class verification
/'spiːkə 'klɑːs verɪfɪ'keɪʃən/ , **/'spi:k@ 'klA:s verIfI'keIS@n/** , [N: [N: speaker][N: class][N: verification]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: evaluation method; speaker recognition. Cohyponym: speaker class identification. Def.: Any decision-making process that uses some features of the speech signal to determine whether the speaker of a given utterance belongs to a given class.

speaker classification
/'spiːkə klæsɪfɪ'keɪʃən/ , **/'spi:k@ kl{sIfI'keIS@n/** , [N: [N: speaker][N: classification]], [plural: -s]. Domain: speaker recognition. Hyperonyms: evaluation method; speaker recognition. Cohyponym: speaker identification, speaker verification. Def.: Any decision-making process that uses some features of the speech signal to determine some characteristics of the speaker of a given utterance.

speaker cluster selection
/'spiːkə 'klʌstə sɪ'lekʃən/, **/'spi:k@ 'klVst@ sI'lekS@n/**, [N: [N: speaker][N: cluster][N: selection]], Domain: speaker recognition: speaker classification. Hyperonyms: speaker classification task. Cohyponym: sex identification, age identification, mood identification, health state identification, accent identification. Def.: The task of classifying a speaker with respect to one of several categories, the characteristics of which cannot be expressed in objective terms, for instance, some speech recognition systems use models of speech units that have variants across several speaker clusters. These clusters may be obtained in an unsupervised manner, and it is usually impossible to find a posteriori an objective attribute that would qualify each cluster. (Gibbon et al. 1997, p. 409)

speaker dependent system
/'spiːkə dɪ'pendənt 'sɪstəm/, **/'spi:k@ dI'pend@nt 'sIst@m/**, [N: [N: speaker][AJ: dependent][N: system]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: speech recognition system. Cohyponym: speaker adaptive system, speaker independent system. Def.: A speaker dependent system needs training for the specific user who is going to use the system.

speaker identification task
/'spiːkər aɪdentɪfɪ'keɪʃən 'tɑːsk/, **/'spi:k@r aIdentIfI'keIS@n 'tA:sk/**, [N: [N: speaker][N: identification][N: task]], [plural: -s]. Domain: speaker recognition. Hyperonyms: recognition task. Cohyponym: speaker verification task. Def.: The goal of a speaker identification task is to classify an unlabelled voice token as belonging to one of a set of n reference speakers. (Gibbon et al. 1997, p. 411)

speaker identification
/'spiːkər aɪdentɪfɪ'keɪʃən/ , **/'spi:k@r aIdentIfI'keIS@n/** , [N: [N: speaker][N: identification]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: evaluation method; speaker recognition. Cohyponym: speaker verification, speaker classification. Def.: a) Any decision-making process that uses some features of the speech signal to determine who the speaker of a given utterance is. b) Task that consists in identifying an unknown speaker as one of a closed set of possible speakers. The typical implementation is carried out by comparing the test utterance with recordings of all known speakers, and choosing the speaker that fits best.

speaker independent system
/'spiːkərɪndɪ'pendənt 'sɪstəm/, **/'spi:k@rIndI'pend@nt 'sIst@m/**, [N: [N: speaker][AJ: independent][N: system]], [plural: -s]. Domain: speech recognition. Hyperonyms: speech recognition system. Cohyponym: speaker adaptive system, speaker dependent system. Def.: A system not trained for a specific user. A speaker independent system is trained in the factory, and can hence be used directly after unpacking. The recognition performance is generally lower than a comparable speaker dependent system.

speaker labelling
/'spiːkə 'leɪbəlɪŋ/, **/'spi:k@ 'leIb@lIN/**, [N: [N: speaker][N: labelling]], [plural: -s]. Domain: speaker recognition. Hyperonyms: evaluation method, recognition task. Cohyponym: speaker matching, speaker alignment, speaker change detection. Def.: When the identity of the speakers taking part in a conversation is known, the goal is to localise when their successive interventions begin and end, including a possible outcome of 'none of the registered speakers', in case of open-set labelling. (Gibbon et al. 1997, p. 411)

speaker matching
/'spiːkə 'mætʃɪŋ/, **/'spi:k@ 'm{tSIN/**, [N: [N: speaker][N: matching]], [plural: -s]. Domain: speaker recognition. Hyperonyms: evaluation method, recognition task. Cohyponym: speaker labelling, speaker alignment, speaker change detection. Def.: The task of choosing a speaker in a closed-set of references which is most similar to a current speaker, even though it is known in advance that the applicant speaker is not a registered speaker. (Gibbon et al. 1997, p. 411)

speaker recognition
/'spiːkə rekəg'nɪʃən/, **/'spi:k@ rek@g'nIS@n/**, [N: [N: speaker][N: recognition]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: evaluation method. Hyponyms: speaker verification, speaker identification, speaker classification; speaker class identification, speaker class verification; spoken language identification, spoken language verification.. Def.: Any decision-making process that uses some features of the speech signal to determine some information on the identity of the speaker of a given utterance.

speaker verification task
/'spiːkə verɪfɪ'keɪʃən 'tɑːsk/, **/'spi:k@ verIfI'keIS@n 'tA:sk/**, [N: [N: speaker][N: verification][N: task]], [plural: -s]. Domain: speaker recognition. Hyperonyms: recognition task. Cohyponym: speaker identification task. Def.: The speaker verification task is to decide whether or not the unlabelled voice belongs to a specific reference speaker. (Gibbon et al. 1997, p. 411)

speaker verification
/'spiːkə verɪfɪ'keɪʃən/ , **/'spi:k@ verIfI'keIS@n/** , [N: [N: speaker][N: verification]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: evaluation method; speaker recognition. Synonyms: speaker authentication. Cohyponym: speaker identification, speaker classification. Def.: Any decision-making process that uses some features of the speech signal to determine whether the speaker of a given utterance is a particular person, whose identity is specified.

specialisation
/speʃəlaɪ'zeɪʃən/, **/speS@laI'zeIS@n/**, [N: specialisation], [plural: none]. Domain: multimodal systems. Hyperonyms: cooperation type. Cohyponym: complementarity, redundancy, equivalence, concurrency, transfer. Def.: A specific chunk of information is always transmitted using the same modality. Specialisation may also manifest itself in user preferences, for example, if users consistently prefer speech over other input modalities for certain tasks. E.g. An information kiosk offers different services which are selected by touching the corresponding button..

specific lexicon theory
/spə'sɪfɪk 'leksɪkən 'θiːəri/, **/sp@'sIfIk 'leksIk@n 'Ti:@ri/**, [N: [AJ: specific][N: lexicon][N: theory]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: lexicon theory. Cohyponym: general lexicon theory. Def.: A specific lexicon formulated in a lexicon formalism on the basis of a lexicon model.

specification and description language
/spesɪfɪ'keɪʃən 'ænd dɪ'skrɪpʃən 'læŋgwɪdʒ/, `/spesIfI'keIS@n '{nd dI'skrIpS@n 'l{NgwIdZ/`, [N: [N: Specification][C: and][N: Description][N: Language]], [plural: -s]. Domain: interactive dialogue systems. Synonyms: SDL. Def.: A graphical language for describing state transition diagrams for event-driven systems. It was standardised by CCITT. (Gibbon et al. 1997, p. 573)

speech act
/'spiːtʃ 'ækt/, `/'spi:tS '{kt/`, [N: [N: speech][N: act]], [plural: -s]. Hyperonyms: action. Def.: A speech act is the informational action that a speaker effects by producing an utterance. For example, asking a question, offering information, and making a promise are three different types of speech act. The basic idea of speech acts is vitally important in work on dialogue systems. Speech acts serve as the base level of categorisation for dialogue work (in much the way that word classes have that function at the lexical level). So, for example, dialogue grammars can be written which describe well-formed sequences of speech acts. Many researchers working on interactive dialogue systems wish to use the notion of speech act without enlisting the whole philosophical apparatus of Speech Act Theory; for this purpose the term dialogue act has been coined and is steadily growing in acceptability.

speech aware consumer electronics
/'spiːtʃ ə'weə kən'sjuːmərɪlek'trɒnɪks/, `/'spi:tS @'we@ k@n'sju:m@rIlek'trQnIks/`, [N: [N: speech][AJ: aware][N: consumer][N: electronics]], [plural: none]. Domain: consumer off-the-shelf products. Def.: Mobile telephone, video recorders, TVs, car radios, etc. that can be controlled by speech input. At the time of writing, most of these systems are under development, and only a few are already available.

speech database
/'spiːtʃ 'deɪtəbeɪs/, `/'spi:tS 'deIt@beIs/`, [N: [N: speech][N: database]], [plural: -s]. Hyperonyms: database. Def.: A systematic database containing a collection of speech signals with header information, transcriptions and signal annotations.

speech output assessment
/'spiːtʃ 'aʊtpʊt ə'sesmənt/ , `/'spi:tS 'aUtpUt @'sesm@nt/` , [N: [N: speech][N: output][N: assessment]], [plural: -s] . Domain: speech synthesis. Hyperonyms: evaluation method. Synonyms: speech output testing. Def.: Determination of the quality of (some aspect(s) of) a speech output system.

speech output evaluation
/'spitʃ 'aʊtpʊt ɪvælju'eɪʃən/, `/'spitS 'aUtpUt Iv{lju'eIS@n/`, [N: [N: speech][N: output][N: evaluation]] , [plural: -s] . Domain: speech synthesis. Hyperonyms: evaluation method. Synonyms: speech output testing. Def.: Determination of the quality of (some aspect(s) of) a speech output system.

speech output system
/'spiːtʃ 'aʊtpʊt 'sɪstəm/, `/'spi:tS 'aUtpUt 'sIst@m/`, [N: [N: speech][N: output][N: system]], [plural: -s]. Domain: speech synthesis. Hyponyms: text-to-speech system, concept-to-speech system. Def.: A device, either a dedicated machine or a computer programme, that produces signals that are intended to be functionally equivalent to speech produced by humans. In the present state of affairs speech output systems generally produce audio signals only, but laboratory systems are being developed that supplement the audio signal with the visual image of the (artificial) talker's face.

speech output testing
/'spiːtʃ 'aʊtpʊt 'testɪŋ/ , `/'spi:tS 'aUtpUt 'testIN/` , [N: [N: speech][N: output][N: testing]] , [plural: -s] . Domain: speech synthesis. Hyperonyms: testing procedure. Def.: Determination of the quality of (some aspect(s) of) a speech output system.

speech pathology
/'spiːtʃ pə'θɒlədʒi/, `/'spi:tS p@'TQl@dZi/`, [N: [N: speech][N: pathology]], [plural: y/-ies]. Domain: corpora, speaker recognition. Def.: Speech pathology is the study of the various types of pathological speech. (Gibbon et al. 1997, p. 92) E.g. hoarseness, aphasia.

speech recogniser training
/'spiːtʃ 'rekəgnaɪzə 'treɪnɪŋ/ , `/'spi:tS 'rek@gnaIz@ 'treInIN/` , [N: [N: speech][N: recogniser][N: training]], [plural: -s]. Domain: speech recognition. Def.: Data for speech recogniser training is generated by means of collecting and analysing samples of real spontaneous speech. (Gibbon et al. 1997, p. 578)

speech recognition system
/'spiːtʃ rekəg'nɪʃən 'sɪstəm/, `/'spi:tS rek@g'nIS@n 'sIst@m/`, [N: [N: speech][N: recognition][N: system]], [plural: -s]. Domain: speech recognition. Hyponyms: discrete speech recognition system, continuous speech recognition system. Synonyms: speech recogniser. Cohyponym: speech understanding system, speech synthesis system. Def.: System that automatically recognises speech. Speech recognition systems support a restricted finite vocabulary, bounded by the limitations of the current technology. (Gibbon et al. 1997, p. 579)

speech synthesis
/'spiːtʃ 'sɪnθəsɪs/, `/'spi:tS 'sInT@sIs/`, [N: [N: speech][N: synthesis]], [plural: speech syntheses]. Domain: speech synthesis. Synonyms: production of speech sounds. Cohyponym: speech recognition, speech understanding. Def.: Speech synthesis is the name given to the production of speech sounds by a machine. Most speech synthesisers take a text string as input and produce a spoken version of the text as output. Some systems allow the text string to be annotated with prosodic markers which result in changes to the intonational pattern of the speech produced. (Gibbon et al. 1997, p. 92)

speech technology
/'spiːtʃ tek'nɒlədʒi/, `/'spi:tS tek'nQl@dZi/`, [N: [N: speech][N: technology]], [plural: none]. Hyperonyms: technology, human language technology. Hyponyms: speech input technology, speech output technology. Synonyms: spoken language technology, spoken language processing, SLP. Cohyponym: text technology, natural language processing, NLP. Def.: The discipline concerned with the research and development of spoken language input and output systems, using contributions from the neighbouring disciplines of acoustics, electrical engineering, statistics, phonetics, natural language processing, and involving system requirements specification, design, implementation and evaluation, corpus and linguistic resource processing, and consumer oriented product evaluation. (Gibbon et al. 1997, p. 578)

speech understanding system
/'spiːtʃ ʌndə'stændɪŋ 'sɪstəm/, `/'spi:tS Vnd@'st{ndIN 'sIst@m/`, [N: [N: speech][N: understanding][N: system]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Cohyponym: speech recognition system, speech recogniser. Def.: A system that not only recognises speech, but also interprets the words. One could call these systems 'speech-to-concept' systems as opposed to 'concept-to-speech' systems. In dialogue systems that ask open questions ("What do you want?") speech understanding plays an important role. If the questions are closed ("Do you want information about trains?") or specific ("Where do you want to go by train?") the system relies on speech recognition to a greater extent or word spotting alone. Generally, however, any speech interactive system that reacts to spoken input sensibly could be called a speech understanding system.

speech-to-speech translation
/'spiːtʃ tə 'spiːtʃ træns'leɪʃən/ , `/'spi:tS t@ 'spi:tS tr{ns'leIS@n/` , [N: [N: speech][PREP: to][N: speech][N: translation]], [plural: -s]. Def.: Translation from spoken utterances in one language directly to spoken utterances in another without intervening textual representations.

spelling alternation
/ˈspelɪŋ ɒltəˈneɪʃən/, /'spelIN Qlt@'neIS@n/, [N: [N: spelling][N: alternation]], [plural: -s]. Domain: lexicon. Synonyms: orthographic alternation. Def.: The differences between spellings of parts of composite words and the spellings of corresponding parts of simplex words. (Gibbon et al. 1997, p. 216) E.g. 'y' - 'i' - 'ie' in English 'fly', 'flier', 'flies'.

spelling rule
/ˈspelɪŋ ˈruːl/, /'spelIN 'ru:l/, [N: [N: spelling][N: rule]], [plural: -s]. Domain: lexicon. Hyperonyms: rule. Def.: Rule which describes spelling alternations. (Gibbon et al. 1997, p. 216)

spoken language corpus
/ˈspəʊkən ˈlæŋgwɪdʒ ˈkɔːpəs/, /'sp@Uk@n 'l{NgwIdZ 'kO:p@s/, [N: [AJ: spoken][N: language][N: corpus]], [plural: spoken language corpora]. Domain: corpora. Hyperonyms: corpus. Synonyms: collection of speech sound recordings. Cohyponym: written language corpus. Def.: Any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use.

spoken language dialogue system
/ˈspəʊkən ˈlæŋgwɪdʒ ˈdaɪəlɒg ˈsɪstəm/ , /'sp@Uk@n 'l{NgwIdZ 'daI@lQg 'sIst@m/ , [N: [AJ: spoken][N: language][N: dialogue][N: system]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: interactive dialogue system. Cohyponym: question/answer system. Def.: A variety of interactive dialogue system in which the primary mode of communication is spoken natural language. Spoken language dialogue systems take human-human conversation as their inspiration, though differences are bound to persist into the forseeable future by virtue of the character of such systems as constrained designed artifact. Spoken language dialogue systems support a much more natural kind of dialogue than Interactive Voice Response systems.

spoken language dialogue
/ˈspəʊkən ˈlæŋgwɪdʒ ˈdaɪəlɒg/, /'sp@Uk@n 'l{NgwIdZ 'daI@lQg/, [N: [AJ: spoken][N: language][N: dialogue]], [plural: -s] . Domain: interactive dialogue systems. Hyperonyms: dialogue. Synonyms: oral dialogue. Def.: A complete spoken verbal interaction between two parties (a system and a human being), each of whom is capable of independent actions. A dialogue is composed of a sequence of steps which are, in some way, related and build on each other. Dialogue systems are thus more sophisticated than question/answer systems, in which one agent may pose a succession of unrelated queries to the other agent.

spoken language identification
/ˈspəʊkən ˈlæŋgwɪdʒ aɪdentɪfɪˈkeɪʃən/ , /'sp@Uk@n 'l{NgwIdZ aIdentIfI'keIS@n/ , [N: [AJ: spoken][N: language][N: identification]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: decision-making process. Cohyponym: spoken language verification. Def.: Any decision-making process that uses some features of the speech signal to determine what language is spoken in a given utterance.

spoken language lexicon
/ˈspəʊkən ˈlæŋgwɪdʒ ˈleksɪkən/, /'sp@Uk@n 'l{NgwIdZ 'leksIk@n/, [N: [AJ: spoken][N: language][N: lexicon]], [plural: spoken language lexica, -s]. Domain: lexicon. Hyponyms: system lexicon, lexical database. Cohyponym: written language lexicon. Def.: 1. A spoken language lexicon may be a component in a system, a system lexicon, or a background resource for wider use, a lexical database, in each case containing information about the pronunciation, the spelling, the syntactic usage, the meaning and specific pragmatic properties of words; lexica containing subsets of this information may also be referred to as spoken language lexica, though the simpler cases are often simply referred to as wordlists. (Gibbon et al. 1997, p. 184) 2. A spoken language lexicon is defined as a list of representations of lexical entries consisting of spoken word forms paired with their other lexical properties such as spelling, pronunciation, part of speech (POS), meaning and usage information, in such a way as to optimise lookup of any or all of these properties. (Gibbon et al. 1997, p. 184)

Spoken Language Processing
/'spəʊkən 'læŋgwɪdʒ 'prəʊsesɪŋ/, /'sp@Uk@n 'l{NgwidZ 'pr@UsesIN/, [N: [AJ: spoken][N: language][N: processing]], [plural: none]. Domain: processing. Hyperonyms: language processing. Hyponyms: automatic speech recognition, automatic speech synthesis, speaker recognition. Synonyms: SLP, speech processing. Cohyponym: natural language processing, text processing, word processing. Def.: Spoken language processing is an area of research and development in the field of human language technologies concerned with input and output systems which process natural human speech.

spoken language technology
/'spəʊkən 'læŋgwɪdʒ tek'nɒlədʒi/, /'sp@Uk@n 'l{NgwIdZ tek'nQl@dZi/, [N: [AJ: spoken][N: language][N: technology]], [plural: y/-ies]. Hyperonyms: technology, human language technology. Hyponyms: speech input technology, speech output technology. Synonyms: spoken language technology, spoken language processing, SLP. Cohyponym: text technology, natural language processing, NLP. Def.: The discipline concerned with the research and development of spoken language input and output systems, using contributions from the neighbouring disciplines of acoustics, electrical engineering, statistics, phonetics, natural language processing, and involving system requirements specification, design, implementation and evaluation, corpus and linguistic resource processing, and consumer oriented product evaluation. (Gibbon et al. 1997, p. 578)

spoken language text corpus
/'spəʊkən 'læŋgwɪdʒ 'tekst 'kɔːpəs/, /'sp@Uk@n 'l{NgwIdZ 'tekst 'kO:p@s/, [N: [AJ: spoken][N: language][N: text][N: corpus]], [plural: spoken language text corpora]. Domain: corpora. Hyperonyms: corpus. Def.: A spoken language text corpus is a collection of data not taken from existing texts but from speech data that are written down in some orthographic or non-orthographic form in order to become part of a data collection. (Gibbon et al. 1997, p. 81)

spoken language verification
/'spəʊkən 'læŋgwɪdʒ verɪfɪ'keɪʃən/ , /'sp@Uk@n 'l{NgwIdZ verIfI'keIS@n/ , [N: [AJ: spoken][N: language][N: verification]], [plural: -s]. Domain: speaker recognition. Hyperonyms: evaluation method; decision-making process. Cohyponym: spoken language identification. Def.: Any decision-making process that uses some features of the speech signal to determine whether the language spoken in a given utterance is a particular language.

spontaneous speech
/spɒn'teɪnɪəs 'spiːtʃ/, /spQn'teInI@s 'spi:tS/, [N: [AJ: spontaneous][N: speech]], [plural: none]. Domain: speech recognition, consumer off-the-shelf products. Hyperonyms: speaking style. Cohyponym: read speech, dictation speech. Def.: This is the most representative speaking style. In everyday life people communicate by talking spontaneously to each other. For a recognition system this talking style is very difficult to recognise. The style is characterised by level variations and the use of unpredictable intonation, hesitations, corrections, and incomplete sentences that are often grammatically incorrect.

SQL
/'eskjuː'el/, /'eskju:'el/, [N: SQL], [plural: none]. Hyperonyms: query language. Synonyms: Standard Query Language. Cohyponym: OSQL. Def.: SQL is the de facto standard language for relational databases, and SQL-3 is currently being standardised by the ISO; important new features are the computation of transitive closure, and object-oriented concepts.

Standard Generalized Markup Language
/'stændəd 'dʒenərəlaɪzd 'mɑːkʌp 'læŋgwɪdʒ/, /'st{nd@d 'dZen@r@laIzd 'mA:kVp 'l{NgwIdZ/, [N: [AJ: Standard][AJ: Generalized][N: Markup][N: Language]], [plural: none]. Hyperonyms: formal language. Hyponyms: eXtended Markup Language (XML). Synonyms: SGML. Def.: An ISO standard for markup (annotation) which describes the structure of a text.

standard procedure
/'stændəd/, /'st{nd@d/, [N: standard], [plural: -s]. Domain: . Cohyponym: ad hoc procedure. Def.: 1. A laboratory procedure conforming to agreed professional best practice. 2. A procedure defined by a standardisation organisation such as ISO, DIN, BSI.

Standard Query Language
/'stændəd 'kwi:əri 'læŋgwɪdʒ/, /'st{nd@d 'kwi:@ri 'l{NgwIdZ/, [N: [AJ: Standard][N: Query][N: Language]], [plural: none]. Hyperonyms: query language. Synonyms: SQL. Cohyponym: OSQL. Def.: The de facto standard language for relational databases, and SQL-3 is currently being standardised by the ISO; important new features are the computation of the transitive closure, and object-oriented concepts.

static-dynamic representation
/'stætɪk 'daɪ'næmɪk reprɪzen'teɪʃən/, /'st{tIk 'daI'n{mIk reprIzen'teIS@n/, [N: [AJ: static][AJ: dynamic][N: representation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: output modality representation. Cohyponym: linguistic representation, analogue representation, iconic representation, arbitrary representation. Def.: The representation is considered static when it can be perceived in an identical form for a certain time. If the representation changes continuously over time, it is called a dynamic representation. A blinking icon is considered static while a movie or music will be characterised as dynamic.

stem lexicon
/'stem 'leksɪkən/, /'stem 'leksIk@n/, [N: [N: stem][N: lexicon]], [plural: stem lexica, -s]. Domain: lexicon. Hyperonyms: lexicon. Cohyponym: morph lexicon, morpheme lexicon, fully inflected form lexicon. Def.: A lexicon in which the basic lexical key or lemma is the stem, which is represented in some kind of normalised notation (e.g. 'infinitive' for verbs, 'nominative singular' for nouns, in a standardised orthographic representation). (Gibbon et al. 1997, p. 199)

stem
/'stem/, /'stem/, [N: stem], [plural: -s]. Domain: lexicon. Cohyponym: affix. Def.: In the most general usage, a stem is any uninflected item, whether morphologically simple or complex. However, intermediate stages in word formation by affixation, and in the inflection of highly inflected languages, are also called stems. The smallest stem is a phonological lexical morph or an orthographic lexical morph, i.e. the phonological or orthographic realisation of a lexical morpheme. Stems may vary in different inflectional contexts. (Gibbon et al. 1997, p. 199)

stochastic grammar
/stə'kæstɪk 'græmə/ , /st@'k{stIk 'gr{m@/ , [N: [AJ: stochastic][N: grammar]], [plural: -s] . Domain: language modelling. Def.: A stochastic grammar is a stochastic language model that is based on a (context free) grammar; the grammar rules are assigned probabilities such that each word string generated by the grammar has a non-zero probability.

stochastic language model
/stə'kæstɪk 'læŋgwɪdʒ 'mɒdəl/, /st@'k{stIk 'l{NgwIdZ 'mQd@l/, [N: [AJ: stochastic][N: language][N: model]] , [plural: -s]. Domain: language modelling. Hyperonyms: language model. Cohyponym: grammar based language model. Def.: A stochastic language model is a language model that assigns probabilities to the allowed word sequences; typically all word sequences have a non-zero probability.

stochastic speech recognition system
/stə'kæstɪk 'spi:tʃ rekəg'nɪʃən 'sɪstəm/, /st@'k{stIk 'spi:tS rek@g'nIS@n 'sIst@m/, [N: [AJ: stochastic][N: speech][N: recognition][N: system]], [plural: -s]. Domain: language modelling, speech recognition. Hyperonyms: speech recognition system. Def.: A stochastic speech recognition system relies on stochastic models which are estimated or trained with (very) large amounts of speech, using some statistical optimalisation procedure. (Gibbon et al. 1997, p. 94) E.g. Hidden Markov Model (HMM), neural network.

stop
/'stɒp/, /'stQp/, [N: stop], [plural: -s]. Hyperonyms: consonant. Def.: Any sound which is produced by a complete closure in the vocal tract, and thus traditionally includes the class of plosives. Both nasal and oral sounds can be classified as stops, though the term is usually reserved for the latter. (Crystal 1988, p. 287)

stroke
/'strəʊk/, /'str@Uk/, [N: stroke], [plural: -s]. Domain: multimodal systems. Hyperonyms: unit. Def.: a) Basic unit of gesture input: trajectory from one touch of the pen/finger on the display to the next lift of pen/finger off the display. b) Concerning 3D gestures: the apex part of the gesture.

structural model
/'strʌktʃərəl 'mɒdəl/, /'strVktS@r@l 'mQd@l/, [N: [AJ: structural][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: synthetic model, physically-based model. Def.: The face is structured as a hierarchy of regions (forehead, brow, cheek, node, lip) and subregions (upper lip, lower lip, left lip corner, right lip corner). Each region corresponds to one muscle or a group of related muscles. These regions can, under the action of a muscle, either contract or be affected by the propagation of movement from adjacent regions. A region is defined by a special point (the point of insertion of the muscle), and its connection information (to which regions it is connected). Connection information is necessary for computing the movement propagation. The muscle is defined by three or five segments that follow the bone structure of the face.

structural property
/'strʌktʃərəl 'prɒpəti/, /'strVktS@r@l 'prQp@ti/, [N: [AJ: structural][N: property]], [plural: y/-ies]. Domain: lexicon. Cohyponym: interpretative property. Def.: Structural (or 'syntactic', in a general sense of the term) properties of a lexical sign are distributional properties (syntactic category and subcategory) and compositional properties (head and modifier constituents (complement or specifier)). (Gibbon et al. 1997, p. 194)

stuttering
/'stʌtərɪŋ/, /'stVt@rIN/, [N: stuttering], [plural: none]. Domain: corpora. Hyperonyms: rhythm disorder. Synonyms: stammering. Cohyponym: cluttering. Def.: Stuttering is a very complex phenomenon that is characterised by, for instance, a repetition of speech segments, abnormal prolongations of sound segments, words being unfinished, or circumlocutions to avoid types of sound that cause problems. Stuttering varies enormously from person to person and from situation to situation. It is, for instance, well known that stutterers almost never stutter when they are singing. Both organic (genetic) causes and functional (environmental) causes are assumed to underlie the stuttering phenomenon. (Gibbon et al. 1997, p. 115)

subject production variable
/'sʌbdʒekt prə'dʌkʃən 'veərɪəbəl/ , /'sVbdZekt pr@'dVkS@n 've@rI@b@l/ , [N: [N: subject][N: production][N: variable]], [plural: -s]. Domain: interactive dialogue systems. Def.: Subject production variables relate to the speech and language produced by the subject insofar as they have implications for the ability of the wizard to recognise and understand the subject's word. (Gibbon et al. 1997, p. 584) E.g. accent, voice quality, dialect, verbosity, politeness.

sublanguage
/'sʌblæŋgwɪdʒ/, /'sVbl{NgwIdZ/, [N: sublanguage], [plural: -s]. Domain: lexicon. Hyperonyms: language variety. Meronym. sup.: natural language. Def.: The subpart of some natural language which is deemed to be relevant to some given task and/or application domain. Interactive dialogue systems are not currently capable of modelling an average speaker's entire linguistic competence, so the normal approach is to identify and model only the sublanguage which is relevant to the function or functions which the interactive dialogue system is intended to perform. The idea of sublanguage is related to, but distinct from the linguistic notion of register. A sublanguage in the context of interactive dialogue systems should not be confused with a sublanguage in the mathematical sense. In the latter case, the language of which the sublanguage is a part is formally well-defined; in the former case it is not.

substitution
/sʌbstɪˈtjuːʃən/, /sVbstI'tju:S@n/, [N: substitution], [plural: -s]. Domain: language modelling, speech recognition, system design, corpora. Hyperonyms: identity assignment. Synonyms: misclassification. Cohyponym: deletion, insertion. Def.: A response, for instance of a speech recogniser, that is different from the response which matches the input.

suffix
/ˈsʌfɪks/, /'sVfIks/, [N: suffix], [plural: -es]. Domain: lexicon. Hyperonyms: affix. Cohyponym: prefix, circumfix. Meronym. sup.: word. Def.: A suffix is an affix attached to the end of a stem; it is a grammatical morpheme used in morphological inflection or derivation. E.g. stem 'cut' + suffix 's' = 'cuts'.

supercardioid microphone
/ˈsuːpəkɑːdiɔɪd ˈmaɪkrəfəʊn/, /'su:p@kA:diOId 'maIkr@f@Un/, [N: [AJ: supercardioid][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: unidirectional microphone. Cohyponym: cardioid microphone, hypercardioid microphone. Def.: Supercardioid microphones are least sensitive at 125 degrees off-axis, 8.7 db down at the sides and approximately 15 db down at the rear. (Gibbon et al. 1997, p. 304)

superfix
/ˈsuːpəfɪks/, /'su:p@fIks/, [N: superfix], [plural: -es]. Domain: lexicon. Def.: Prosodic realisation of an inflectional or derivational morpheme.

syllabic orthography
/sɪˈlæbɪk ɔːˈθɒgrəfi/, /sI'l{bIk O:'TQgr@fi/, [N: [AJ: syllabic][N: orthography]], [plural: y/-ies]. Domain: lexicon. Hyperonyms: orthography. Cohyponym: logographic orthography, alphabetic orthography. Def.: In syllabic orthography characters are closely related to phonological syllables. (Gibbon et al. 1997, p. 188) E.g. Japanese 'Kana'.

syllable monitoring
/ˈsɪləbəl ˈmɒnɪtərɪŋ/, /'sIl@b@l 'mQnIt@rIN/, [N: [N: syllable][N: monitoring]], [plural: none]. Domain: speech synthesis. Hyperonyms: monitoring. Cohyponym: phoneme monitoring, word monitoring. Def.: Testing the intelligibility of combinations of sounds. (Gibbon et al. 1997, p.490)

syllable
/ˈsɪləbəl/, /'sIl@b@l/, [N: syllable], [plural: -s]. Domain: speech synthesis, corpora, lexicon. Hyperonyms: unit of speech. Hyponyms: stressed syllable, unstressed syllable. Meronym. sup.: word. Def.: A vowel optionally preceded and/or followed by one or more consonants, e.g. V, CV, CVC, VC, CCV, etc.; phonological unit used for describing the structure of words form the point of view of their pronunciation, without direct reference to meaning. (Gibbon et al. 1997, p. 212) The principle underlying syllable structure is the sonority hierarchy: vowels are the most sonorous sounds, and the consonant sequence from left margin or right margin of a syllable to the vowel proceeds from least (e.g. voiceless stops) to most (e.g. liquids) sonorous. Syllable structure is an important criterion in language typology for differences between languages, with a hierarchy of preferences: all languages have CV and V structures, fewer have CVC, still fewer have CCVC and so on. Since there is a small finite upper bound on the length of syllables, and a finite vocabulary of sounds at each syllable position, there is a finite set of syllables in any given language. A distinction must be made between the actual (i.e. lexically attested) syllables of a language and the potential syllables which can be constructed in principle by combining consonants and vowels according to the phonotactic principles of a language. The potential syllables form one of the potential sources of new words in a language.

symbolic gesture
/sɪmˈbɒlɪk ˈdʒestʃə/, /sIm'bQlIk 'dZestS@/, [N: [AJ: symbolic][N: gesture]], [plural: -s]. Domain: Spoken Language Technolgy: multimodal systems. Hyperonyms: gesture. Cohyponym: deictic gesture, iconic gesture, metaphoric gesture. Def.: Symbolic gestures can be translated directly to some meaning. E.g. thumb-up gesture to indicate agreement.

synonym
/'sɪnənɪm/, **/'sIn@nIm/**, [N: synonym], [plural: -s]. Domain: lexicon. Hyperonyms: word, lexical item. Hyponyms: full synonym, partial synonym. Cohyponym: antonym. Def.: Two words are synonyms if and only if they have the same meaning (or at least have one meaning in common), i.e. if the meaning of each entails the meaning of the other. (Gibbon et al. 1997, p. 850) Full synonyms are hard to find, except in very restricted domains; partial synonymy in which two words share at least one meaning is more common. Abbreviations and their full versions are perhaps the 'purest' synonyms.

synonymy
/sɪ'nɒnəmi/, **/sI'nQn@mi/**, [N: synonymy], [plural: none]. Domain: lexicon. Hyperonyms: semantic relation. Hyponyms: partial synonymy, full synonymy. Cohyponym: antonymy. Def.: The semantic relation that holds between two words that have the same meaning (or at least have one meaning in common), i.e. if the meaning of each entails the meaning of the other.

syntactic word
/sɪn'tæktɪk 'wɜːd/, **/sIn't{ktIk 'w3:d/**, [N: [AJ: syntactic][N: word]], [plural: -s]. Domain: lexicon. Hyperonyms: word. Cohyponym: orthographic word, phonological word, morphological word, prosodic word. Def.: Word based on its distribution in sentences. (Gibbon et al. 1997, p. 197)

syntax
/'sɪntæks/, **/'sInt{ks/**, [N: syntax], [plural: none]. Hyponyms: dialogue syntax, phrasal syntax, sentence syntax. Cohyponym: morphology, phonology, phonetics, semantics, pragmatics. Meronym. sup.: linguistics. Def.: 1. The study of the rules governing the way words are combined to form sentences in a language. 2. The study of the interrelationships between elements of sentence structure, and of the rules governing the arrangement of sentences in sequences. (Crystal 1988, p. 300)

synthesis by rule
/'sɪnθəsɪs 'baɪ 'ruːl/, **/'sInT@sIs 'baI 'ru:l/**, [N: [N: synthesis][PREP: by][N: rule]], [plural: none]. Domain: speech synthesis. Hyperonyms: speech generation, speech synthesis. Def.: Synthesis by rule is a method of generating computerised speech. (Gibbon et al. 1997, p. 93)

synthetic agent
/sɪn'θetɪk 'eɪdʒənt/, **/sIn'TetIk 'eIdZ@nt/**, [N: [AJ: synthetic][N: agent]], [plural: -s]. Domain: multimodal systems. Cohyponym: talking head, talking face. Def.: A whole synthetic persona including the whole body.

system capability profile
/'sɪstəm keɪpə'bɪlɪti 'prəʊfaɪl/, **/'sIst@m keIp@'bIlIti 'pr@UfaIl/**, [N: [N: system][N: capability][N: profile]], [plural: -s]. Domain: system design. Cohyponym: application requirement profile. Def.: The system capability profile indicates the available technology through commercial products as well as through pre-industrial laboratory prototypes (the last stage of the prototyping process). It exhibits what can be done. (Gibbon et al. 1997, p. 32)

system correction rate
/'sɪstəm kə'rekʃən 'reɪt/, **/'sIst@m k@'rekS@n 'reIt/**, [N: [N: system][N: correction][N: rate]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: ratio. Def.: Percentage of all system turns which are correction turns.

system-driven dialogue
/'sɪstəm 'drɪvən 'daɪəlɒg/, **/'sIst@m 'drIv@n 'daI@lQg/**, [N: [N: system][AJ: driven][N: dialogue]], [plural: -s]. Domain: interactive dialogue systems. Def.: A type of human-machine dialogue control in which the system always determines which information items can be input in response to the system prompt. System driven dialogues can be menus (i.e. an interaction in which the legal selections are explicitly presented by the system) or selections from implicit lists (if the number of options is too large to allow explicit presentation).

system-in-the-loop method
/ˈsɪstəm ˈɪn ðə ˈluːp ˈmeθəd//, /'sIst@m 'In D@ 'lu:p 'meT@d//, [N: [N: system][PREP: in][DET: the][N: loop][N: method]], [plural: none]. Domain: interactive dialogue systems. Hyperonyms: collection method. Def.: A speech data collection method which involves getting subjects to use an existing spoken language dialogue system, and recording what they say. According to this method, which is used for the purpose of collecting speech data for training and testing recognisers, users interact with an existing dialogue system while the data generated is collected. (Gibbon et al. 1997, p. 581)

Tadoma method
/təˈdəʊmə ˈmeθəd/, /t@'d@Um@ 'meT@d/, [N: [N: Tadoma][N: method]], [plural: -s]. Domain: multimodal systems. Hyperonyms: speech recognition method. Def.: Tactile perception of speech by using the finger tips to sense the vibrations of the throat and face and jaw positions of the speaker. Tadoma is used by deaf-blind people.

tagging scheme
/ˈtægɪŋ ˈskiːm/, /'t{gIN 'ski:m/, [N: [N: tagging][N: scheme]], [plural: -s]. Domain: dialogue representation. Def.: A list of annotation tags together with their definitions and the guidelines needed to map them on to a corpus.

tagset
/ˈtægset/, /'t{gset/, [N: tagset], [plural: -s]. Domain: dialogue representation. Def.: The set of tags used for labelling words in a particular language and in a particular corpus.

talk through
/ˈtɔːk ˈθruː/, /'tO:k 'Tru:/, [N: [V: talk][PREP: through]], [plural: none]. Synonyms: barge-in. Def.: Talk through is assumed to be of great importance in spoken dialogue systems for frequent users. Two types of talk through must be distinguished, one in which the human can only interrupt the system output, but without being understood; and another in which the human can stop the system output by starting to speak and the speech is understood.

talking face
/ˈtɔːkɪŋ ˈfeɪs/, /'tO:kIN 'feIs/, [N: [AJ: talking][N: face]], [plural: -s]. Domain: multimodal systems. Synonyms: talking head. Cohyponym: synthetic agent. Def.: Synthetic face.

talking head
/ˈtɔːkɪŋ ˈhed/, /'tO:kIN 'hed/, [N: [AJ: talking][N: head]], [plural: -s]. Domain: multimodal systems. Synonyms: talking face. Cohyponym: synthetic agent. Def.: Synthetic face.

tap
/ˈtæp/, /'t{p/, [N: tap], [plural: -s]. Hyperonyms: consonant; manner of articulation. Cohyponym: plosive, nasal, trill, fricative, lateral fricative, approximant, lateral approximant. Def.: Tap is a term used in the phonetic classification of consonant sounds on the basis of their manner of articulation: it refers to any sound produced by a single rapid contact with the roof of the mouth by the tongue, resembling a very brief articulation of a stop. (Crystal 1988, p. 304)

target-based model
/ˈtɑːgɪt ˈbeɪst ˈmɒdəl/, /'tA:gIt 'beIst 'mQd@l/, [N: [N: target][AJ: based][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: look-ahead model. Cohyponym: feature-based model, goal-based model. Def.: In the target-based model, positions are invariant in the sense that the articulator (lip shape) is forced to assume a given target without regard of the pattern of muscle contraction or how such a position might be achieved. Only the final target is considered. Depending on the context (i.e. the surrounding segments), a given target may be executed differently and different muscular contractions may be involved.

task
/ˈtɑːsk/, /'tA:sk/, [N: task], [plural: -s]. Def.: A task consists of all the activities which a user must develop in order to attain a fixed objective in some domain.

task-dependent vocabulary
/ˈtɑːsk dɪˈpendənt vəˈkæbjʊləri/, /'tA:sk dI'pend@nt v@'k{bjUl@ri/, [N: [N: task][AJ: dependent][N: vocabulary]], [plural: y/-ies]. Domain: speech recognition. Hyperonyms: vocabulary. Cohyponym: task-independent vocabulary. Def.: A task-dependent vocabulary is designed for a specific recognition task.

task-oriented dialogue
/ˈtɑːsk ɒriˈentɪd ˈdaɪəlɒg/ , /'tA:sk Qri'entId 'daI@lQg/ , [N: [N: task][AJ: oriented][N: dialogue]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: dialogue. Def.: A dialogue concerning a specific subject, aiming at an explicit goal (such as resolving a problem or obtaining specific information). For example, dialogues concerned with obtaining travel information or booking theatre tickets are task-oriented.

taxonomic relation
/tæksəˈnɒmɪk rɪˈleɪʃən/, /t{ks@'nQmIk rI'leIS@n/, [N: [AJ: taxonomic][N: relation]], [plural: -s]. Domain: terminology. Hyperonyms: semantic relation. Synonyms: taxonymic relation, ISA relation. Cohyponym: meronomic relation, mereonomic relation, PARTOF relation, meronymic relation. Def.: The term is rather general, and covers relations which have been referred to in other formalisms and theoretical frameworks with terms such as: paradigmatic relation, classification, taxonomy, field, family, similarity, set partition, subset-set inclusion, element-set membership, generalisation, property, implication, inheritance. Typical ISA relations define, in phonology, the natural classes characterised by distinctive feature vectors or by distributional classes based on syllable or word positions; in morphology, affix and stem classes; in phrasal syntax, parts of speech and constituent categories; in semantics, synonym, antonym and hyponym sets, or semantic fields.

taxonomy
/tæˈksɒnəmi/, /t{'ksQn@mi/, [N: taxonomy], [plural: y/-ies]. Domain: terminology. Hyperonyms: hierarchy. Synonyms: ISA hierarchy, generic concept hierarchy, logical concept hierarchy. Cohyponym: mereonomy, meronomy, PARTOF hierarchy, ontological hierarchy, partitive hierarchy. Def.: A hierarchy defined by the relation of generalisation and its inverse, specialisation.

taxonymic relation
/tæksəˈnɪmɪk rɪˈleɪʃən/, /t{ks@'nImIk rI'leIS@n/, [N: [AJ: taxonymic][N: relation]], [plural: -s]. Domain: terminology. Hyperonyms: semantic relation. Synonyms: taxonomic relation, ISA relation. Cohyponym: meronomic relation, meronymic relation, mereonomic relation, PARTOF relation. Def.: The term is rather general, and covers relations which have been referred to in other formalisms and theoretical frameworks with terms such as: paradigmatic relation, classification, taxonomy, field, family, similarity, set partition, subset-set inclusion, element-set membership, generalisation, property, implication, inheritance. Typical ISA relations define, in phonology, the natural classes characterised by distinctive feature vectors or by distributional classes based on syllable or word positions; in morphology, affix and stem classes; in phrasal syntax, parts of speech and constituent categories; in semantics, synonym, antonym and hyponym sets, or semantic fields.

TEI P3
/ˈtiːiːˈaɪ ˈpiːˈθriː/, /'ti:i:'aI 'pi:'Tri:/, [N: TEI P3], [plural: none]. Hyperonyms: TEI. Synonyms: Guidelines for Electronic Text Encoding and Interchange. Def.: These Guidelines are the result of over five years' effort by members of the research and academic community within the framework of an international cooperative project called the Text Encoding Initiative (TEI), established in 1987 under the joint sponsorship of the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing.

TEI
/ˈtiːiːˈaɪ/, /'ti:i:'aI/, [N: TEI], [plural: none]. Hyperonyms: SGML. Hyponyms: TEI P3. Synonyms: Text Encoding Initiative. Def.: The Text Encoding Initiative (TEI) is an international project to develop guidelines for the preparation and interchange of electronic texts for scholarly research, and to satisfy a broad range of uses by the language industries more generally.

template matching
/ˈtempleɪt ˈmætʃɪŋ/, /'templeIt 'm{tSIN/, [N: [N: template][N: matching]], [plural: none]. Domain: multimodal systems. Hyperonyms: face recognition. Hyponyms: Principle Component Analysis, PCA; geometric template matching, optical flow technique, deformable template matching, neural network based approach. Cohyponym: feature-based recognition. Def.: Images, represented as a two dimensional array of intensity values, are compared with an initial set of images, using adequate metric measurements. Template-based recognition represents images as an array of pixel values. Subimages can be masks of the eyes, nose, or mouth. The pixel value can be intensity values or may have been pre-processed by gradient or Laplacian filters to achieve scale, translation, and rotation independency. The recognition is performed by computing a normalised cross-correlation for each template, and finding the highest cumulative score.

term
/ˈtɜːm/, /'t3:m/, [N: term], [plural: -s]. Domain: terminology. Hyperonyms: word. Def.: The verbal representation of a technical concept.

termbank
/ˈtɜːmbæŋk/, /'t3:mb{Nk/, [N: termbank], [plural: -s]. Domain: terminology. Hyperonyms: database. Synonyms: termbase. Def.: A database containing the vocabulary of a special subject field.

termbase
/ˈtɜːmbeɪs/, /'t3:mbeIs/, [N: termbase], [plural: -s]. Domain: terminology. Hyperonyms: database. Synonyms: termbank. Def.: A database containing the vocabulary of a special subject field.

terminology science
/tɜːmɪˈnɒlədʒi ˈsaɪəns/, /t3:mI'nQl@dZi 'saI@ns/, [N: [N: terminology][N: science]], [plural: -s]. Domain: terminology. Def.: Science studying the structure, formation, development, usage and management of terminologies in various subject fields. (ISO CD 1087-1: 1997)

terminology
/tɜːmɪˈnɒlədʒi/, /t3:mI'nQl@dZi/, [N: terminology], [plural: y/-ies]. Domain: terminology. Def.: The set of designations belonging to one special language.

Text Encoding Initiative
/ˈtekst enˈkəʊdɪŋ ɪˈnɪʃətɪv/, /'tekst en'k@UdIN I'nIS@tIv/, [N:[N: Text][N: Encoding][N: Initiative]], [plural: none]. Hyperonyms: SGML. Hyponyms: TEI P3. Synonyms: TEI. Def.: The Text Encoding Initiative (TEI) is an international project to develop guidelines for the preparation and interchange of electronic texts for scholarly research, and to satisfy a broad range of uses by the language industries more generally.

text preprocessing
/ˈtekst priːˈprəʊsesɪŋ/, /'tekst pri:'pr@UsesIN/, [N: [N: text][N: preprocessing]], [plural: none]. Domain: speech synthesis. Meronym. sup.: linguistic interface. Def.: The first stage of the linguistic interface of a text-to-speech system, which handles punctuation marks and other non-alphabetic textual symbols (e.g. parentheses), and expands abbreviations, acronyms, numbers, special symbols, etc. to full-blown orthographic strings (Gibbon et al. 1997, p. 851).

text-dependent speaker recognition system
/ˈtekst dɪˈpendənt ˈspiːkə rekəgˈnɪʃən ˈsɪstəm/, **/’tekst dI’pend@nt ’spi:k@ rek@g’nIS@n ’sIst@m/**, [N: [N: text][AJ: dependent][N: speaker][N: recognition][N: system]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker recognition system. Cohyponym: text-independent speaker recognition system . Def.: A speaker recognition system for which the training and test speech utterances are composed of exactly the same linguistic material, in the same order (typically, a password).

text-independent speaker recognition system
/ˈtekst ɪndɪˈpendənt ˈspiːkə rekəgˈnɪʃən ˈsɪstəm/ , **/’tekst IndI’pend@nt ’spi:k@ rek@g’nIS@n ’sIst@m/** , [N: [N: text][AJ: independent][N: speaker][N: recognition][N: system]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: speaker recognition system . Hyponyms: unrestricted text-independent speaker recognition system, event-dependent speaker recognition system. Cohyponym: text-dependent speaker recognition system. Def.: A speaker recognition system for which the linguistic content of test speech utterances varies across trials.

text-prompted speaker recognition system
/ˈtekst ˈprɒmptɪd ˈspiːkə rekəgˈnɪʃən ˈsɪstəm/ , **/’tekst ’prQmptId ’spi:k@ rek@g’nIS@n ’sIst@m/** , [N: [N: text][AJ: prompted][N: speaker][N: recognition][N: system]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: speaker recognition system . Cohyponym: voice-prompted speaker recognition system, unprompted speaker recognition system. Def.: A speaker recognition system for which, during the test phase, a written text is prompted (through an appropriate device) to the user, who has to read it aloud.

text-to-speech system
/ˈtekst tə ˈspiːtʃ ˈsɪstəm/ , **/’tekst t@ ’spi:tS ’sIst@m/** , [N: [N: text][PREP: to][N: speech][N: system]], [plural: -s]. Domain: speech synthesis. Hyperonyms: speech output system. Cohyponym: concept-to-speech system. Def.: Speech output system that converts orthographic text (generally stored in a computer memory as ASCII codes) into speech.

text-to-visual-speech face synthesis
/ˈtekst tə ˈvɪʒʊəl ˈspiːtʃ ˈfeɪs ˈsɪnθəsɪs/, **/’tekst t@ ’vIZU@l ’spi:tS ’feIs ’sInT@sIs/**, [N: [N: text][PREP: to][AJ: visual][N: speech][N: face][N: synthesis]], [plural: none]. Domain: multimodal systems. Hyperonyms: face synthesis. Cohyponym: puppeteer control face synthesis, performance-driven face synthsis, audio-driven face synthesis. Def.: The input of the system is plain text. The input text is first decomposed into its phonetic representation. Information about phonemes and their duration are automatically generated from the text. Formants and other speech parameters (frequency, pitch, pitch range and so on) are then computed. The text-to-visual-speech technique is suited when parametric facial models are used. Parameters defining facial animation are added to the set of speech parameters: lip shape, facial expressions, jaw rotation, etc. As a novel approach, speech synthesis systems have been extended to include facial parameters in their speech output parameters. The parallel computation of the auditory and visual parameters ensures a perfect synchronisation of the two channels, which is an advantage of such a technique. But different sampling rates of the speech synthesiser and of the animation system have to be reconciled. While the animation system uses 25-30 frames/sec, an acceptable audio system requires at least 50-60 frames/sec. To avoid temporal aliasing effect of the visual images, motion blur between successive frames can be used. Parameter values driving the facial model are blurred with their neighborhood (corresponding to the precedent and successive frames) parameters using a Gaussian filter. Text-to-visual-speech systems may be enhanced by adding markers describing intonation, speech rhythm, type of voice to the input text. Speech would be of better quality and such parameters could be used to get a more complex facial animation. For example, accents could be synchronised with raised eyebrows and head nods. Different facial models corresponding to different types of voice have also been explored.

theory-based evaluation
/ˈθiːəri ˈbeɪst ɪvæljʊˈeɪʃən/, /'Ti:@ri 'beIst Iv{ljU'eIS@n/, [N: [N: theory][AJ: based][N: evaluation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: evaluation methodology. Cohyponym: expert-based evaluation, user-based evaluation. Def.: Theory-based evaluation involves a designer or evaluator who models task and user, based on the system specification. This ultimately generates quantitative values for interaction times, learnability or usability of the evaluated system. The evaluation involves neither a user-computer interaction nor a system prototype.

time-locked model
/ˈtaɪm ˈlɒkt ˈmɒdəl/, /'taIm 'lQkt 'mQd@l/, [N: [N: time][AJ: locked][N: model]], [plural: -s]. Domain: multimodal systems. Hyperonyms: coarticulation model. Cohyponym: look-ahead model, hybrid model, expansion model. Def.: The time-locked model is based on the principle that an event starts from an inherent time (a locked time). The protrusion influence due to a vowel appears at a given time before the vowel.

ToBI
/ˈtəʊbi/, /'t@Ubi/, [N: ToBI], [plural: none]. Hyponyms: E_ToBI, GlaToBI, J_ToBI,. Synonyms: Tone and Break Indices. Def.: A system of prosodic transcription which concentrates exclusively on represeting perceived pitch patterns (tones, 'To') in terms of target tone heights (usually two) and a hierarchy of boundary indices ('BI'). It has become perhaps the most popular and consistently applicable variety of prosodic transcription.

topic identification
/ˈtɒpɪk aɪdentɪfɪˈkeɪʃən/, /'tQpIk aIdentIfI'keIS@n/, [N: [N: topic][N: identification]], [plural: -s]. Hyperonyms: task. Synonyms: topic spotting. Def.: The determination of the topic of some speech or text material.

topline reference (condition)
/ˈtəplaɪn ˈrefərəns/, /'t@plaIn 'ref@r@ns/, [N: [N: topline][N: reference][N: condition]], [plural: -s]. Domain: speech synthesis. Hyperonyms: speech output. Cohyponym: baseline reference condition. Def.: Speech output that represents optimum performance, typically by a professional human talker.

tracking accuracy
/ˈtrækɪŋ ˈækjʊrəsi/, /'tr{kIN '{kjUr@si/, [N: [N: tracking][N: accuracy]], [plural: y/-ies]. Domain: multimodal systems. Def.: Percent deviation from true (facial) feature position.

tracking success
/ˈtrækɪŋ səkˈses/, /'tr{kIN s@k'ses/, [N: [N: tracking][N: success]], [plural: -es]. Domain: multimodal systems. Def.: Ratio of time when feature is tracked and time when feature is lost.

training
/ˈtreɪnɪŋ/, /'treInIN/, [N: training], [plural: -s]. Domain: speech recognition. Hyperonyms: process. Def.: The process in which a speech recognition system learns the pronunciation of words to be recognised at a later instance.

transaction
/trænsˈækʃən/, /tr{ns'{kS@n/, [N: transaction], [plural:s]. Domain: interactive dialogue systems. Meronym. sup.: dialogue . Def.: The part of a dialogue devoted to a single high-level task (for example, making a travel booking or checking a bank account balance). Atransaction may be coextensive with a dialogue, or a dialogue may consist of more than one transaction.

TRANSCRIBER
/træn'skraɪbə/, /tr{n'skraIb@/, [N: TRANSCRIBER], [plural: none]. Hyperonyms: annotation tool. Def.: A public domain tool for segmenting, labelling, and transcribing speech. It is written in Tcl/tk script language and is freely available as free software. TRANSCRIBER allows segmenting, labelling, and transcribing long duration signals. The output is in a standard SGML format. Multiple languages are supported. The tool can be ported to various platforms and is very flexible so that new functions can be easily added.

transcription
/træn'skrɪpʃən/, /tr{n'skrIpS@n/, [N: transcription], [plural: -s]. Hyperonyms: notation. Hyponyms: phonetic transcription, phonemic transcription. Cohyponym: orthographic transcription. Def.: A method of writing down the names of speech sounds in a systematic and consistent way. (Crystal 1988, p. 313)

transfer
/'trænsfɜː/, /'tr{nsf3:/, [N: transfer], [plural: none]. Domain: multimodal systems. Hyperonyms: cooperation type. Cohyponym: complementarity, redundancy, equivalence, specialisation, concurrency. Def.: A chunk of information in one modality triggers an event in another modality. E.g. In hypermedia interfaces: a mouse click provokes the display of an image..

trigram
/'traɪgræm/, /'traIgr{m/, [N: trigram], [plural: -s]. Domain: language modelling. Cohyponym: zerogram, unigram, bigram. Def.: In language modelling a trigram is sequence of three words. (Gibbon et al. 1997, p. 94)

trill
/'trɪl/, /'trIl/, [N: trill], [plural: -s]. Hyperonyms: consonant; manner of articulation. Cohyponym: plosive, nasal, tap, flap, fricative, lateral fricative, approximant, lateral approximant. Def.: Trill is a term used in the phonetic classification of consonant sounds on the basis of ther manner of articulation: it refers to any sound made by the rapid tapping of one organ of articulation against another. (Crystal 1988, p. 318) E.g. Example in English: [r].

turn duration
/'tɜːn dju'reɪʃən/, /'t3:n dju'reIS@n/, [N: [N: turn][N: duration]], [plural: –]. Domain: interactive dialogue systems. Def.: Turn duration is a measure of the average duration of one turn in a corpus of dialogues. (Gibbon et al. 1997, p. 606)

turn
/'tɜːn/, /'t3:n/, [N: turn], [plural: -s]. Domain: interactive dialogue systems. Meronym. sup.: dialogue. Def.: A stretch of speech, spoken continuously by one party in a dialogue. A stretch of speech may contain several linguistic acts or actions. A dialogue consists of a sequence of turns produced alternately by each party. Turns are also known as utterances.

UCR
/'juːsiː'ɑː/, /'ju:si:'A:/, [N: UCR], [plural: -s]. Hyperonyms: correction rate. Synonyms: user correction rate. Def.: Percentage of all user turns which are correction turns.

ultradirectional microphone
/ʌltrədaɪ'rekʃənəl 'maɪkrəfəʊn/, /Vltr@daI'rekS@n@l 'maIkr@f@Un/, [N: [AJ: ultradirectional][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: microphone. Cohyponym: unidirectional microphone, bidirectional microphone, omnidirectional microphone, pressure zone microphone, headset microphone. Def.: The ultradirectional microphone is designed for distant pickup, e.g. in film or TV productions. It strongly attenuates off-axis sound by means of multipath interference at a long slotted tube mounted in front of a unidirectional microphone. Compared to omni- and unidirectional microphones the sound quality is relatively poor since it has been traded against good directivity. The ultradirectional microphone is not recommended for high-quality speech recordings. (Gibbon et al. 1997, p. 305)

unacquainted impostor
/ʌnə'kweɪntɪd ɪm'pɒstə/, /Vn@'kweIntId Im'pQst@/, [N: [AJ: unacquainted][N: impostor]], [plural: -s]. Domain: speaker recognition. Hyperonyms: intentional impostor. Cohyponym: acquainted impostor. Def.: An unacquainted impostor has never been in contact with the genuine or authentic user. (Gibbon et al. 1997, p. 422)

uncooperative speaker
/ʌnkəʊ'ɒpərətɪv 'spiːkə/, /Vnk@U'Qp@r@tIv 'spi:k@/, [N: [AJ: uncooperative][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: registered speaker. Cohyponym: cooperative speaker. Def.: When the user's goal and the system's purpose are inverse, an uncooperative (registered) speaker knows that he is being verified but wants the system to reject him. (Gibbon et al. 1997, p. 422)

unidirectional microphone
/juːnɪdaɪ'rekʃənəl 'maɪkrəfəʊn/, /ju:nIdaI'rekS@n@l 'maIkr@f@Un/, [N: [AJ: unidirectional][N: microphone]], [plural: -s]. Domain: physical characterisation. Hyperonyms: microphone. Hyponyms: cardioid microphone, hypercardioid microphone, supercardioid microphone. Cohyponym: bidirectional microphone, omnidirectional microphone, ultradirectional microphone, pressure zone microphone, headset microphone. Def.: The unidirectional type of microphone is most sensitive to sound arriving from one direction and more or less attenuates incident sound from other directions. Thus, unidirectional microphones will suppress intended sound when pointed at the wanted sound source, i.e. the speaker. (Gibbon et al. 1997, p. 303)

uniform language model
/'juːnɪfɔːm 'læŋgwɪdʒ 'mɒdəl/, /'ju:nIfO:m 'l{NgwIdZ 'mQd@l/, [N: [AJ: uniform][N: language][N: model]], [plural: -s]. Domain: language modelling. Hyperonyms: language model. Def.: Here, the idea is to use the same probability for all events; events can be either the words of the vocabulary or the sentences, if the number of sentences is limited. (Gibbon et al. 1997, p. 243)

Uniform Resource Locator
/'juːnɪfɔːm rɪ'sɔːs ləʊ'keɪtə/, /'ju:nIfO:m rI'sO:s l@U'keIt@/, [N: [AJ: Uniform][N: Resource][N: Locator]], [plural: -s]. Synonyms: URL. Def.: 1. Generally, a type of combined file name, data access pointer, and access parameters, for files located at arbitrary positions in a network. 2. Specifically, an Internet address, supplemented by detailed access information. A web client (WWW browser) requests a document via a URL from a WWW server. A URL has the form 'protocol://address:port/path/file#anchor?value_list'. 'protocol': an Internet protocol such as 'http', 'ftp', 'news', etc.; 'address': either an IP number or IP address; 'port': an operating system communications port number; 'path': a path name relative to the web server's root directory; 'file': a file name; 'anchor': a named position within the file; 'value_list': a list of attribute-value pairs written as 'attribute=value'. Attribute-value pairs are separated by '&'. URLs can be partial only; missing parts are substituted with default values by the server. The server interprets the URL and returns the requested document to the client.

unigram
/'juːnɪgræm/, /'ju:nIgr{m/, [N: unigram], [plural: -s]. Domain: language modelling. Cohyponym: bigram, trigram, zerogram. Def.: In language modelling: single word.

unimodal input event
/juːni'məʊdəl 'ɪnpʊt ɪ'vent/, /ju:ni'm@Ud@l 'InpUt I'vent/, [N: [AJ: unimodal][N: input][N: event]], [plural: -s]. Domain: multimodal systems. Hyperonyms: input event. Synonyms: monomodal input event. Cohyponym: multimodal input event. Def.: Set of user input events that belong together and are intended to convey a chunk of information, consisting of at least two parts in one modality.

unintentional impostor
/ʌnɪn'tenʃənəl ɪm'pɒstə/ , **/VnIn'tenS@n@l Im'pQst@/** , [N: [AJ: unintentional][N: impostor]], [plural: -s]. Domain: speaker recognition]. Hyperonyms: impostor. Cohyponym: intentional impostor. Def.: An unintentional impostor does not have the clear goal of being identified or verified or to be identified as somebody else.

unprompted speaker recognition
/ʌn'prɒmptɪd 'spiːkə rekəg'nɪʃən/, **/Vn'prQmptId 'spi:k@ rek@g'nIS@n/**, [N: [AJ: unprompted][N: speaker][N: recognition]], [plural: –]. Domain: speaker recognition. Hyperonyms: speaker recognition. Cohyponym: voice-prompted speaker recognition, text-prompted speaker recognition. Def.: Speaker recognition where totally spontaneous speech is used, i.e. for which the user is totally free to utter what he wants. (Here, a further distinction could be made between language dependent and language independent systems), or for which the system has no control over the speaker. (For instance, in forensic applications, the speaker may not be physically present, or may not be willing to cooperate.)

unreliable speaker
/ʌnrɪ'laɪəbəl 'spiːkə/, **/VnrI'laI@b@l 'spi:k@/**, [N: [AJ: unreliable][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker. Synonyms: goat. Cohyponym: dependable speaker. Def.: A speaker with a high misclassification rate. (Gibbon et al. 1997, p. 432) called {em a goatindex{goat}}.}

unrestricted text-independent speaker recognition system
/ʌnrɪ'strɪktɪd 'tekst ɪndɪ'pendənt 'spiːkə rekəg'nɪʃən 'sɪstəm/ , **/VnrI'strIktId 'tekst IndI'pend@nt 'spi:k@ rek@g'nIS@n 'sIst@m/** , [N: [AJ: unrestricted][N: text][AJ: independent][N: speaker][N: recognition][N: system]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: text-independent speaker recognition system. Meronym. sup.: text-independent speaker recognition. Def.: A text-independent speaker recognition system for which no constraints apply regarding the linguistic content of the test speech material.

unseen event
/'ʌnsiːn ɪ'vent/, **/'Vnsi:n I'vent/**, [N: [AJ: unseen][N: event]], [plural: -s]. Domain: language modelling. Hyperonyms: event. Cohyponym: singleton event, doubleton event. Def.: Event not observed in the training data. (Gibbon et al. 1997, p. 248, 249)

URL
/'juːɑːr'el/, **/'ju:A:r'el/**, [N: URL], [plural: -s]. Synonyms: Uniform Resource Locator. Def.: 1. Generally, a type of combined file name, data access pointer, and access parameters, for files located at arbitrary positions in a network. 2. Specifically, an Internet address, supplemented by detailed access information. A web client (WWW browser) requests a document via a URL from a WWW server. A URL has the form 'protocol://address:port/path/file#anchor?value_list'. 'protocol': an Internet protocol such as 'http', 'ftp', 'news', etc.; 'address': either an IP number or IP address; 'port': an operating system communications port number; 'path': a path name relative to the web server's root directory; 'file': a file name; 'anchor': a named position within the file; 'value_list': a list of attribute-value pairs written as 'attribute=value'. Attribute-value pairs are separated by '&'. URLs can be partial only; missing parts are substituted with default values by the server. The server interprets the URL and returns the requested document to the client.

user correction rate
/'juːzə kə'rekʃən 'reɪt/, **/'ju:z@ k@'rekS@n 'reIt/**, [N: [N: user][N: correction][N: rate]], [plural: -s]. Hyperonyms: correction rate. Synonyms: UCR. Def.: Percentage of all user turns which are correction turns.

user study
/'juːzə 'stʌdi/, /'ju:z@ 'stVdi/, [N: [N: user][N: study]], [plural: y/-ies]. Hyperonyms: experimental technique. Cohyponym: benchmark evaluation, simulation study, iterative design, rapid prototyping. Def.: If a prototype of a multimodal application has been implemented, informal or formal studies of users performing real tasks using the system can be performed. User studies typically yield a rich set of data, ranging from quantitative measures to informal observations. Additionally, user studies can be employed to build up a database of multimodal interactions for later benchmark evaluations. User studies however are quite costly, and require a careful experimental design of the study.

user vocabulary size
/'juːzə və'kæbjʊləri 'saɪz/, /'ju:z@ v@'k{bjUl@ri 'saIz/, [N: [N: user][N: vocabulary][N: size]], [plural: -s]. Domain: consumer off-the-shelf products, speech recognition. Hyperonyms: vocabulary size. Synonyms: extension vocabulary size, exception vocabulary size. Cohyponym: active vocabulary size, passive vocabulary size. Def.: The number of words a user may add to the lexicon of a speech recogniser.

user-added word
/'juːzərædɪd 'wɜːd/, /'ju:z@r{dId 'w3:d/, [N: [N: user][AJ: added][N: word]], [plural: -s]. Domain: speech recognition. Def.: Words added to the vocabulary of a speech recogniser by the user (i.e. the caller in the case of a telephone-based system).

user-based evaluation
/'juːzə 'beɪst ɪvæljʊ'eɪʃən/, /'ju:z@ 'beIst Iv{ljU'eIS@n/, [N: [N: user][AJ: based][N: evaluation]], [plural: -s]. Domain: multimodal systems. Hyperonyms: evaluation methodology. Cohyponym: theory-based evaluation, expert-based evaluation. Def.: The user-based approach involves one or more users completing one or more tasks. Task, user, and environment characteristics must match those for which the system is being designed. Data on how user and system behave are collected while the user performs experimental tasks.

uvular consonant
/'juːvjʊlə 'kɒnsənənt/, /'ju:vjUl@ 'kQns@n@nt/, [N: [AJ: uvular][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, labiodental consonant, dental consonant, alveolar consonant, postalveolar consonant, retroflex consonant, palatal consonant, velar consonant, pharyngeal consonant, glottal consonant. Def.: Uvular consonant is a term used in the phonetic classification of consonant sounds on the basis of their place of articulation: it refers to a sound made by the back of the tongue against the uvula. (Crystal 1988, p. 322)

validation
/vælɪ'deɪʃən/, /v{lI'deIS@n/, [N: validation], [plural: -s]. Domain: corpora. Hyperonyms: evaluation technique. Cohyponym: monitoring (antonym of 'validation' in definition 1); evaluation (antonym of 'validation' in definition 2). Def.: 1. Validation relates to an off-line (or post hoc) technical or phonetic evaluation of the material recorded. (Gibbon et al. 1997, p. 129) 2. Process of determining whether a given resource meets its formal specification.

velar consonant
/'viːlə 'kɒnsənənt/, /'vi:l@ 'kQns@n@nt/, [N: [AJ: velar][N: consonant]], [plural: -s]. Hyperonyms: consonant. Cohyponym: bilabial consonant, labiodental consonant, dental consonant, alveolar consonant, postalveolar consonant, retroflex consonant, palatal consonant, uvular consonant, pharyngeal consonant, glottal consonant. Def.: Velar consonant is a term used in the phonetic classification of consonant sounds on the basis of their place of articulation: it refers to a sound made by the back of the tongue against the soft palate or velum. (Crystal 1988, p. 324) E.g. Examples in English: [k, g].

verb
/ˈvɜːb/, **/'v3:b/**, [N: verb], [plural: -s]. Domain: lexicon. Hyperonyms: lexical category. Cohyponym: noun, adjective, adverb. Def.: One of the four main lexical categories (parts of speech), typically occurring as main predicate of a sentence, modified by adverbs and by noun and sentence complements, and in many languages inflecting (inter alia) for person, number, and tense.

vernacular
/vəˈnækjʊlə/, **/v@'n{kjUl@/**, [N: vernacular], [plural: -s]. Hyperonyms: language variety. Cohyponym: standard, Lingua Franca (Crystal 1988). Meronym. sup.: natural language. Def.: Vernacular is a term used in sociolinguistics to refer to the indigenous language or dialect of a speech community. (Crystal 1988) E.g. vernacular of Liverpool, Berkshire, Jamaica (Crystal 1988).

violated speaker
/ˈvaɪəleɪtɪd ˈspiːkə/, **/'vaI@leItId 'spi:k@/**, [N: [AJ: violated][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: registered speaker. Def.: The registered speaker owning the identity assigned erroneously to an impostor in open-set speaker identification. The registered speaker owning the identity claimed by a successful impostor in speaker verification. (Gibbon et al. 1997, p. 414)

viseme
/ˈvɪziːm/, **/'vIzi:m/**, [N: viseme], [plural: -s]. Domain: multimodal systems. Def.: Smallest unit of lip movement (while speaking) that can make a difference in meaning.

visual output device
/ˈvɪʒʊəl ˈaʊtpʊt dɪˈvaɪs/, **/'vIZU@l 'aUtpUt dI'vaIs/**, [N: [AJ: visual][N: output][N: device]], [plural: -s]. Domain: multimodal systems. Hyperonyms: output device. Cohyponym: acoustic output device, haptic output device. Def.: Visual display by using a monitor is the most used means of communication via computer. Virtual reality, stereoscopic monitors, and immersive systems enhance spatial information by displaying data in 3D.

Viterbi alignment
/vɪˈtɜːbi əˈlaɪnmənt/, **/vI't3:bi @'laInm@nt/**, [N: [N: Viterbi][N: alignment]], [plural: -s]. Domain: language modelling. Hyperonyms: alignment algorithm. Synonyms: Viterbi decoding, Viterbi approximation, maximum approximation. Meronym. sup.: HMM recogniser. Def.: A widely used alignment algorithm that finds the best path through a probability graph. Viterbi alignment is usually applied to HMM output.

Viterbi approximation
/vɪˈtɜːbi əprɒksiːˈmeɪʃən/, **/vI't3:bi @prQksi:'meIS@n/**, [N: [N: Viterbi][N: approximation]], [plural: -s]. Domain: language modelling. Hyperonyms: algorithm. Synonyms: maximum approximation, Viterbi alignment, Viterbi decoding. Def.: In the so-called Viterbi or maximum approximation, the sum over all paths is approximated by the path which has the maximum contribution to the sum.

Viterbi decoding
/vɪˈtɜːbi diːˈkəʊdɪŋ/, **/vI't3:bi di:'k@UdIN/**, [N: [N: Viterbi][N: decoding]], [plural: –]. Domain: language modelling. Hyperonyms: alignment algorithm. Synonyms: Viterbi alignment, Viterbi approximation, maximum approximation. Meronym. sup.: HMM recogniser. Def.: An algorithm used in pattern recognition that, under certain assumptions finds the best non-linear alignment of two sequences so as to maximise their similarity.

vocabulary size
/vəˈkæbjʊləri ˈsaɪz/, /v@’k{bjUl@ri ’saIz/, [N: [N: vocabulary][N: size]], [plural: -s]. Domain: speech recognition, consumer off-the-shelf products. Hyponyms: active vocabulary size, passive vocabulary size, user vocabulary size, extension vocabulary size, exception vocabulary size. Synonyms: coverage, vocabulary coverage . Meronym. sub.: extensional coverage, intensional coverage. Def.: The size of the vocabulary is defined as the number of words that a speech recogniser can handle.

vocabulary
/vəˈkæbjʊləri/, /v@’k{bjUl@ri/, [N: vocabulary], [plural: y/-ies]. Hyponyms: active vocabulary, backup vocabulary. Synonyms: dictionary. Meronym. sup.: speech recogniser. Def.: The set of words in a lexicon, such as the set of words that an automatic speech recognition system is capable to recognise.

voice characteristic
/ˈvɔɪs kærəktəˈrɪstɪk/, /’vOIs k{r@kt@’rIstIk/, [N: [N: voice][N: characteristic]], [plural: -s]. Def.: Those aspects of speech which remain relatively constant over longer stretches of speech, and constitute the background against which segmental and prosodic variation is produced and perceived (e.g. mean pitch level, mean loudness, mean tempo, harshness, creak, whisper, tongue body orientation, dialect).

voice disorder
/ˈvɔɪs dɪsˈɔːdə/, /’vOIs dIs’O:d@/, [N: [N: voice][N: disorder]], [plural: -s]. Domain: corpora. Hyperonyms: speech disorder. Synonyms: disphonia. Cohyponym: articulation disorder, resonance disorder, language disorder, rhythm disorder. Def.: A voice disorder involves lesions of the vocal cords. The voice may emerge as a whisper (no vocal-cord vibration), for instance due to paralysis; or vocal-cord vibration may be present to some degree, but accompanied by excessive air flow (a “breathy” voice); or there may be irregular and therefore aperiodic vocal fold vibration, for instance due to the growth of abnormal tissue (nodules) on the vocal folds, resulting in a “hoarse” voice quality. Dysphonia may be caused by psychological and emotional factors, such as a severe shock, or by organic factors. A serious voice disorder is cancer of the vocal cords. (Gibbon et al. 1997, p. 115)

Voice Manager
/ˈvɔɪs ˈmænɪdʒə/, /’vOIs ’m{nIdZ@/, [N: [N: Voice][N: Manager]], [plural: -s]. Domain: consumer off-the-shelf products, speech recognition. Hyperonyms: tool. Meronym. sup.: speech recogniser. Def.: a commercial off-the-shelf (COTS) The voice commanding part of a recognition system that can track the contents of the currently active application (the window that has the focus), and dynamically adapts its active vocabulary. All normal Windows widgets can be read, such as buttons, menus, check boxes, radio buttons, pull down menus etc.

voice onset time
/vɔɪs ˈɒnset taɪm/, /vOIs ’Qnset taIm/, [N: [N: voice][N: onset][N: time]], [plural: -s]. Synonyms: voice onset time. Def.: The delay, which may be negative, between the release of a plosive and the onset of voicing; if there is perceptible delay, this may result in aspiration of the plosive, as with initial pre-vocalic plosives in English or German words.

voice-prompted speaker recognition system
/ˈvɔɪs ˈprɒmptɪd ˈspiːkə rekəgˈnɪʃən ˈsɪstəm/ , /’vOIs ’prQmptId ’spi:k@ rek@g’nIS@n ’sIst@m/ , [N: [N: voice][AJ: prompted][N: speaker][N: recognition][N: system]] , [plural: -s] . Domain: speaker recognition. Hyperonyms: speaker recognition system . Cohyponym: text-prompted speaker recognition system, unprompted speaker recognition system. Def.: A speaker recognition system for which, during the test phase, the user has to repeat a speech utterance, which he listens to through an audio device.

voicing decision
/ˈvɔɪsɪŋ dɪˈsɪʒən/, /’vOIsIN dI’sIZ@n/, [N: [N: voicing][N: decision]], [plural: -s]. Def.: The decision, used, for example, in certain speech coders, as to whether a particular section of speech is voiced or unvoiced.

voicing
/ˈvɔɪsɪŋ/, /'vOIsIN/, [N: voicing], [plural: –]. Domain: corpora, speech synthesis. Meronym. sup.: speech sound. Def.: A vibration of the vocal cords during the production of vowels and some consonants.

VOT
/ˈviː əʊ ˈtiː/, /'vi: @U 'ti:/, [N: VOT], [plural: -s]. Synonyms: voice onset time.. Def.: The delay, which may be negative, between the release of a plosive and the onset of voicing; if there is perceptible delay, this may result in aspiration of the plosive, as with initial pre-vocalic plosives in English or German words.

vowel
/ˈvaʊəl/, /'vaU@l/, [N: vowel], [plural: -s]. Cohyponym: consonant. Meronym. sup.: speech sound. Def.: A speech sound produced with relatively open vocal tract, without constriction or blockage of the airflow.

vulnerable speaker
/ˈvʌlnərəbəl ˈspiːkə/, /'vVln@r@b@l 'spi:k@/, [N: [AJ: vulnerable][N: speaker]], [plural: -s]. Domain: speaker recognition. Hyperonyms: speaker. Synonyms: lamb. Cohyponym: resistant speaker. Def.: A speaker with a high mistrust rate.(Gibbon et al. 1997, p. 433)

well-intentioned impostor
/ˈwel ɪnˈtenʃənd ɪmˈpɒstə/, /'wel In'tenS@nd Im'pQst@/, [N: [AV: well][AJ: intentioned][N: impostor]], [plural: -s]. Domain: speaker recognition. Hyperonyms: impostor. Def.: An impostor having the goal of being rejected. (Gibbon et al. 1997, p. 422)

white noise
/ˈwaɪt ˈnɔɪz/, /'waIt 'nOIz/, [N: [AJ: white][N: noise]], [plural: –]. Domain: physical characterisation. Hyperonyms: noise. Def.: Noise-like sound in which all possible frequencies in the range of hearing are randomly present, at random amplitudes and in random phase relationships. (Clark & Yallop, p. 218)

WIMP
/ˈwɪmp/, /'wImp/, [N: WIMP], [plural: -s]. Domain: multimodal systems. Hyperonyms: interface. Def.: An interface with windows (W), icons (I), menus (M) and pointing (P) facilities. The association is with the English word 'wimp', a favourite of Margaret Thatcher's, meaning 'a male person of weak character'.

wizard production variable
/ˈwɪzəd prəˈdʌkʃən ˈveərɪəbəl/, /'wIz@d pr@'dVkS@n 've@rI@b@l/, [N: [N: wizard][N: production][N: variable]], [plural: -s]. Domain: interactive dialogue systems. Cohyponym: wizard recognition variable. Def.: Speech generation variable such as voice quality, intonation, syntax, register. One production variable of particular interest is the wizard's response time. (Gibbon et al. 1997, p. 587)

wizard recognition variable
/ˈwɪzəd rekəgˈnɪʃən ˈveərɪəbəl/, /'wIz@d rek@g'nIS@n 've@rI@b@l/, [N: [N: wizard][N: recognition][N: variable]], [plural: -s]. Domain: interactive dialogue systems. Cohyponym: wizard production variable. Def.: Wizard recognition variables define the ranges of acoustic, lexical, syntactic and pragmatic phenomena which the wizard is allowed to recognise.

Wizard-of-Oz simulation
/ˈwɪzəd əv ˈɒz sɪmjuːˈleɪʃən/, /'wIz@d @v 'Qz sImju:'leIS@n/, [N: [N: Wizard][PREP: of][N: Oz][N: simulation]], [plural: -s]. Domain: interactive dialogue systems. Hyperonyms: simulation. Synonyms: human-machine-communication simulation. Def.: Simulation of the behaviour of an interactive automaton by a human being. This can be done (i) by speaking to the user in a disguised voice, (ii) by choosing and triggering system predefined responses, (iii) by manually modifying some parameters of the simulation system, or (iv) by using a person to simulate the integration of existing system components (a bionic Wizard-of-Oz simulation)

wolf
/ˈwʊlf/, /'wUlf/, [N: wolf], [plural: wolves]. Domain: speaker recognition. Hyperonyms: impostor. Synonyms: skilled impostor. Cohyponym: poor impostor, badger. Def.: Impostor with a high success rate in claiming an identity averaged over each claimed identity. (Gibbon et al. 1997, p. 441)

word error rate
/ˈwɜːd ˈerə ˈreɪt/, /'w3:d 'er@ 'reIt/, [N: [N: word][N: error][N: rate]], [plural: -s]. Domain: speech recognition. Hyperonyms: error rate. Cohyponym: sentence error rate. Def.: Proportion of the words in a test that are misrecognised.

word formation
/wɜːd fɔːˈmeɪsən/, /w3:d fO:'meIs@n/, [N: [N: word][N: formation]], [plural: none]. Domain: lexicon. Hyperonyms: morphological operation. Cohyponym: inflection. Meronym. sup.: morphology. Meronym. sub.: compounding, derivation. Def.: Word formation deals with the construction of words from smaller meaningful parts. (Gibbon et al. 1997, p. 214) E.g. 1. re + activate = reactivate (derivation) 2. wind + mill = windmill (compounding).

word graph
/ˈwɜːd ˈgrɑːf/, /'w3:d 'grA:f/, [N: [N: word][N: graph]], [plural: -s]. Domain: language modelling. Meronym. sup.: search in speech recognition. Def.: A word graph or lattice is used in the context of search in speech recognition to provide an explicit interface between the acoustic recognition and the application of the language model. The word graph or lattice should contain the most likely word hypotheses where in addition to the word hypothesis the start and end times, the nodes and an acoustic probability are given.

word lattice
/ˈwɜːd ˈlætɪs/, /'w3:d 'l{tIs/, [N: [N: word][N: lattice]], [plural: -s]. Domain: language modelling. Meronym. sup.: search in speech recognition. Def.: A word graph or lattice is used in the context of search in speech recognition to provide an explicit interface between the acoustic recognition and the application of the language model. The word graph or lattice should contain the most likely word hypotheses where in addition to the word hypothesis the start and end times, the nodes and an acoustic probability are given.

word monitoring
/ˈwɜːd ˈmɒnɪtərɪŋ/, /'w3:d 'mQnIt@rIN/, [N: [N: word][N: monitoring]], [plural: none]. Domain: speech synthesis. Hyperonyms: monitoring. Cohyponym: phoneme monitoring, syllable monitoring. Def.: Testing the intelligibility of whole words in isolation as well as in various types of context. (Gibbon et al. 1997, p. 490)

word spotting
/ˈwɜːd ˈspɒtɪŋ/, /'w3:d 'spQtIN/, [N: [N: word][N: spotting]], [plural: none]. Domain: speech recognition. Hyperonyms: isolated word speech recognition. Def.: The procedure of checking a speech signal - 'listening' - for a small subset of the words used in producing the signal. (Gibbon et al. 1997, p. 95)

word
/'wɜːd/, /'w3:d/, [N: word], [plural: -s]. Hyperonyms: lexical unit. Hyponyms: orthographic word, phonological word, morphological word, syntactic word, prosodic word. Def.: 1. The basic type lexical sign in a language, smaller than an idiom and larger than a morpheme; a word may be simplex or complex (compound or derived). 2. Linguistic textbooks distinguish between several different views of words as lexical units, depending on which kind of lexical sign information is regarded as primary: see 'phonological word, 'morphological word, 'orthographic word, 'prosodic word, syntactic word. (Gibbon et al. 1997, p. 196) 3. In automatic speech recognition: every sequence of characters between blanks is a word. (Gibbon et al. 1997, p. 246)

written language corpus
/'rɪtən 'læŋgwɪdʒ 'kɔːpəs/, /'rIt@n 'l{NgwIdZ 'kO:p@s/, [N: [AJ: written][N: language][N: corpus]], [plural: written language corpora]. Domain: corpora. Hyperonyms: corpus. Cohyponym: spoken language corpus. Def.: A written language corpus consists of data material collected from text sources which already exist and often are available in published form. (Gibbon et al. 1997, p. 81)

WWWTranscribe
/'wɜːld 'waɪd 'web træn'skraɪb/, /'w3:ld 'waId 'web tr{n'skraIb/, [N: WWWTranscribe], [plural: none]. Hyperonyms: transcription system. Def.: A transcription system based on the WWW. It is platform independent and allows network access to speech databases. It consists of a number of template HTML files and cgi-scripts written in perl that instantiate the template files with current variable values. Its modular structure makes it flexible, and it connects easily to existing signal processing applications or database management systems.

XML
/'eks 'em 'el/, /'eks 'em 'el/, [N: XML], [plural: none]. Hyperonyms: Standard Generalized Markup Language (SGML). Synonyms: eXtended Markup Language. Def.: XML is a simplified and flexible SGML (Standard Generalized Markup Language, ISO 8879) derivate which many expect to become the standard language for describing WWW documents.

List of abbreviations

ACL	Association of Computational Linguistics
ACT	Advanced Crew Terminal
AI	Artificial Intelligence
AMA	Abstract Muscle Action
ANN	Artificial Neural Network
API	Application Programming Interface
ARPA	Advanced Research Projects Agency
ASL	American Sign Language
ASR	automatic speech recognition
ATIS	Air Travel Information Servic
ATR	Interpreting Telecommuncations Research
AU	Action Unit
AVSP	Audio–Visual Speech Processing
BAS	Bavarian Archive for Speech Signals
BNC	British National Corpus
BRI	Base Rate Interface
CAPI	Common-ISDN-API
CART	Classification And Regression Tree
CES	Corpus Encoding Standard
CGU	Common Ground Unit
CHILDES	Child Language Data Exchange System
CMU	Carnegie Mellon University
CNRS	Centre National de la Recherche Scientifique
COCOSDA	International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques for Speech Input/Output
CORBA	Common Object Request Broker Architecture
COTS	Consumer Off-The-Shelf
CP	(set of) Correct Positive (event hypotheses)
CREA	Corpus de Referencia del Español Actual
CSCS	Corpus of Spoken Contemporary Spanish
CSLU	Centre for Spoken Language Understanding
CTS	Concept-To-Speech
DAMSL	Dialog Act Markup in Several Layers
DARPA	Defense Advanced Research Projects Agency
DAVID	Digital Audio-Visual Integrated Database
DBMS	Database Management System
DCG	Definite Clause Grammars
DECT	Digital Enhanced Cordless Telecommunications
DGI	Defense Group Inc.
DRI	Discourse Resource Initiative
DTD	Document Type Definition
DVD	Digital Versatile Disk
EACL	European chapter of the ACL
EAGLES	Expert Advisory Group on Language Engineering Standards
ELRA	European Language Resources Association
ELSNET	European Network in Language and Speech

EMG	Electromyograph
ESCA	European Speech Communication Association
ESPS	Entropic Signal Processing System
ETSI	European Telecommunications Standards Institute
EU	European Union
F_0	F zero (fundamental frequency)
FACS	Facial Action Coding System
FAP	Facial Animation Parameter
FAPU	Facial Animation Parameter Unit
FAQ	Frequently Asked Questions
FDP	Facial Definition Parameter
FEM	Einite Element Method
FN	(set of) False Negative (event hypotheses)
FP	(set of) False Positive (event hypotheses)
FSM	Finite State Machine
FST	Finite State Transducer
FYI	For Your Information
GI	Generic Identifier
GSM	Global System for Mobile communication
GUI	Graphic User Interface
HCI	human–computer interaction
HLT	Human Language Technology
HMM	Hidden Markov Model
HTK	Hidden Markov Toolkit
HTML	Hypertext Markup Language
HUD	Head-up Display
ICASSP	International Conference on Acoustics, Speech and Signal Processing
ICPhS	International Congress of the Phonetic Sciences
ICSLP	International Conference on Spoken Language Processing
IDS	Institut für deutsche Sprache
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IMMPS	intelligent multimedia presentation system
INTSINT	International Transcription System for Intonation
IPA	International Phonetic Alphabet
IPSK	Institut für Phonetik und Sprachliche Kommunikation
ISDN	Integrated Services Digital Network
ISO	International Organisation for Standardisation
ITU	International Telecommunication Union
JSGF	Java Speech Grammar Format
JSML	Java Speech Markup Language
KIM	Kiel Intonation Model
LDC	Linguistic Data Consortium
LE	Language Engineering
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
LREC	Language Resources and Evaluation Conference

MARCLIF	MAchine-Readable Concept- and Lexicographically oriented Interchange Format
MARSEC	MAchine Readable Spoken English Corpus
MARTIF	Machine-Readable Terminology Interchange Format
MASK	Multimodal-multimedia Automated Service Kiosk
MATE	Multi-Level Annotation Tools Engineering
MIME	Multi-purpose Internet Mail Extension
MIT	Massachusetts Institute of Technology
MPEG	Motion Picture Experts Group
MS-MIN	Multi-State Mutual Information Network
MSE	Mean-Squared Error
NCL	Nominal Complement Clause
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OAA	Open Agent Architecture
OCR	Optical Character Recognition
OGI	Oregon Graduate Institute
OOV	Out-Of-Vocabulary
PCA	Principal Component Analysis
PDA	Personal Digital Assistant
PDF	Portable Document Format
POS	Part Of Speech
PREMO	PResentation Environments for Multimedia Objects
PRI	Primary Rate Interface
PVM	Parallel Virtual Machine
RAID	Rapid Array of Inexpensive Disks
RFC	Request for Comments
RIM	Repair Interval Model
RP	Received Pronunciation
SAM	Speech Assessment Methods
SAMPA	SAM Phonetic Alphabet
SAPI	Speech Application Programming Interface
SEC	Spoken English Corpus
SFS	Speech Filing System
SGML	Standard Generalized Markup Language
SIL	Summer Institute of Linguistics
SL	Spoken Language
SLP	Spoken Language Processing
SMS	Short Message Services
SNHC	Synthetic and Natural Hybrid Coding
SNR	Signal-to-Noise Ratio
SOX	Sound Exchange
SPEX	Speech Expertise Centre
SQL	Standard Query Language
STD	Standard
TCP/IP	Transmission Control Protocol / Internet Protocol
TDNN	Time Delay Neural Network
TEI	Text Encoding Initiative

ToBI	Tones and Break Indices
TOC	Table Of Contents
TSM	Tonetic Stress Marks
TTS	Text-To-Speech
URL	Uniform Resource Locator
VQ	Vector Quantization
WER	Word Error Rate
WHG	Word Hypotheses Graph
WIMP	Windows, Icons, Menus, Pointing
WOZ	Wizard of Oz
WWW	World Wide Web
XML	eXtensible Markup Language

Index

CD–ROM disclaimer