To appear in: Lew Zybatow, ed., *Forum Translationswissenschaft*, Innsbruck

Spoken language lexicography: an integrative framework

Dafydd Gibbon

1 Integrating lexical information

Everybody knows (so they say) that 'the dictionary' is the ultimate repository of wisdom about a language: not only for Scrabble players but also to solve disputes about pronunciation, including prosody, as in the *conTROversy CONtroversy* which flares up from time to time in readers' letters to highbrow British weeklies. Lexicography, the science and technology of lexicon construction, is an extremely heterogeneous discipline, like Applied Linguistics in general. Lexica have applications for human and computational use ranging from 'the dictionary' on the office or living room shelf to written and spoken translation aids and second language learning materials, and are realised in a variety of media from print materials to electronic hyperlexica and lexical databases.

Lexicon applications involve a wide variety of types of lexical information: orthography, phonological structure in terms of speech sounds and word prosody, morphological structure, part-of-speech information, definitions, semantic relations, examples of occurrences in context, and metadata about the origin of the lexical information and conventions for interpretating the lexicon. Lexica are typically oriented towards written language. However, recent developments of lexica for use in spoken language Information and Communication Technology (ICT) systems have motivated the inclusion of interaction oriented spoken and multimodal lexical information: information about tone and body language, i.e. manual and facial gesture, gaze, posture and spatial configuration.

The thesis of this study is that it is possible to integrate these different kinds of lexicon, types of lexical information and lexical structures, and the objective of the study is to outline an approach called the Integrated Lexicon Framework and apply it in the description of a case study.

Before presenting the case study of lexicography in speech-to-speech translation in the Verbmobil project, the background to spoken language lexicography will be discussed. First, a range of perspectives on the lexicon is presented (Section 2), followed by a discussion of a variety of kinds of lexical information which are characteristic of non-written, i.e. spoken and multimodal language (Section 3). Then a model of lexicon types of increasing complexity is presented in order to clarify the problems and stages of creating a lexicon based on corpus data (Section 4). The structure of lexica is introduced (Section 5) from the point of view of a lexicon as a special text type, with its own internal structure, semantics, and physical rendering, and continues (Section 6) with a discussion of lexical structure, with the four components of megastructure (Section 7), macrostructure (Section 8), microstructure (Section 9) and mesostructure (section 10). A special case of a theoretically motivated computational lexicon, the inheritance lexicon, is dealt with (Section 11) before practical issues of lexical implementation are dealt with (Section 12). The case study of spoken language lexicon implementation in Verbmobil (Section 13) is dealt with in some detail, and finally (Section 14) a conclusion and future prospects are discussed.

2 Perspectives on the lexicon

Traditionally, 'the dictionary' on the office or living-room shelf is a consumer article like a

recipe book, and it is typically used as a source of instructions about the 'proper' use of vocabulary. This is particularly true of dictionaries for language learners and of technical dictionaries, which are designed explicitly for prescriptive use. The core of lexicography is descriptive, not prescriptive, however; this is the perspective taken in the present contribution, which attempts to provide a background for a theory of the Integrated Lexicon, in which the aspects of the currently heterogeneous situation can be related to each other on a principled basis.

In the Integrated Lexicon approach (familiarly known as 'ILEX'), a descriptive dictionary is initially defined as a metalinguistic document which enumerates the vocabulary of a language in a structured fashion and has a printed or electronic rendering. This essentially semiotic and universalistic definition of the meaning, structure and form of a lexicon as a document has a background characterised by a set of more far-reaching questions:

1. Can all languages be described lexicographically in the same way?
2. Is it possible to find a principled way of relating the heterogeneous lines of lexicography on the basis of a common ontology in an integrated model?
3. How are practical lexicographic concerns related to linguistic theories of the lexicon?
4. In epistemological terms, can an analogue of the lexicon in be defined in terms of human lexical knowledge?
5. In cognitive theoretic terms, how does linguistic description of lexical information relate to the mental lexicon?
6. More generally, what are the properties of the Universal Lexicon within the theory of Universal Grammar?

These are questions for long-term research; the present contribution addresses a number of points raised by the first three questions. In order to do this, the contribution builds on a number of previously created sources of information[1] (cf. Gibbon et al. 1997, Gibbon et al. 2000, Gibbon 2000, Gibbon & Lüngen 2000). The operational contexts for this work, in addition to computational linguistic work on lexicon theory, were four projects in the *Verbmobil* speech-to-speech consortium and its precursor, the *Architectures for Speech and Language* consortium,[2] a project in the SAM (*Speech Assessment Methodologies*) consortium, two projects in the EAGLES (*Expert Advisory Groups for Language Engineering Systems*) spoken language standards and resources consortia and a consultancy in the ISLE (*International Standards for Language Engineering*) consortium[3], a project in the DoBeS (Dokumentation bedrohter Sprachen) consortium[4], a project in the *Texttechnologische Grundlagen der Informationsmodellierung* basic research consortum[5], a joint curriculum development project[6] with the University of Uyo, Nigeria, and the Université de Cocody/Abidjan, Côte d'Ivoire, and a consultancy in the E-MELD (*Electronic Metastructures for Endangered Languages Data*) project.[7]

3 Spoken and multimodal lexical information

Although 'the dictionary' generally contains information about the segments, syllables and

---

1 If citations of other authors in this contribution are relatively sparse, it is because extensive citations and discussions are included in the earlier studies on which the present contribution is based.

2 Funded by the German Federal Ministry of Education and Research (BMBF), 1991-2000.

3 Funded by the European Commission DG XIII, 1999-2002.

4 Funded by the Volkswagen-Stiftung, 2001-2002.

5 *ModeLeX* project, Universität Bielefeld, funded by the Deutsche Forschungsgemeinschaft (DFG).

6 *M.A. Documentation of Local Languages*, funded by the Deutscher Akademischer Austauschdienst (DAAD).

7 Funded by the National Science Foundation, at Wayne State University, Detroit, Eastern Michigan University, University of Arizona, USA, and University of Melbourne, Australia.

word prosody of spoken language, this information is generally quite minimal. A far wider range of detail and of conventionalised, ritualised and lexicalised communicative events is associated with spoken language than is usually found in 'the dictionary'. But, first of all, why should spoken language and multimodal information be at all important for our text-oriented civilisation, where writing is our criterion for objective knowledge, the medium of our laws and most of our historical records?

The need for these kinds of information is motivated by general linguistic considerations but is also driven by requirements of modern interactive ICT systems. The fact is that most of the world's 6000 or so languages are unwritten. Further, everyday communication in both written and unwritten languages is spontaneous, transient and multimodal, combining spoken and gestural modalities, whether face-to-face or via modern voice telecommunication. Even lack of a visual channel does not inhibit gesture on the telephone. The internet and text messaging certainly encourage the use of writing - and members of unwritten language communities or communities with non-alphabetic scripts have been quick to apply English-based orthography conventions to writing their own languages. But writing also has its limits in scenarios where fast, direct, personal contact is necessary. In this domain, voice support by means of Information and Communication Technology (ICT) is quickly developing for both global and local languages in many advanced industrial and service communicative activities, as well as for basic health and agricultural information services to remote rural areas of the world in pre-literate societies.

These technologies require access to lexical information about spoken and multimodal language. By 'spoken and multimodal language' (often referred to just as 'speech') is meant language as used in face-to-face oral-acoustic-auditory and gestural-visual-optical communication. The term 'multimodal' refers to the use of more than one parallel communication modality. The term 'modality' means a pair of human motor-sensory output-input devices (e.g. oral-acoustic; manual-visual). Multimodal communication may be *face-to-face*, using speech and gesture or (as in a lecture, or in dictation software) speech and text, or *teleglossic*, via writing or electronic media. The term 'submodality' has been introduced to mean parallel modulations of the same modality, such as the locution and prosody submodalities in speech, or the text and graphics submodalities in writing, or relatively independent facial, manual and body gesture systems (cf. Gibbon, Kölsch, Mertins, Schulte & Trippel 1999, Gibbon, Moore & Winski 1997, Gibbon, Mertins & Moore 2000).

A number of lexicographically relevant issues with regard to kinds of lexical object and types of lexical information are raised by including spoken and multimodal items in the lexicon. Here are some examples of such lexical objects and their properties.

1. Forms of oral communication (speech) in relation to visual communication (gesture, writing), including the segmentation of these forms in speech and writing (e.g. spoken and written syllabification, stress assignment), prosody (Bleiching 1992, Gibbon 2002b) and gesture representation (Gibbon et al. 2004).

2. Referentially light nouns with vague referents, such as *whatsit*, *thingummy-jig*, *doodah*, ... and nonce formations (often supplemented or replaced by partly idiosyncratic paralinguistic hisses, tongue clicks finger-clicking and other gestures). Light nouns are used in time-critical situations as surrogates for nouns which cannot be retrieved from the mental lexicon in real time, or which may not exist for a concept meant by the speaker. Light nouns are characteristic of informal situations in which the speaker is in a hurry; there may also be other gender-specific and age-specific sociolinguistic constraints associated with this class of words.

3. Discourse particles of various kinds, including single-word greetings and disfluency markers: *well*, *erm*, *er*, *aha*, *hi*, *bye*, *cheers*, and word fragments (interrupted words) which are often associated with disfluency markers (Tseng 1999, Fischer 2000) and gestures.

4. Features of the 'restricted codes' of everyday speech, including pronouns whose

referents can only be resolved by knowledge of extra-communicative situation features, elliptical speech, indirect speech acts, routine expressions, from proverbs and idioms to temporary or long-lasting 'private vocabularies' within families or other small but intensively interacting groups. Proverbs, sayings, idioms and other kinds of fixed expression which typically characterise interactive use of spoken and multimodal language also figure in this category.

5. Taboo nouns and verbs with appraisive, emotive and generally insulting connotations, including the notorious '(near) four-letter words' in English (with analogues in other languages) such as *arse/ass*, *crap*, *cunt*, *frig*, *fuck*, *prat*, *prick*, *shit*, *twat*, *wank*, and others (sometimes derived from the four-letter words) such as *bugger*, *crappy*, *frigging*, *fucker*, *fucking*, *shitty*, *wanker*, *wanking*. Some are associated with characteristic gestures. Words of this class are used in sociolinguistically highly specific informal situations, mostly by males to underline stereotypical masculine attitudes, are regarded as highly impolite or insulting, and are frowned on in formal situations and in polite circles, and hence also in writing.

6. Specific words, word classes, grammatical constructions have different frequencies in spoken and written language: first and second person pronouns, pragmatic idioms, interrogatives (including tag questions) and imperative constructions.

7. Gestural enhancements of speech (not considering special sign languages for acoustically deprived or hostile environments), including gestural idioms, some of which are iconic, others indexical: waving, gestures for success/failure, insults, eating, drinking, shapes, sizes, pointers to objects in the communication environment (which are even used in non-multimodal communication on the telephone).

8. Multiple interconnected layers of temporally parallel lexical information (Witt, Lüngen & Gibbon 2000), encompassing the locutionary, prosodic, paralinguistic and gestural information which characterise multimodal communication by speech, gesture, and in some contexts, e.g. classroom teaching, also simultaneous writing.

All of these features are perfectly normal in varieties of spontaneous speech, and all share the key problem of obtaining authenticity, in the sense of attestation in corpora. Traditional introspective methods in linguistics are highly biassed towards idiolect and personal experience, which is unsuitable as an empirical basis, particularly for spoken language and multimodal lexicography. Modern varieties of 'the dictionary' are associated with very large authentic corpora, and with spoken and multimodal language lexica the authenticity requirement is even stronger, as custom-made transcribed and annotated audio-visual material is required: spoken and multimodal language data are not *objets trouvés*, as in written corpora. Ideally there will also be a route from the dictionary back to the corpus contexts which were used in creating the entry, i.e. authentic examples in the dictionary entries themselves; for spoken and multimodal language lexica this requires special software, an audio-visual concordance (Gibbon et al. 2001, Gibbon & Trippel 2002). Example citations in written language dictionary entries are simple cases of corpus linking, and the translation memory databases used in professional large-scale translation applications are complex cases.

ICT systems have particularly stringent requirements: a computer system cannot use common sense intuition about taboo items, inexplicitness, and gestural meaning. An ICT system requires very precisely identified and defined exemplars of corpus occurrences of anything it might encounter. The occurrences must be in sufficient quantity to be able to induce prototypical statistical models of the speech signal aligned with units of spoken language such as words; the study of gesture is only gradually reaching a stage where gestural lexical items can be recognised automatically. For this purpose, recorded speech signals have to be annotated with lexical items by attaching time-stamps to the items. This applies both to the 'talking dictionary' with custom-recorded citation forms of words, a conservative extension of the classical dictionary, and to the pronunciation models of speech recognition and speech

synthesis ICT systems.

4 From corpus to lexicon: an integrative multi-layer model

Lexical information is obtained from examining speech and text corpora; this principle is valid for all large modern lexica. In the Integrated Lexicon approach, a structured and ranked scale of abstraction from corpus to lexicon is defined in order to clarify the relation between corpus data on the one hand, and lexical descriptions on the other.. The major distinction is between corpus information and lexical information, further distinguising primary, secondary and tertiary corpus information, and four orders of lexical information abstraction. Figure 1 shows the overall structure of this abstraction-based generic lexicon typology.

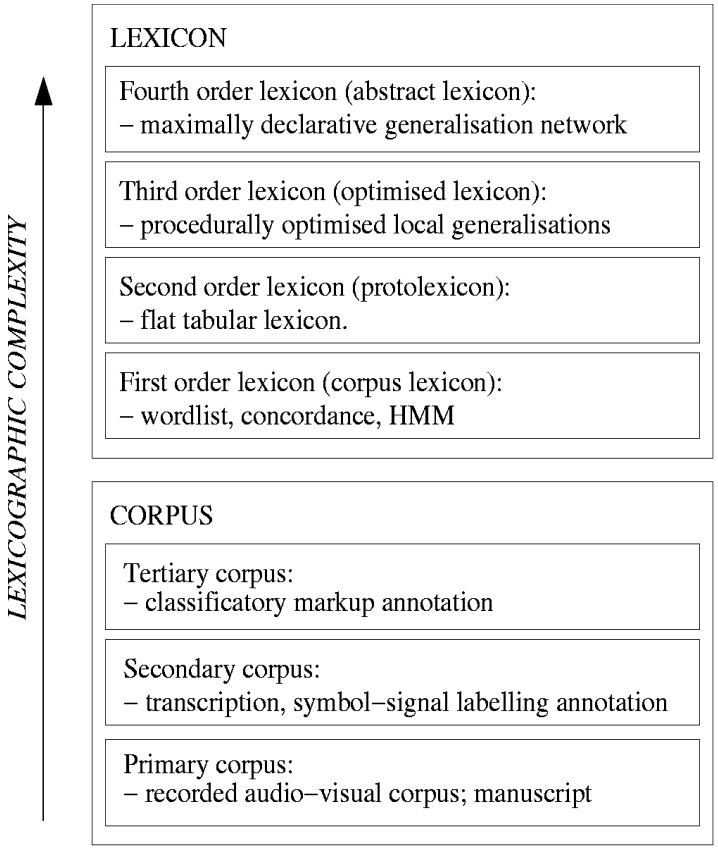

Figure 1: Layers in a generic lexicon typology.

Spoken and multimodal language corpora have three layers, defined by degrees of abstraction from primary utterance and text data:

C1. Primary corpus information: The signal level of an audio-visual recording, or a handwritten or printed manuscript, with metadata on the situation of recording (participants, equipment, scenario) and the format of the recording. Legacy written data in electronic form may also be regarded as primary corpus information.

C2. Secondary corpus information: The symbolic digital level of phonological, phonetic, orthographic or other transcription of the audio-visual recording in an accepted and well-defined notation, and an annotation of the audio-visual recording in the form of a mapping between segmented tokens in the transcription and temporal intervals in the audio-visual recording, with metadata on the theoretical assumptions (e.g. as an ontology of descriptive

categories) which underlie even transcriptions and annotations. The analogue for written corpora is somewhat different. If a C1 corpus exists, the analogue is a comparable OCR scan with an electronic text version and appropriate annotations. If the text is already in a well-defined electronic version, the C1 level is vacuous, and effectively identical with the transcription at the C2 level.

C3. Tertiary corpus information: The annotation of the tokens in a segmented corpus with classificatory linguistic information (morphological, syntactic and semantic tagging, treebank construction), which presupposes extensive prior linguistic analysis. (The boundary between orders C2 and C3 is fuzzy as C2 also presupposes at least some phonological analysis.)

The corpus is segmented into tokens which are classified into lexical types at four different orders of lexical abstraction.

L1.The first order of lexical abstraction from the corpus is the form-based *corpus lexicon* which summarises the *types* (i.e. the forms) of unit whose *tokens* occur in the corpus. In the case of words, the types are fully inflected forms. The corpus lexicon may be simply a wordlist of these types. A basic KWIC (KeyWord In Context) concordance is a corpus lexicon consisting of a list of types, each with a list of the contexts in the corpus in which each token occurs. Coverage and type-token statistics may also be included; the statistical models of language units which are used in the Human Language Technologies, such as Hidden Markov Models (HMMs) for word recognition, are essentially more sophisticated versions of a form-based corpus lexicon. The overall structure, the macrostructure, of a dictionary of this kind is simply a set of entries, perhaps ordered in some way, such as alphabetically. The structure of the lexical entries themselves, the microstructure, is also very simple: a list or vector of basic information, such as word frequency, type token ratio, concordance context, statistics of cooccurrence in context. The overall structure of macrostructure and microstructure can be represented as a table and stored as a simple database relation.

L2.The second order of lexical abstraction is the *protolexicon* which pairs lexical types with other linguistic analyses and interpretations of more complex other kinds according to a semiotic model of form, structure and meaning, and may add out-of-corpus vocabulary (sometimes known just as out-of-vocabulary items). This additional lexical information includes the lexical lemma or lexeme, the morphological class and structure (e.g. simplex, derived, compound, inflected), syntactic category (POS, part of speech) and subcategory, definitions, and semantic relations with other lexical types, such as synonymy and antonymy. A KWIC concordance which accesses this kind of information is a form of protolexicon. The microstructure of a protolexicon is essentially an extension of the microstructure of the corpus lexicon. Lexical ambiguities (polysemy, homonymy, inflexional syncretism, alternative pronunciations) may be treated as unrelated items; in such cases the list of entries will be huge. For example, *either*, pronounced /aɪðə/ or /iːðə/, would then be two separate entries, and all possible meanings of "fast" (as in *fast boat*, *fast reader*, ...) would be listed separately (the decision to separate ambiguous forms into separate entries is a heuristic one, since it is more than doubtful whether all polysemous shades of meaning can be enumerated). The protolexicon is not organised from any particular procedural perspective, and contains no generalisations about the lexical object language, i.e. about sets of lexical entries or relations between them (definitional generalisations about the lexical metalanguage for transcription, parts of speech etc., are necessary). Essentially, a protolexicon is a collection of elementary lexical facts which are partly corpus-based, but which are only useful to the lexicographer and are not of much interest to other categories of dictionary user. The contemporary version of the traditional card-index protolexicon is the lexicographer's lexical database.

L3.The third order of lexical abstraction is the *optimised lexicon* in which some

generalisations, including lemma abstraction, are made, based on selected procedural criteria of access to lexical information. Examples of such generalisations are summaries of alternative pronunciations or of polysemous meanings, or the inclusion of related derived and compound or synonymous words in one entry. A procedurally optimised lexicon may still have a simple macrostructure in the form of a list or set, but the microstructure of the entries will be more complex, and contain shallow hierarchies of information options. The two most well known kinds of procedural optimisation for different functionalities are the (somewhat outdated) *semasiological* perspective, in which the macrostructure and microstructure of a lexicon are optimised in order to find the meanings of known lexical forms (the form-based decoding lexicon or reader's lexicon), and the *onomasiological perspective*, in which the macrostructure and microstructure of a lexicon are optimised in order to find the lexical forms associated with known meanings (the function-based encoding lexicon or writer's lexicon) as in a hierarchically organised thesaurus. Dictionaries or lexical databases organised for a particular functionality may still contain exactly the same declarative information, but procedurally optimised for this functionality.

L4.The fourth order of lexical abstraction is the *abstract lexicon* in which a maximum of theoretically motivated generalisations of all kinds is incorporated into the lexicon, each lexical entry is minimally specified under an *abstract lemma*, and inherits generalised information from a hierarchy of lexical classes. In this kind of lexicon, each data category (type of lexical information) in the microstructure is associated with a hierarchy of generalisations which link lexical entries with similar properties. For example, a semantic hierarchy of lexical entries may constitute a classificatory taxonomy from the general ('entity') via intermediate levels (e.g. 'animal', 'mammal', 'dog') to the specific ('poodle', ...) or a compositional meronymy of whole units ('house') and their parts ('roof', 'storey', 'room', 'corridor', ...), or be based on other kinds of lexical relation (Cruse 1986). Lexica of this kind are typically found in linguistic theories of the lexicon, and are purely declarative; the best-known example of a linguistic theory with a hierarchical system of lexical category types is *Head-driven Phrase Structure Grammar*, HPSG (cf. the introduction in Sag & Wasow 1999). In computational linguistics many theoretically well-founded practical applications with declarative hierarchical lexica have been developed, such as the lexicon representation language DATR (Evans & Gazdar 1996; see also Gibbon 1991, 2002a). The fourth level of abstraction is procedurally neutral, subject only to logical inference types.

Consumer lexica are third order lexica; the microstructure are typically unevenly weighted in that one type of lexical information is used as a search key and constitutes the headword of the entry; the rest of the entry constitutes the article describing the other lexical properties of this headword. In 'the dictionary' this would be the orthography of a canonical morphological form such as nominative singular for nouns, or the infinitive for verbs; this applies to English and other European languages; for other morphological language types other criteria may be used. The canonical form is used as a *headword* for matching with search keys, and the rest of the lexical entry is formulated systematically as an *article* about this headword. Lexical databases for the lexicographer's workbench, for archiving, and for dissemination to consumers may be very differently organised.

In order to approach the Integrated Lexicon theory more closely, a generic document-theoretic approach to the lexicon is taken. The increasing convergence of lexicon-theoretic and practical lexicographic issues through the use of efficient computational lexicographic methods has resulted in a situation today in which theoretical underpinnings are needed in order to develop complex computational tools for lexicography. The theoretically motivated fourth order lexicon is used as a source of information for third order lexicon products.

5 A document theoretic approach to the lexicon

A descriptive dictionary was defined above as a metalinguistic document which enumerates the vocabulary of a language in a structured fashion and has a printed or electronic rendering. In this definition, a document is presented as a complex semiotic object or sign, with a meaning, a form and a structure. Documents are complex signs within the domain of disciplines such as text linguistics and text technology; a lexicon document is a special case of complex sign in its own right, composed of smaller signs such as lexical entries, with the same kinds of defining property as other kinds of document.

For the purposes of Integrated Lexicon theory, a sign is defined as an abstract semiotic object with three main properties of meaning, form and structure. These properties are not of the same kind: the structural properties of the sign are distinguished from the two interpretative properties of form and meaning, both of which which relate the sign to reality: form to the acoustic or visual rendering of the sign and meaning to the denotation of the sign. It is these two interpretative properties which constitute the semiotic core of the sign: the relation between form and meaning. This sign model applies to large and complex semiotic objects, such as documents, just as it applies to smaller semiotic objects, such as sentences or words. The two kinds of interpretative property, for example, correspond to the functions of semantic interpretation and phonetic interpretation in mainstream linguistics; documents are simply complex signs which are within the domain of text linguistics and text technology, not the linguistics of sentences and words.

The form of a lexicon document is the implementation of the document with a printed or electronic rendering. The meaning of a lexicon document is the vocabulary of a language (or an excerpt from the vocabulary); a lexicon document is evidently a metalinguistic document, because it denotes the vocabulary of a language, unlike documents which have non-linguistic content. The structure of a lexicon document concerns the organisation of lexical entries, of lexical information about these entries, and of metadata about the lexicon and its components. The structure of lexicon documents is the topic of this section.

The main semiotic properties of signs, including lexicon documents as complex signs, can be described by means of a model with a compositionality dimension of external and internal structure, and an interpretative dimension relating the sign to reality in two domains, the content and the form modality. The pragmatic properties of signs are not represented in the model; the entire model is to be understood as being embedded in a pragmatic situational structure. The sign model is visualised in Figure 2 and described below.

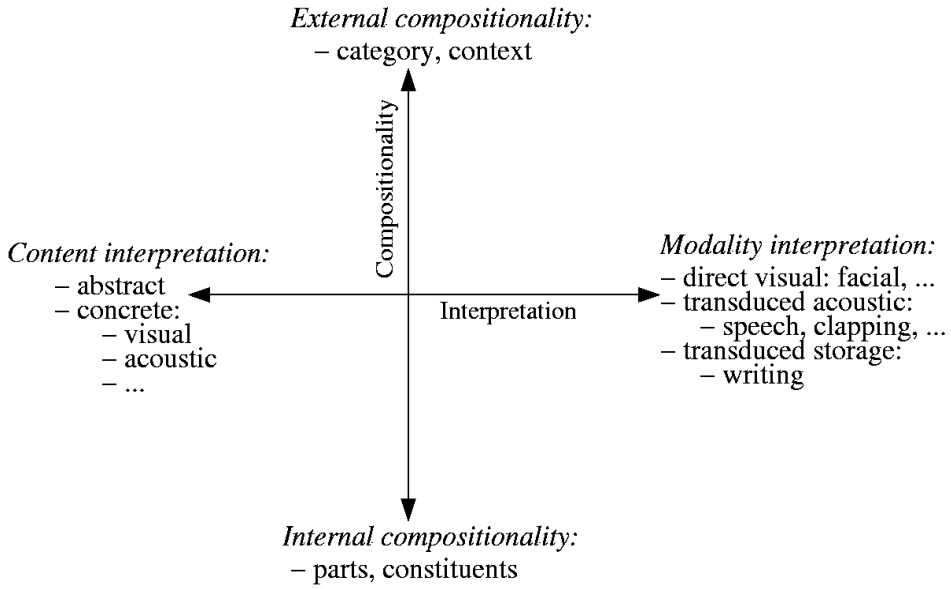

Figure 2: Structural and interpretative dimensions of signs.

1.External compositionality: the structural context in which the sign occurs. A lexicon document occurs within a larger context of lexicographic objects such as corpora and databases; components of a lexicon document, such as lexical entries, occur within the larger context of the lexicon itself. The objects denoted by lexical documents and their components are themselves signs; a lexical entry may denote a word, for example, which is itself a sign with form, meaning and structure, and whose external structure is constituted by its syntagmatic relations within phrases and sentences.

2.Internal compositionality: a lexical document has internal structure, i.e. parts, components, constituents, which are systematically related. The internal structure of a lexical document and its parts is defined at different levels called the *megastructure*, the *macrostructure*, the *microstructure* and the *mesostructure*, of the lexical document. These structures will be dealt with below. Other signs have internal structures: sentences and phrases, including idioms which are lexicalised sentences and phrases, have phrases and words; words have stems and morphemes.

The interpretative properties of signs correspond to the two interpretative components of mainstream grammars, semantic interpretation and phonetic interpretation, except that phonetic interpretation is generalised to modality interpretation in order to cover the manual-visual (writing) and gestural-visual modalities:

1.Content interpretation: the relation of the sign to the reality which it denotes (e.g. in terms of definitions of objects, properties of objects, relations between objects). In the case of the lexicon and its components, this reality is the vocabulary of a language (or an excerpt from the vocabulary). The content of the lexicon ranges from general domains to restricted semantic word-fields, technical domains as in glossaries of technical terminology, and specialised lexica such as pronunciation lexica.

2.Modality interpretation: the relation of the sign to the reality in which it is realised (e.g. acoustically in terms of pronunciation, or visually in terms of orthography or gestures). The modality interpretation is the implementation of a lexicon document in some accessible medium-specific format, such as print, a relational database (Gibbon 2000), or an inheritance network (Gibbon 2002a), with a 'user interface' in the form of

1.a working format for the lexicographer (Gibbon & Lüngen 2000),

2.an archival format whose structure is marked up in XML (eXtensible Markup Language), or

3.a dissemination format such as a book or an electronic hyperlexicon.

The structural and interpretative types of information constitute multidimensional semiotic space in which a lexicon document and its components, such as lexical entries, are located as complex metalinguistic signs.

6 Lexicon structure

In the following sections, the four organisational principles of lexical documents are discussed: megastructure, macrostructure, microstructure and mesostructure; a fifth (numbered 2 in the Figure) is the metainformation which is also contained in the megastructure. The architecture is shown in Figure 3.

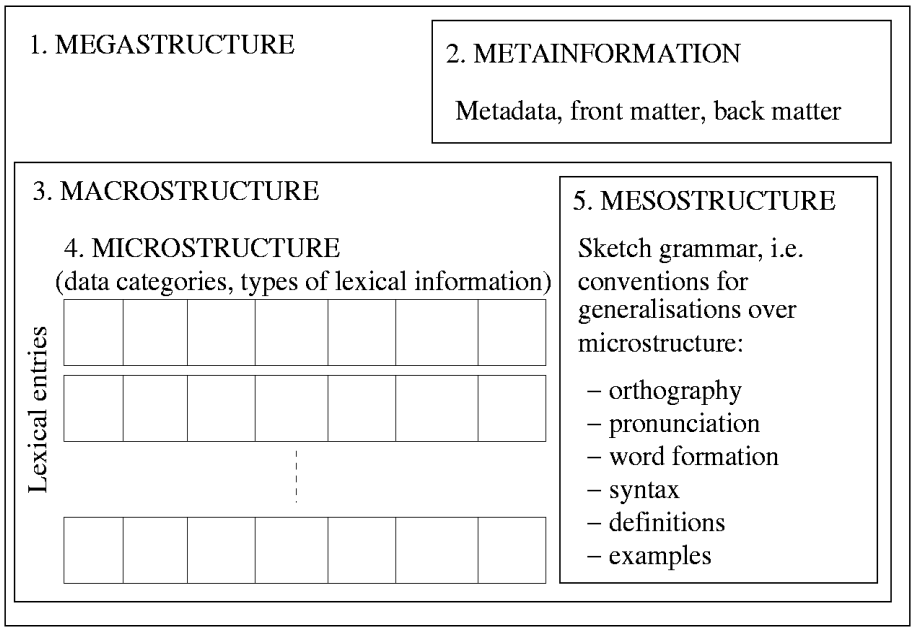

Figure 3: The four organisational components of a lexicon: megastructure, macrostructure, microstructure and mesostructure.

1. Megastructure is the overall structure of a lexicon, including general metainformation about the lexicon and its components, as well as the macrostructure, the microstructure and the mesostructure.
2. Metainformation specifies general information about the origin and format of the lexicon, including formal metadata for archiving purposes, publisher's information, the other front matter and back matter of a dictionary, and the body of the lexicon containing the lexical entries.
3. Macrostructure is the organisation of the body of a lexicon which includes the lexical entries and lexical information, and generalisations about lexical information. The main mode of macrostructure organisation in third order consumer dictionaries, for example, is semasiological. This is the standard mode of organisation of 'the dictionary', in which entries are ordered by headwords according to their canonical form (generally orthographic), which is used as a search key, and the information accessed is the meaning. Two important kinds of macrostructure are the onomasiological and semasiological optimisation principles which have already been introduced. But there are many other kinds of macrostructure, as in bilingual dictionaries, or pronunciation dictionaries.
4. Microstructure is the internal structure of a lexical entry, which lists the data categories of the types of lexical information associated with the lexical entry. The data categories are the dimensions of the semiotic space in which the language entity described by the lexical entry, such as the word or idiom, is located, e.g. orthography, pronunciation, etymology, parts of speech, morphological categories, definitions, cross-references, contexts of use, and glosses in another language. There are many kinds of microstructure; in a thesaurus the microstructure is typically just a set of near-synonyms; in 'the dictionary' the dictionary is typically lexical information ordered according to data categories. The microstructure may incorporate mesostructural elements such as generalisations over polysemous sub-entries, or sub-entries for derived or compound words which include the headword.
5. Mesostructure is a system of generalisations about classes of lexical entries based on shared microstructure properties, including generalisations about orthography and

pronunciation, word structure (inflected, compound, derived or simple) or parts of speech contained in sketch grammars, cross-references to other lexical entries, and references to corpus sources. Some mesostructural generalisations contained in definitions, abbreviation conventions and the like, is often taken to be part of the metainformation, but the latter term is best reserved for formal types of metainformation (origin, format, etc.) which are not specific to the lexical object language.

Most lexicographic work is concerned with defining microstructure and data categories standing for the types of lexical information which are organised by microstructure data categories. Part of the typical microstructure of 'the dictionary' is shown, simplified slightly from the Longmans Dictionary of Contemporary English (LDOCE, Procter 1978).

> **an·i·mal** /ˈænɪməl/ *n* **1** a living creature, not a plant, that has senses and is able to move itself when it wants to: *Snakes, fish, and birds are all animals.* | *Humans are the most intelligent of all the animals.* | *Man is a political animal.* **2** all this group except human beings: *farm animals* | *Should animals be kept in cages?* **3** a MAMMAL **4** a person considered as behaving like a wild non-human ...

The lexical entry denotes the word 'animal', and has a complex rendering format which consistently reflects the structure of the entry, i.e. the microstructure of the lexicon. The headword doubles as the orthographic representation of the canonical inflexional form of the word (for English nouns: non-plural, non-genitive), showing hyphenation breaks. The next elements show the phonemic represantation of the pronunciation, with primary stress, in IPA notation, followed by the part of speech, noun. The generalisations expressed by the orthographic, phonemic and categorial information are described implicitly in a mesostructure embedded in the front matter of the dictionary. The definitional parts of the microstructure are divided into four numbered polysemous readings, each of which consists of a definition by *genus proximum et differentiae specificae* and a set of examples of use in a phrasal context, the members of which are separated by a vertical bar. The third reading is a microstructural cross-reference to a synonym of the reading, represented in small capitals.

The following sections treat the four structural components in more detail.

7 Megastructure and metadata

The first use of the term 'megastructure' for dictionaries has been attributed to Hartmann (1983) in connection with printed dictionaries, and covers the body of the lexicon proper together with ancillary information which is often called the front matter (e.g. publisher's information, editorial material, preface, introduction, sketch grammar, list of abbreviations), and the back matter (e.g. appendices various kinds). As it has been used previously, the term applies to dictionaries in conventional print medium renderings, but can be generalised unproblematically to include electronic lexical repositors, including lexical databases and hypertext lexica (*hyperlexica*) on the web or CD-ROM (Gibbon & Trippel 2000, Gibbon, Trippel & Sasaki 2004). In the Integrated Lexicon context, the megastructure is defined as containing the macrostructure of the lexicon and the lexical metadata (which in turn can be complex).

What kind of megastructure does a lexical database for spoken language have in ICT contexts? The terminology used in ICT contexts is often somewhat different from that used in mainstream lexicography. Depending on the level of abstraction, various terms such as 'database documentation', 'file header information', 'library catalogue information' are found. These notions have the generic label of metadata, and are sometimes referred to in a more homely fashion as housekeeping information. Using the concept of metadata, the definition of megastructure can be generalised in a highly convenient and topical fashion: lexical metadata provide formal information about lexical data, regarding the content of a lexicon as data from a practical computational point of view (rather than metalinguistic information about the

language itself):

1. publishing and editorial history (author, date, place, etc. - i.e. library catalogue data),
2. languages (lexicographic metalanguage, object source language, object target language for bilingual lexica, controlled language),
3. organisational principles,
4. domain, coverage,
5. formatting conventions,
6. abbreviations,
7. user rights and obligations.

In short, the metadata may cover any information necessary for interpreting the lexical information in the dictionary, lexicon, or lexical database and the conditions of its use. In the case of spoken language lexica, additional information about recording conventions such as the signal characteristics (stereo/mono, sampling rate, amplitude resolution) and the encoding format (WAV, WMA, MP3, OGG etc.) of the recorded corpus will be included.

It is only during the past few years that the study of metadata has come into its own as part of a specific information and library science discipline. The reason for this is clear: the internet is a universal archive access mechanism, providing sophisticated generic strategies for information retrieval from a straightforwardly structured network, using sophisticated string search techniques, but supplementing these with very specific metadata to support cataloguing of the data and keyword based information retrieval. What applies to texts on the internet in general applies even more to lexical information on the internet, i.e. information which is designed to be accessed in a structured and systematic fashion, unlike free text, graphical or audio information. A number of metadata conventions have been established for text documents in general, which are also applicable at a highly generic level to lexica. The main convention for highly generic metadata is the *Dublin Core* (*DC*) metadata set of 15 categories: Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type.

An extension which was specifically designed for linguistic documents such as dictionaries is the *Open Language Archive Community* (*OLAC*) metadata set. The OLAC metadata subset[8] for the general linguistic data type *Lexicon* is reproduced in tabular form in Table 1.

Table 1: OLAC Linguistic Data Type Vocabulary.

Name	*Lexicon*
Definition	The resource includes a systematic listing of lexical items.
Comments	Lexicon may be used to describe any resource which includes a systematic listing of lexical items. Each lexical item may, but need not, be accompanied by a definition, a description of the referent (in the case of proper names), or an indication of the item's semantic relationship to other lexical items.
Examples	Examples include word lists (including comparative word lists), thesauri, wordnets, framenets, and dictionaries, including specialized dictionaries such as bilingual and multilingual dictionaries, dictionaries of terminology, and dictionaries of proper names. Non-word-basedexamples include phrasal lexicons and lexicons of intonational tunes.

The OLAC lexicon metadata terminology includes references to spoken and multimodal language items, but more explicit listing of subtypes is needed, both inside and outside the spoken and multimodal language areas: wordlist, glossary, and concordance types, as well as cross-classifying parameters such as print vs. machine-readable and available media (partly covered by the DC metadata category set), and the modality of the object languages. Even

[8] Source: <www.language-archives.org>.

more important is information about the organisational principle of the lexicon: semasiological vs. onomasiological vs. form-based (as in many bilingual lexica and other specialised types such as pronouncing dictionaries). This area can evidently bear more research, and the point will be taken up from a more theoretical point of view below.

8 Macrostructure and lexical entries

Lexicon macrostructure is the information structure into which lexical entries, their microstructures and mesostructural generalisations (if any) are embedded. The main representatives of the traditional, procedurally optimised macrostructures are the onomasiological and semasiological macrostructure types. Other kinds of macrostructure are the various kinds of network organisation to be found in semantic networks, wordnet lexica, and inheritance lexica, including framenet lexica. In many traditional semasiological dictionaries which contain sub-entries, the macrostructure intrudes into the microstructure (or vice-versa): sub-entries are at the same time lexical objects which are organised by the macrostructure, and types of lexical information which are organised by the microstructures.

In the Integrated Lexicon framework, the basic macrostructure for fourth order lexica is simply a set of attribute-value structures for abstract lemmata and lexical classes, ordered in terms of the implication relations characteristic of inheritance lexica. If the lexical entries are fully specified with their entire complements of lexical information (an ideal assumption - this is never the case in any but the most trivial lexica), then there is no other organisation except that the entries are in a set. In this basic macrostructure, the microstructure is simply a feature structure, there are no generalisations, optionalities, no cross-references between lexical objects, and no generalisations over categories of lexical objects.

Equivalently, in this basic macrostructure, the macrostructure and the microstructure, taken together, constitute a table, in which an arbitrary ordering is imposed on the lexical objects in the macrostructure, and on the types of lexical information in the microstructure: the columns specify the types of lexical information in the microstructure, and the rows represent the microstructures of the particular lexical objects.

In practice, many simple varieties of lexicon have just this kind of basic macrostructure: pronunciation lexica and glossaries are two obvious examples. For a generic and integrative approach, however, this basic macrostructure concept is too primitive.

An important concept which pertains to macrostructure is the coverage of the lexicon. This is a not uncontroversial notion (cf. the discussion in Landau 1983). A useful distinction between types of coverage was introduced in the projects which provided the background for the development of the Integrated Lexicon framework:

1. *extensional coverage*: the number of lexical entries;
2. *intensional coverage*: the number of data categories and their values which are provided for the lexical information.

Despite problems in defining coverage, it is still a basic criterion for evaluation lexica of all kinds, relative of course to the type of lexicon: a pronunciation lexicon obviously cannot be compared with a thesaurus.

9 Microstructure and lexical data categories

The lexical information associated with a lexical entry is organised in the microstructure of the lexical entry; the microstructure, in the general case, is a list of data categories representing types of lexical information such as orthography, pronunciation, part of speech, definition. Like the lexicon document as a whole, as a complex semiotic object, each lexical entry is a sign type; the microstructure can consequently be ordered in terms of groups of data

categories corresponding to the structural and interpretative dimensions of the sign. This view is fundamentally in accord with traditional approaches in lexicography, but it is also in accord with current theoretical views of the lexicon as a part of lingusitic theory, particularly in the HPSG paradigm.

A semasiological lexical entry in a traditional third order lexicon is the word, whether simplex, derived or compound. In a third order optimised lexicon, a lexical entry is identified by a headword, represented basically by the orthography of a canonical inflexional form (a conflation of two functions: headword and a type of lexical information). The semiotic properties of the entry, the types of lexical information, are represented in the data categories of the microstructure of the lexical entry. The microstructure is described in an article associated with the headword.

In fourth order lexica in the Integrated Lexicon framework, a lexical object is either a lexical entry (known in this context as an abstract lemma) or a class of lexical entries. The lexical objects have either an arbitrary 'headword' (a unique code, for instance a number) or no headword at all. The headword is, declaratively speaking, redundant, since every lexical entry is uniquely characterised by the structural and interpretative properties of the lexical entry: when the headword function is conflated with the orthography data category, as in 'the dictionary' which we know and love, the headword is evidently part of a third order procedural optimisation for semasiological lexical lookup.

A major difference between a lexical entry in the Integrated Lexicon framework and the lexical entries in conventional lexica is that the lexical object may not necessarily be a word: a lexical object may be any unit of language which is stereotyped, ritualised, fixed and inventarised - in other words, lexicalised. So an abstract lemma may be a word, of course, but also a larger unit such as an idiom or a proverb, or even larger units of ritualised communication such as greeting dialogues, liturgies, performable fixed texts, or a smaller unit such as a morpheme, a morphosyntactic tone (as in many African languages, for example), an intonational contour or tone sequence (in all languages), an iconic gesture, or even a phoneme (whose 'distinctive meaning' is rather minimal, as meanings go). In the context of practical lexicography, this generous view of abstract lemmata is likely to lead to some controversy. But there are indeed 'dictionaries' of proverbs, quotations, and the like, and inventories of all kinds of other units of language may go under the name of dictionary: this is not just a fortuitous or metaphorical usage, or based on a family resemblance, but a consequence of a common view of lexical entries as semiotic units.

The article associated with the abstract lemma lists the properties of the abstract lemma and the relations into which the abstract lemma enters with other objects (other abstract lemmata as well as classes) in the lexicon. For the representation of the properties attribute-value structures are used (sometimes also called feature structures, or attribute-value matrices). Attribute-value structures organise the set of properties into subsets of mutually exclusive properties such as *noun*, *verb* or *animate*, *inanimate*.

The structural properties of an abstract lemma define two levels of structure:

1. Internal structure (endotaxis), for example the morphological structure of a word or the syntactic structure of an idiom. The internal structure of a word determines its internal composition based on morphological rules which define inflexional patterns, derivational patterns and compounding patterns.
2. external structure (exotaxis), i.e. its distribution in larger contexts, including its modifer, complement, argument and role structures, but also in principle situation-dependent usage constraints. The external structure of a lexical object such as a word thus determines its category as a part of speech such as *noun* or *ditransitive verb* on the basis of the role of the word in context.

The lexical entry is a complex metalinguistic sign which denotes linguistic sign such as a

word, an idiom, a morpheme, etc. (depending on the domain of the lexicon as a whole). As a sign, the lexical entry has the same kind of semiotic structure which other signs have. The lexicon document is a complex metalinguistic sign, as already noted, with various components, including lexical entry signs. The microstructure of a lexical entry defines the semiotic space in which the denotation of the lexical entry is located. The position in semiotic space of a particular linguistic lexical object, in this case a simple one which is characteristic of spoken language and uncharacteristic of most registers of written language is illustrated here: the lexical entry for the interjection "Hi!", defined as an attribute-value matrix (Figure 4).

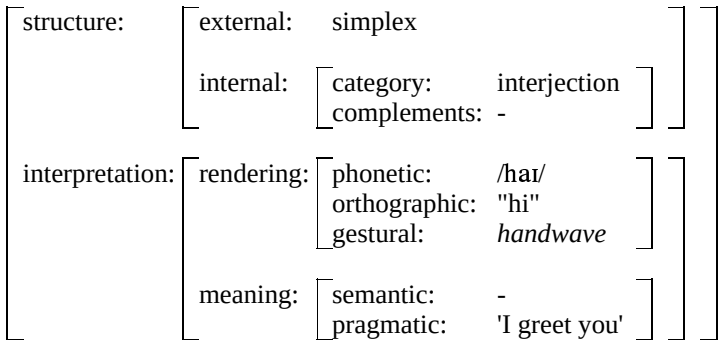

Figure 4: Attribute-value representation for a simple lexicon microstructure.

10 Mesostructure and lexical generalisations

Far from being simply repositories of idiosyncratic information, in dictionaries of all kinds lexical generalisations of many different varieties are to be found. The main kinds of ubiquitous generalisation are expressed as follows:

1. Lexical metalanguage generalisations:
    1. *character encoding* of lexical entries (characters standing for standard orthographies, syllabaries, logographic sets, pronunciations; fonts and font highlights standing for particular types of lexical information),
    2. *abbreviations* (e.g. for parts of speech, styles).
2. Lexical object language generalisations:
    1. *linguistically significant generalisations*:
        1. *classification*, i.e. assignment of lexical entries to classes based on similarity (paradigmatic relations),
        2. *composition*, i.e. role in a structure (syntagmatic relations).
    2. *cross-references* (expressing dependencies between lexical entries and other lexical entries such as co-hyponyms, synonyms, antonyms),
    3. *illustrations*, including examples, citations and graphical illustrations.

All of these generalisations constitute the mesostructure of a lexicon: the common feature of devices for expressing these kinds of generalisation is that they are all relations between objects, whether lexical entries, or classes of lexical entries, or definitions of properties of lexical entries. The conventions for expressing these generalisations are often contained in the metainformation in the megastructure of a lexicon, in traditional dictionaries for example in the front matter.

Classificatory mesostructural generalisations over lexical objects express paradigmatic relations, i.e. relations of similarity and difference, which are used to define sets of lexical objects. These sets can overlap in many ways: paradigmatic relations of inclusion involve

taxonomies, i.e. hyponymic hierarchies: sets of hyponyms (e.g. names of kinds of dog) are included in (and thus totally overlap with) sets defined more generally (e.g. names of kinds of mammal). There are structures, not only in lexical semantics but also in describing syntactic and phonotactic relations, for example, which are not expressible in terms of tree graph structures, but require cross-classification in terms of orthogonal properties.

Compositional mesostructural generalisations express syntagmatic relations, i.e. relations between parts, and between parts and wholes, including constituency relations and dependency relations, (and anaphoric relations, which are less relevant to lexicographic issues). Syntagmatic inclusion relations are also represented by tree graph structures (constituent structures, parse trees). In the context of spoken and multimodal lexical objects, the realisation of syntagmatic relations is quite problematic: in texts, syntagmatic relations are formulated by means of the concatenation operation over characters and sequences of characters, expressed as one-dimensional linear spatial layouts. This is not the case in multidimensional hypertext layouts. Nor is it the case in spoken and multimodal lexical objects, which are in general mapped to both time and space, and are associated with simultaneous as well as sequential events. In these cases, a different operation, that of *temporal overlap* is defined in addition to the operation of *temporal precedence*. Simple cases such as primary and secondary stress are rather well-behaved, and are associated with specific vowels or syllables in the pronunciation representation. More complex cases, such as the association of a lexical entry with both pronunciation and manual gesture, require more complex representations as graph structures; adequate representation of these is only feasible electronically, with suitable procedural navigation methods.

11 Lexical representation: mesostructure and the inheritance lexicon

One of the most widespread techniques for representing lexical generalisations is the inheritance lexicon, first developed in the 1970s as a way of implementing certain kinds of systematic inference in Artificial Intelligence. An inheritance lexicon provides a declarative representation with the following properties:

1. The basic lexical objects are the lexical entries and the lexical classes.
2. Each lexical entry has a microstructure which is assigned properties of two types:
    1. Types of lexical information (often as attribute-value pairs, sometimes as atomic properties); this relation is traditionally known, in the Artificial Intelligence community, as the *HASPROP* ('has property') relation.
    2. An assignment to a class of similar lexical objects (in the case of simple inheritance) or more than one class assignment (in the case of multiple inheritance); this relation is traditionally known, in the Artificial Intelligence community, as the *ISA* ('is a') relation.
3. Each class of lexical entries is assigned the same kinds of property as the lexical entries themselves:
    1. Characteristic types of structural and interpretative lexical information which the members of the class have in common.
    2. Assignment of similar lexical classes to a superclass for simple inheritance (or more than one superclass for multiple inheritance).
4. An inference theory for *inheriting* properties of higher level classes, consisting of
    1. An interpretation of the superclass-subclass relation, and the class-entry relation in terms of implication: *IF a class to which a lexical entry belongs has a certain property, THEN the lexical entry also has this property.*
    2. Use of the transitivity of implication, so that *IF an arbitrarily higher level class has a certain property, THEN all lower classes also have this property.*

3. A redundancy removing convention, such that

   IF a property can be inherited from a higher level class, THEN it need not be specified for any lexical entry or any subclass below this higher level class.

The example which is systematised in Table 2 is taken from the *Longmans Dictionary of Contemporary English* (LDOCE, Procter 1978), with relevant definitions extracted from the articles and presented in plain glossary form using the simple table model for basic macrostructures. The table of defining terms could be extended, of course: many of the terms remain undefined. The definitions in this example would leave a lot to be desired in a terminological lexicon, but will be fine for present purposes.

Table 2: Tabular glossary for 'flageolet' and defining terms

Definiendum	*Definiens*
flageolet:	a small wind-instrument like a whistle, with 6 holes for the fingers
wind-instrument:	any musical instrument played when air is being blown through it
musical:	of, like, or producing music
instrument:	an object which is played to give musical sounds
sound:	a sensation in the ear
whistle:	a simple musical instrument for making a high sound by passing air or steam through
hole:	an empty space within something solid
finger:	one of the 5 movable parts with joints, at the end of each human hand

In the Integrated Lexicon framework, the abstract lemma corresponds to the *definiendum* component of the definition (i.e. the headwords in these cases), and the *definiens* corresponds to the microstructure (in this case just the lexical definition). The important point is that the type of definition is the classical *definitio per genus proximum et differentiae specificae*[9] which involves relations of generality and specificity on the basis of which implications can be formulated and used in inference. In the case of 'flageolet', the nearest kind is represented by the word 'wind-instrument', itself a lexical entry; the specific differences are that it is like (but not identical to) a whistle, has 6 holes for the fingers. The implication relation between the nearest kind (the *genus proximum*) in the *definiens* and the *definiendum* permits inference, for instance, that the flageolet is a musical instrument played when air is being blown through it, though this is not stated explicitly in the lexical entry for 'flageolet'; the lexical entry for 'flageolet' is underspecified, and the entry is completed by inferring the generalisable information from higher classes.

This kind of definition by nearest kind and specific differences fulfils all the conditions for an inheritance lexicon: lexical objects are assigned properties and classes (in this case, the classes are also associated with lexical entries - this need not be the case). In traditional terms, the definition determines a hierarchy of implication relations between properties of lexical entries, i.e. a taxonomy. The 'nearest kind' assignment fulfils the class assignment condition, and the 'specific differences' assignment fulfils the non-redundant property assignment condition. The subordinate term can then inherit properties from the higher level term. In this case, an additional special kind of metonymic inheritance ('like a whistle') is included.

A simplified inheritance structure which has been reconstructed from the implication relations in the LDOCE definitions is shown in Figure 5. The 'ISA' inheritance relation is shown by the solid arrows; metonymic inheritance from peers (objects of the same type) is shown by the dashed arrow.

Hierarchical inheritance relations can be defined for other kinds of lexical information than

[9] Glosses: *definiendum* = 'that which is to be defined', *definiens* = 'that which defines'; *definitio per genus proximum et differentiae specificae* (sometimes: *differentia specifica*) = 'definition by nearest kind and specific differences'.

semantic definitions, including features which are characteristic of the forms of spoken and multimodal lexical items, such as prosody. Other areas in which inheritance lexica have been developed, for example, have included phonological and prosodic hierarchies (Gibbon 2001), and morphological hierarchies of paradigms and sub-paradigms (Bleiching et al. 1996, Lüngen 2002, Lüngen & Sporleder 1999, Sporleder 2004).

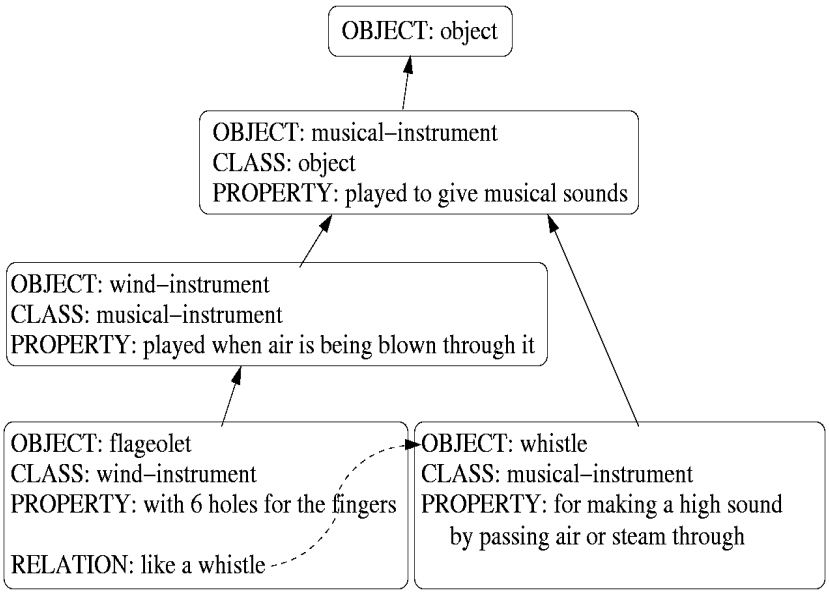

Figure 5: Inheritance structure for 'flageolet'.

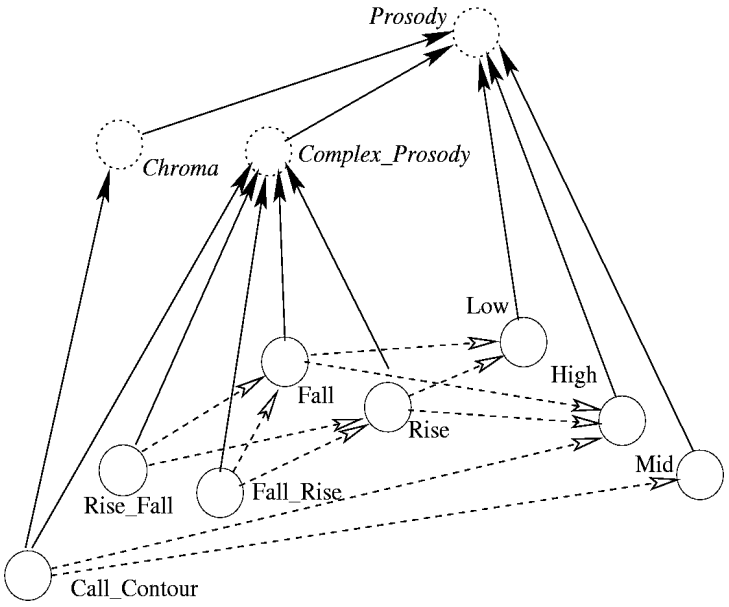

Figure 6: Prosodic inheritance hierarchy for 'call contour' lexical entry.

A hierarchy for the English 'call contour' intonational lexical entry (as in the chanted call *Johnny, where are you?*) and other intonational lexical entries ('rise-fall', 'fall-rise', 'rise', 'fall', 'low', 'high', 'mid') is outlined in Figure 6. The lexical entry nodes in the inheritance graph are

represented by solid circles, the ISA inheritance relations are solid arrows, the abstract lexical class nodes are represented by dotted circles, and the syntagmatic or compositional inheritance of properties of parts ('percolation') is represented by dashed arrows. The chanted call contour, for example, inherits properties from the abstract class *Chroma* (the chant property) which inherits from *Prosody*, and it inherits compositional properties from its parts, the *High* and the *Mid* tones of which it is composed.

In the present context it is not possible to do more than hint at extensions beyond prosodic inheritance to multimodal inheritance relations. Pioneering work was done by Bleiching (1992) in designing the Verbmobil German lexicon of fully inflected forms on Integrated Lexicon principles, including lexical stress patterns, using an inheritance lexicon. In this lexicon, abstract lemmata are underspecified and inflections are inherited from a complex inheritance hierarchy of paradigms and sub-paradigms for nouns and verbs; the correct stress patterns are assigned recursively to compound words by syntagmatic, compositional or percolation inheritance of properties of constituents (see also Carson-Berndsen & Gibbon 2002b, Gibbon 2002a).

12 Functionalities and formats: implementing a lexicon document

The implementation of a lexicon document follows general principles of software development projects, whether the implementation is a traditional printed book or an electronic document. One very simple model for an implementation procedure requires six phases:

1. specification of functional requirements (i.e. who is the product for, how it is to be used),
2. design of the document from the points of view of content, structure and rendering,
3. the implementation itself in some hardware and software environment,
4. the evaluation of the product,
5. the marketing of the product,
6. the maintenance of the product.

The following subsections are concerned with phases 1, 2 and 3 in the context of the Verbmobil speech-to-speech translation project of the 1990s.

There are many ways of specifying functionality of lexica; the constitutive factors of Jakobson (Jakobson 1960), are still useful as a source of analogy for this purpose: the sender is the producer, including the lexicographer, the receiver is the consumer, the message is the lexicon, the context is the vocabulary domain, the code is the metalanguage of the lexicon, and the channel is the dissemination medium; other functionalities, familiar from lexical verb frames, include the instruments of lexicography and lexicography dissemination, and various time and space requirements. These are extensional properties, as opposed to intensional properties such as purpose and intention of producer and consumer. Focussing on extensional properties, the lexicon has the following dimensions of functionality:

1. Producer: lexicon producers range from individual linguists, language learners and teachers to large commercial lexicography units. The tools of the trade range from paper notebooks to lexicography workbench software applications.
2. Consumer: roughly speaking, there is a distinction between human use, with paper or electronic dissemination formats, and machine use in speech and language engineering systems, with machine processable lexical databases for speech recognition (for grapheme-phoneme conversion, word models, language models), speech synthesis (for text parsing, grapheme-phoneme conversion, word prosody assignment), and machine translation (multilingual dictionaries, translation memories, terminology databases, speech-specific vocabulary - cf. Gibbon & Lüngen 2000). There are also intermediate uses, such as the dictionaries and thesauri in word processors and dictation software, and utilities for

hyphenation, realtime contextual character replacement, spell-checking, word completion and grammar or style checking,

3.Context: the content of the lexicon, i.e. a specified vocabulary, is specified according to the criteria discussed in preceding sections, in terms of document structure and its semantic and modality interpretations.

4.Code: the technical language of lexicography, the language used for the metainformation of the lexicon metadata and front matter, and the natural language used to express these, have been introduced implicitly and non-exhaustively in this contribution, but require detailed specification for the implementation of a given lexicon.

5.Channel: the media of lexicon work, storage and dissemination also require specification:

1.*Lexicographer's formats*: the database formats used by the lexicographer. Typically, the lexicographer's working format will include all kinds of housekeeping data and corpus source information (metadata), and be in a flexibly accessible and maintainable database environment.

2.*Archivist's formats*: the database used for sustainable, reusable and interpretable storage and access for future applications. The most widely held view is currently that the optimal archivist's format is defined in XML (eXtensible Markup Language), and for archiving purposes the dictionary text is provided with XML markup.

3.*Dissemination formats*: a wide range of printed and electronic dictionary formats for the consumer, with personal or professional, academic or industrial applications.

An basic distinction is made between procedural and declarative information in lexicon development. This distinction corresponds roughly to the distinction between the functionality of a lexicon for a particular use *how* - and the information in a lexicon. Traditionally, this kind of distinction is referred to as 'knowledge *how* vs. knowledge *that*' or '*how-to* information' vs. '*factual* information'. No lexical (or other) representation is entirely declarative and non-procedural: some procedure for creating or extracting information is always required, whether straightforward consultation (for a Level 2 protolexicon) or sophisticated logical inference (for Level 4 abstract representations). A declarative representation is perhaps best thought of as a 'minimally procedural' representation with no particular application in mind, whereas a procedurally optimised representation will require more complex construction and access algorithms and strategies.

In the domain of the lexicon, procedural information pertains to the acquisition of lexical information, access to lexical information, the archiving and maintenance of lexical information, and the production and dissemination of lexical information in print or electronic media. In the context of lexica for spoken and multimodal language, procedures will differ greatly from one modality to another. The semasiological-onomasiological distinction is, for example, a purely procedural distinction based on different access procedures required for accessing from the form or content perspectives.

There are very important declarative and procedural issues from a computational point of view. One type of procedural optimisation, for example, is quite common in ICT systems. Given a particular mapping between types of lexical information, the structure of the lexicon by using a decision tree. In a form-based lexicon, starting from a common root node, the tree then branches into possible first characters (standing for first orthographic letters) in lexical entries, then possible next characters, and so on, until all lexical entries have been integrated into the tree. The conceptual hierarchy in a thesaurus is also, from a declarative point of view, a decision-tree based form of optimisation. This form of procedural optimisation leads to very efficient searches, provided that left-right form-based search is required: the time taken by the search depends primarily on the length of the entry and not primarily on the size of the dictionary, which is the usual case in human search. The cost of this procedurally driven optimisation is that other functionalities are not efficiently supported.

Hierarchical representation techniques have been developed both in linguistic contexts (e.g. in the *Head-driven Phrase Structure Grammar*, *HPSG*, paradigm) and in computational contexts (e.g. the DATR lexical representation language and the Integrated Lexicon modelling conventions associated with it). A special case of hierarchically organised database structure is found in the increasingly popular *XML* (*eXtended Markup Language*) paradigm for the hierarchical representation of textual and other information on the internet and elsewhere, including lexical information. The tree structure of the XML markup language is generally used to represent hierarchies of textual objects (entities), with properties which are represented as attribute-value pairs, and consequently provides a near-ideal data structure for the representation of many structural properties of lexica.

A hierarchical document description language such as XML has the advantage of simplicity and simple computational processing, and powerful query techniques are available in XML-related technologies (Sasaki, Witt, Gibbon & Trippel 2004). But XML also has the disadvantage of being unable to cope with structures which are not trees without additional interpretations imposed informally by the user. Examples of such structures are:

1. Embedded table structures which are used, for example, in the following lexical contexts:
    1. Lexical entries in procedurally optimised lexica (Level 3 lexicon already described). The sub-entries have the same data category vector structure, of the same length, i.e. if represented by trees, the branches of the tree have equal numbers of sub-branches, i.e. equal fan-out. This is a dependency relation which cannot be given a general description in terms of simple tree structures of the kind which are expressible by means of XML (though of course in a given lexicon, the entries can easily be formulated in terms of a specific tree structure, and thus also in XML).
    2. Morphological paradigm descriptions, which contain recursively embedded tables; this is the general case of structures with equal child branching width. Tables of this kind are also found in L3 lexica with procedurally optimised representations of lexical entries with sub-entries: the sub-entries have equal numbers of internal branches for the same types of lexical information. Trees with equal child branching width constraints are more complex than general tree structures.
2. Arbitrarily linked networks, i.e. networks which are not restricted to tree-structures but may correspond to general acyclic (even cyclic) graphs, and are constructed on the basis of formal relations and/or hyperlinks, for example for the purpose of representing hierarchical lexica with additional dependencies, or cross-references between lexical entries. Network lexica of this kind include inheritance lexica (Gibbon 1991) wordnets (Fellbaum 1998), termnets and framenets (Fillmore & Atkins 1998, Boas 2002).
3. Partially synchronised parallel events involved in prosody and multimodal communication (the lattices involved in representing such events are not trees). For example, a gesture pointing to an object which is referred to in an open utterance will almost certainly not occur in exactly the same interval as an utterance of "That!", particularly if the deictic pronoun is embedded in a longer utterance. Consequently, the pointing gesture and the word cannot conveniently be included in exactly one segment of time (which conventional constituent structures generally imply): the gesture overlaps different locutionary events, and the locutionary events overlap different portions of the gesture, and a more complex kind of directed acyclic graph is needed to represent the partially synchronised parallel events.

From a computational point of view, information which may be described by means of tree structures can be straightforwardly modelled by a standard formal language type called a *context-free language*, which can be described by a formal grammar type called a *context-free grammar*. XML is a context-free language. The more general graph structures described above require more powerful grammars; for example, equal branching width for sibling

nodes, and specifically embedded table structures, require a more complex form of grammar called a *context-sensitive* grammar (in fact, a restricted type of context-sensitive grammar called an *indexed grammar*) to describe them adequately.

13 Case study: aspects of lexicography in speech-to-speech translation

Lexicography in the Verbmobil speech-to-speech translation consortial project (Wahlster 2000) was based on the Integrated Lexicon concept, and used lexical databases of all four orders in the rank scale of lexical abstraction.

The Verbmobil project lasted for 8 years, in two 4 year phases. The consortium required lexicographic coordination for most of the 32 partners involved in the consortium. In Verbmobil Phase I, there were 16 subprojects; project 5, coordinated by the author, was concerned with lexicon and morphology, interacted with all the other projects and contributed to several system modules, in particular the speech recognition, the deep syntactic analysis, semantic construction, dialogue semantics, transfer, generation and speech synthesis modules. The lexicographic tasks will be discussed in the following sections (cf. also Gibbon 2000 on the treatment of the forms of lexical entries, and Gibbon & Lüngen 1999, 2000 on the translation context).

The main lexicographic requirements of the partners were the following:

1. German speech recognition (for the grapheme-phoneme pronunciation lexicon - the speech recognition components for other languages were independently developed),
2. prosodic analysis (for word stress),
3. deep analysis (for parts of speech),
4. dialogue-act based translation modules, semantic construction and dialogue semantics (for semantic interpretations),
5. transfer rules for German-English and English-German translation,
6. generation component (for onomasiological information in German as a target language).

The lexicographic products of Project 5, orientated towards these needs, included the following:

1. Lexicographic data:
    1. A lexical database for fully inflected German words, oriented towards speech recognition systems and machine translation (with web interface).
    2. An on-demand concordance for the corpus transcriptions (with web interface).
    3. Morphological knowledge base (Bleiching 1992).
    4. Printed version of the substantive fields of the lexicon for demonstration purposes.
2. Development tools:
    1. System for identifying phonologically confusable words in order to trigger clarification dialogues (with web interface).
    2. Syllable generator and checker (Gibbon, Simões & Matthiesen 2000)
3. System components:
    1. Morphological components, including the repair of fragmented words in disfluent contexts (Althoff 1997; Langer 1990).
    2. Design prototype for cascaded finite state phonological and morphological processing in speech recognition (Carson-Berndsen 1998, Pampel 1999).
4. Computational linguistic components:
    1. A morphological inheritance network implemented as a Level 4 lexicon in DATR, later ported to Prolog (Bleiching 1992, Bleiching et al. 1996, Lüngen 2002).

1. Grapheme-phoneme transducer for unknown (out of vocabulary) words.
2. Morphophonological, phonological and syllabification components for surface phonetic form generation (Gibbon et al. 2000; Matthiesen 1998).

In declarative terms, the lexicographic work in Bielefeld started with corpus input from Verbmobil partners in Munich, who created a spoken language corpus consisting of about 20,000 recordings of bilingual appointment scheduling dialogues in the languages German, English and Japanese. The main lexical database was designed in the following phases.

1.The initial stage was to construct a first order corpus lexicon for the German components of the transcribed recordings of a appointment scheduling scenario.

2.From the corpus lexicon, a second order lexicographer's protolexicon was constructed, by integrating information from Verbmobil partners. The second order protolexicon has option-free microstructure and no generalisations over lexical entries: one lexical entry for each assignment of a fully inflected orthographic form to its inflexional category. The initial set of fully inflected forms is determined by the set of fully inflected orthographic forms which occurred in the corpus transcriptions.

3.The inflexion-free stems were extracted from items in the second order lexicon, and a hierarchical morphological knowledge base for German was developed, constituting a fourth order hierarchical inheritance lexicon implemented in the lexical representation language DATR. The hierarchical lexicon contains morphological generalisations which project the extant forms to fully populated paradigms of inflected forms for verbs, nouns and adjectives. The knowledge base correctly covers the entire range of German inflexional forms, and remained unchanged until the end of the project, despite the continual addition of new words (Bleiching & al. 1996). A list of proper names was also included.

4.The fourth order lexicon was used to infer the full set of inflectional forms for all stems, which were fed back into the second order lexicon, and a degree of generalisation was introduced in order to be able to list the full range of morphological categories of syncretistic (morphologically ambiguous) fully inflected forms in a single fully inflected entry. The Verbmobil lexical database is thus a third order optimised lexical database based on fully inflected forms.

5.The transfer lexicon for the machine translation stage of the Verbmobil prototype system was constructed mainly by partners in the Tübingen Verbmobil project in cooperation with other projects specialising in semantics, and integrated into the Bielefeld third order database.

The extensional coverage of the database (i.e. the number of entries) is about 10,000 fully inflected corpus-derived forms, corresponding to approximately the same number of stems, extended to a full set of 50,000 inflected items by inference from the inheritance lexicon. The size of the lexicon is much smaller than a typical text-oriented lexicon because of the constraints determined by the capabilities of the speaker-independent speech recognisers at the time (mid 1990s): the lexicon is based on transcriptions of a specially recorded corpus of negotiation dialogues in an appointment scheduling domain. The morphological full paradigm projection using the fourth order lexicon, together with the proper name list, contributes towards a solution of the out-of-vocabulary word problem.

The intensional coverage of the Phase I database, i.e. the number of data categories with assigned values in the microstructure, has 35 data categories, divided into 3 groups:

1. Morphology, morphophonology, morphosemantics: orthography, morphologically segmented orthography, phonemic transcription with prosodic information (syllable boundaries, word stress), orthographic stem, phonemic stem, inflexional categories, canonical lemma form, spelling (a comment field for corpus transcription errors), proper name tag, compound word semantics.

2. Corpus distribution, selection, tagging: quantitative information about occurrences in different corpora and subcorpora.

3. Syntax, Semantics, Transfer, Dialogue, Glossary: local orthographies used by partner projects; syntactic and semantic analyses.

The microstructure of the third order optimised lexicon is therefore not designed to produce a specific dictionary in the traditional sense, but to be a reference tool for consortium partners which integrates many types of practically useful information, from phonemic transcriptions to frequencies in different corpora, into a single database.

On the procedural side, lexical data acquisition and integration took place in a complex normalisation process (Lüngen & Gibbon 2000). Information for the data categories was provided by partners in a wide range of formats, all of which were standardly encoded with UNIX conventions in keyboard-friendly (and lexicographer-friendly) 7-bit ASCII characters.

At the character level, the orthography for German was encoded in an agreed canonical 7-bit ASCII format. The German non-ASCII characters were represented as shown in Table 3 (adopting the LaTeX German style convention, which was also a common email convention at the time). The phonemic representations were encoded in the keyboard-friendly encoding of the IPA known as SAMPA (cf. Gibbon et al. 1997, Appendix B), using a slightly modified version of phonemic German SAMPA with extensions to indicate prosodic information. These were straightforward to process with standard UNIX script prototyping techniques. However, the data structures in the different databases supplied were very different, in some cases actually being complex hierarchically structured inheritance lexicon formats with embedded feature structures.[10] For all these operations, a suite of UNIX shell scripts was developed, using standard UNIX tools.

Table 3. LaTeX-style ASCII encoding of German special characters.

Lower Case		*Upper Case*	
Normal	*Encoding*	*Normal*	*Encoding*
ä	"a	A	"A
ö	"o	O	"O
ü	"u	U	"U
ß	"s		

The main lexical database is implemented as a relational UNIX database. A single lexical relation represents the sequence of records of lexical entries, the columns representing the lexical microstructure. Straightforward techniques were used for rapid prototyping and constant updating as the corpus grew and new data categories as well as additional values for the data categories became available. All representations are encoded in 7-bit ASCII, and a standard CSV (character separated value) format is used for the relational database table, following conventional UNIX practice.

A major requirement for procedural optimisation of the lexicon macrostructure was stipulated by the speech recognition partners: the lexicon was to be based on fully inflected forms for straightforward grapheme-phoneme conversion in the training of Hidden Markov Models. From the spoken language perspective this constraint is unfortunate, because it introduced orthographic noise, i.e. heterophonic homographs (spellings with different pronunciations) and heterographic homophones (pronunciations with different spellings). However, the statistical methods used in speech recognition minimise the disadvantages.

The second order lexical database, a protolexicon with no optionalities or ambiguities or generalisations, was compressed automatically into a much smaller third order optimised

[10] In one such case, the structure was based on company-internal specifications which could not be distributed. For the lexicon database, the format therefore had to be reverse engineered (by permission) and normalised to fit the Verbmobil lexical database. After normalisation the database was integrated automatically.

lexicon, also in tabular form, by combining into one entry cases of inflexional syncretism, i.e. ambiguous fully inflected forms with different morphological categories. For example, German *Tage* is nominative, accusative and genitive plural, and, in formal styles, also dative singular. The inflexional syncretism information is compressed into a list under the data category for inflexion in the microstructure, creating a flat hierarchy, representing disjunctions, for the relevant data categories. Semantic ambiguities are compressed similarly. Expansion to the full second order lexical structure containing separate entries for every homophonous option is always possible. Partners did not in general require this full expansion, however, but restricted their attention to ambiguities in specific microstructure data categories.

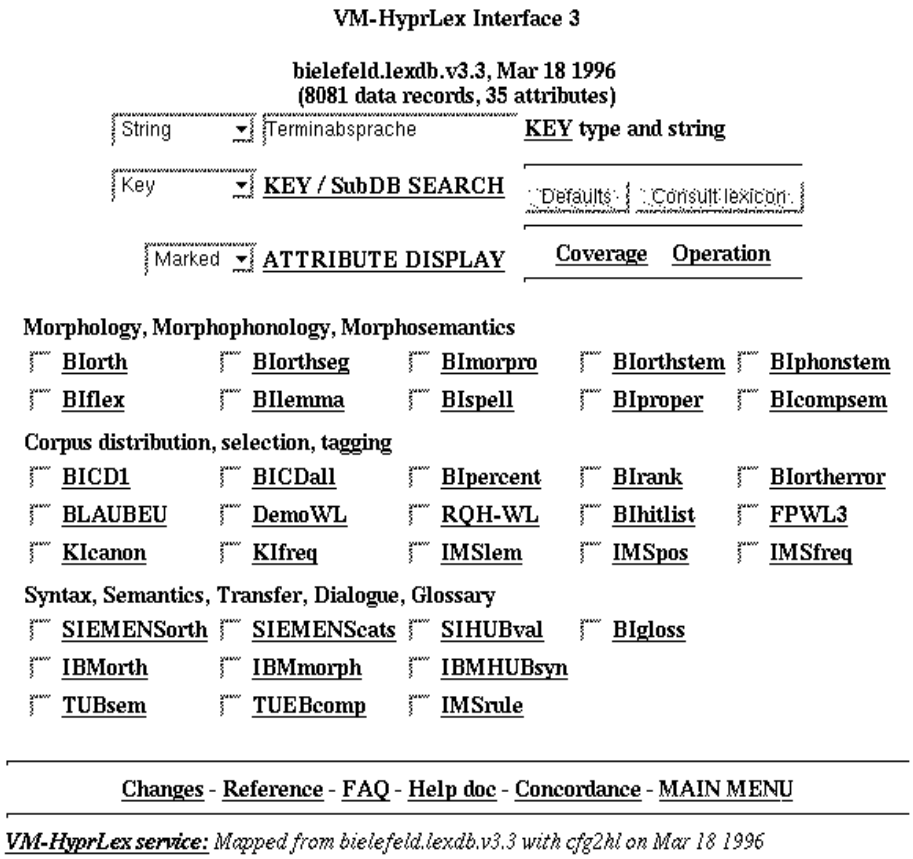

Figure 7: Web interface to Verbmobil Phase I database.

The database was to be made available to all Verbmobil partners worldwide, so the problem of consistency between database copies arose. This problem was solved in the simplest possible way: just one token of each version of the database was made available to all partners simultaneously on the World-Wide Web via a hypertext lexicon interface or hyperlexicon (known as *HyprLex*) thus guaranteeing version consistency. Access to this single token was provided via the central Verbmobil lexicography server in Bielefeld, using HTML forms and CGI scripts. With a single token, by definition no version inconsistencies can arise since no copies exist (though later some partners started using mirroring techniques). With the spread of large-scale commercial and other database applications on the web, varieties of this technique are now commonplace, of course, and much more sophisticated. However, at the time of inauguration (1994) the Verbmobil lexical database web facility was said to be among

the largest and most complex anywhere.[11] The HyprLex service included extensive metadata (including help information and intensional and extensional coverage statistics) on the current state of the database, as well as an on-the-fly dynamic concordancer covering the entire transcribed corpus, and a number of other tools for spoken language lexicography.

Figure 7 shows a screenshot of an early Verbmobil Phase I version of the web interfaces to the database. The interface form was generated automatically from a template, using the data category set for automatic construction of selectable output filters.

The output from the query shown in the Key field of Figure 7 is given in Table 4, which shows the full microstructure of the database in the optimised form described above.

Table 4: Database response to query with all data categories selected.

Category	*Value*
BIorth:	Terminabsprache
BIorthseg:	Termin#ab#sprach#+e
BImorpro:	tE6.m'i:n#?''ap#Spr''a:.x#+@
BIorthstem:	Termin#ab#sprach
BIphonstem:	tE6.m'i:n#?''ap#Spr''a:x
BIflex:	N,akk,sg,fem;N,dat,sg,fem;N,gen,sg,fem;N,nom,sg,fem
BIlemma:	Terminabsprache
BIspell:	--
BIproper:	--
BIcompsem:	ObjEreig
BICD1:	cd1=2_cd12=7_cd3=2_cd4=3_cd5=1
BICDall:	15
BIpercent:	0.00568005%
BIrank:	977
BIortherror:	Termin-Absprache,-
BLAUBEU:	--
DemoWL:	demo-wl
RQH-WL:	--
BIhitlist:	hit#977=15
FPWL3:	fpwl
KIcanon:	tE6m'i:n#Q"ap#Spr"a:x@
KIfreq:	14
IMSlem:	Terminabsprache
IMSpos:	NN
IMSfreq:	8
SIEMENSorth:	Terminabsprache
SIEMENScats:	sem_lex(nr,terminabsprache) & nr:rel=terminabsprache&sortal_Terminabsprache(nr) & count_noun_norm(nr) & subst_klasse2_1(nr)terminabsprache & sortal_einigen_auf & count_noun_norm&subst_klasse2_1
SIHUBval:	--
BIgloss:	appointment_scheduling
IBMorth:	--
IBMmorph:	--
IBMHUBsyn:	[gender:fem,number:sg,case:ncase_v, syn_ibm: [phon:'Terminabsprache',cuf_macro:common_noun_syn], person:3]
TUBsem:	terminabsprache & communicating & -
TUEBcomp:	terminabsprache: compound(terminwoche,first(termin), second(absprache), semrel(arg3_rel)).
IMSrule:	terminabsprache:[H:terminabsprache(I)] <-> [H:scheduling(I), H1:indef(Y,H2), H2:appointment(Y), H3:of(I,Y)].

14 Conclusion and prospects

A systematic conceptual and terminological approach to lexical acquisition and lexical representations was introduced in this contribution in the form of the semiotically motivated Integrated Lexicon (ILEX) framework. The framework includes a scale of corpus-to-lexicon abstraction and a standard lexicon structure consisting of megastructure, macrostructure,

[11] The original lexicon server is still running: <www.spectrum.uni-bielefeld.de/VM-HyprLex/>.

microstructure and mesostructure. A range of theoretical issues involved in developing Integrated Lexicon theory was discussed, and a practical lexical database application in speech-to-speech translation context was described, using the corpus-to-lexicon abstraction scale as an organising principle. It has been shown that a comprehensive integrative approach towards a theory of multimodal lexicon construction and lexical representation and processing is possible, though some rapidly progressing areas (such as lexical machine learning and automatic distributional analysis) were inadequately dealt with, many open questions remain unanswered, and many details need to be filled in and inhomogeneities to be smoothed out. The field is developing rapidly in terms of new intellectual questions to be addressed in these respects, but new strategies are becoming available for answering the questions, and the development of practical tools for supporting the strategies is accelerating.

When, for example, the projects on which the developments reported in this contribution were first launched, in the late 1980s, the World Wide Web did not exist. It could not have been predicted at the start of the Verbmobil project that the lexicography database would be networked and used on a daily basis for spoken language resource and system development by a distributed world-wide consortium, with instant availability of consistent updates.

Nor can it be predicted now, except in the most general terms, what kinds of facilities for spoken language and multimodal lexicography will be available four years after the publication of this contribution. Developments in data-mining and machine learning for lexical acquisition (Matthiesen 1998, Lüngen & Sporleder 1999, Trippel, Sasaki, Hell & Gibbon 2003, Sporleder 2004), and in multimodal data processing and rendering for lexical access, as well as further miniaturisation and wireless operation of computing devices will certainly revolutionise the field during this period. But how this will happen is anybody's guess. At the end of the Verbmobil project in 2000, only the boldest could have predicted that wireless internet multimodal database consultation via miniature handheld devices such as mobile phones and personal digital assistants (PDAs) would be commonplace four years later, or that these tiny devices would store collections of much larger databases - including video databases and lexical databases with audio renderings - than the Verbmobil lexicon.

15 References

Althoff, Frederek (1997). *MEWES: Ein Modul für den Einsatz Morphologischen Wissens bei der Erkennung gesprochener Sprache*. M.A. Thesis, Universität Bielefeld.

Bleiching, Doris (1992) Prosodisches Wissen in Lexicon. In G. Goerz, ed., *Proceedings of KONVENS-92*, Berlin: Springer-Verlag, pp. 59-68.

Bleiching, Doris, Guido Drexel, Dafydd Gibbon (1996). Ein Synkretismusmodell für die deutsche Morphologie. In: Dafydd Gibbon, ed. *Natural Language Processing and Speech Technology. Results of the 3rd Conference "Verarbeitung Natürlicher Sprache" (KONVENS)*, pp. 237-248.

Boas, Hans C. (2002). Bilingual FrameNet Dictionaries for Machine Translation, I n *Proceedings of LREC 2002, Las Palmas,* pp. 1364-1371.

Carson-Berndsen, Julie (1998). *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition.* Dordrecht: Kluwer Academic Publishers.

Carson-Berndsen, Julie & Dafydd Gibbon (2002). Visualising lexical prosodic representations for speech applications. In: Paul McKevitt, Seán Ó Nualláin & Conn Mulvihill, eds. *Language, Vision and Music*. Amsterdam: John Benjamins.

Evans, Roger & Gerald Gazdar (1996). DATR: A language for lexical knowledge representation. *Computational Linguistics* 22.2, 167-216.

Fellbaum, Christiane, ed. (1998). WordNet: An Electronic Lexical Database.MIT Press.

Fillmore, Charles J. and Atkins, B. T. S. (1998). FrameNet and lexicographic relevance,

Proceedings of LREC 1998, Granada.

Fischer, Kerstin (2000). Functional Polysemy of Discourse Particles. Mouton de Gruyter: Berlin, New York (also: Ph.D. thesis, Universität Bielefeld: *A Cognitive Lexical Pragmatic Approach to the Functional Polysemy of Discourse Particles*. PhD thesis, Universität Bielefeld, 1998).

Gibbon, Dafydd (1991). ILEX: A linguistic approach to computational lexica. In: U. Klenk, ed. *Computatio Linguae. Zeitschrift für Dialektologie und Linguistik*, Beiheft 73.

Gibbon, Dafydd (2000). Computational lexicography. In: Frank van Eynde & Dafydd Gibbon, eds. *Lexicon Development for Speech and Language Processing*. Dordrecht: Kluwer Academic Publishers, pp. 1-42.

Gibbon, Dafydd (2001). Phonological preferences as defaults. In: Katarzyna Dziubalska-Kołaczyk, ed. (2001) *Constraints and Preferences*. Berlin: Mouton de Gruyter, 143-199.

Gibbon, Dafydd (2002a). Compositionality in the Inheritance Lexicon: English Nouns. In: Leila Behrens & Dietmar Zaefferer, eds. *The Lexicon in Focus: Competition and Convergence in Current Lexicology*. Frankfurt am Main: Lang.

Gibbon, Dafydd (2002b). Prosodic information in an integrated lexicon. *Proceedings of Speech Prosody 2002*. Aix-en-Provence, 335-338.

Gibbon, Dafydd, Roger Moore & Richard Winski (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Gibbon, Dafydd, Silke Kölsch, Inge Mertins, Michaela Schulte & Thorsten Trippel (1999). Terminology principles and support for spoken language system development. In: *Proceedings of EUROSPEECH 99 Budapest*.

Gibbon, Dafydd & Harald Lüngen (1999). Consistent Vocabularies for Spoken Language Machine Translation Systems. In: Jost Gippert, ed., *Multilinguale Corpora. Codierung, Strukturierung, Analyse*. 169-178. Prague: Enigma Corporation.

Gibbon, Dafydd & Harald Lüngen (2000). Speech Lexica and Consistent Multilingual Vocabularies. In: Wolfgang Wahlster, ed. *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer Verlag.

Gibbon, Dafydd, Inge Mertins & Roger Moore, ed. (2000). *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Dordrecht: Kluwer Academic Publishers. 2000.

Gibbon, Dafydd, Ana Paula Quirino Simões & Martin Matthiesen (2000). An optimised FS resource generator for highly inflecting languages. *Proceedings of LREC 2000, Athens*.

Gibbon, Dafyddd & Thorsten Trippel (2000): A multi-view hyperlexicon resource for speech and language system development. *Proceedings of LREC 2000, Athens*, pp. 1713-1718.

Gibbon, Dafydd, Thorsten Trippel & Serge Sharoff (2001). Concordancing for Parallel Spoken Language Corpora. In: *Proceedings of Eurospeech 2001*, Aalborg, Denmark, III: 2059 - 2062.

Gibbon, Dafydd & Thorsten Trippel (2002). Annotation driven concordancing: the PAX toolkit, *Proceedings of LREC 2002, Las Palmas de Gran Canaria*, pp. 1568-1572.

Gibbon, Dafydd, Ulrike Gut, Benjamin Hell, Karin Looks, Jan-Torsten Milde, Alexandra Thies & Thorsten Trippel (2004).. CoGesT: a formal transcription system for conversational gesture. In: *Proceedings of LREC 2004, Lisbon.*

Gibbon, Dafydd, Thorsten Trippel & Felix Sasaki & (2004). Consistent storage of metadata in inference lexica: the MetaLex approach. In: *Proceedings of LREC 2004, Lisbon*.

Hartmann, Reinhard R. K., ed. (1983). *Lexicography. Principles and Practice*. London:

Academic Press.

Landau, Sidney I. (1983). *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Langer, Hagen (1990). Syntactic normalization of spontaneous speech. In: *Proceedings of the 13th conference on Computational linguistics.* Volume 3: 180-183.

Lüngen, Harald (2002). *A hierarchical model of German morphology in a spoken language lexicon environment.* Ph.D. thesis, Universität Bielefeld.

Lüngen, Harald & Caroline Sporleder (1999). Automatic Induction of Lexical Inheritance Hierarchies. In: Jost Gippert, ed. *Multilinguale Corpora. Codierung, Strukturierung, Analyse* pp. 42-52. Prague: Enigma Corporation.

Matthiesen, Martin (1999). *Morphologie im Textmining.* M.A. thesis, Universität Bielefeld.

Pampel, Martina (1999). *Morphologische Wortmodellierung und automatische Spracherkennung*. Ph.D. thesis, Universität Bielefeld.

Procter, Paul, ed. (1978). *The Longman Dictionary of Contemporary English.* Harlow & London: Longman.

Sasaki, Felix, Andreas Witt, Dafydd Gibbon & Thorsten Trippel (2004). Concept-based Queries: Combining and Reusing Linguistic Corpus Formats and Query Languages. In: *Proceedings of LREC 2004, Lisbon*. Mit Thorsten Trippel, Felix Sasaki, Andreas Witt.

Sporleder, Caroline. *Discovering Lexical Generalisations. A Supervised Machine Learning Approach to Inheritance Hierarchy Construction.* PhD Thesis, School of Informatics, University of Edinburgh, 2004.

Trippel, Thorsten, Felix Sasaki, Benjamin Hell & Dafydd Gibbon (2003). Acquiring Lexical Information from Multilevel Temporal Annotations. In: *Proceedings of Eurospeech 2003*, Geneva.

Tseng, Shu-Chuan (1999). *Grammar, Prosody and Speech Disfluencies in Spoken Dialogues*. Ph.D. thesis, Universität Bielefeld.

Witt, Andreas, Harald Lüngen & Dafydd Gibbon (2000). Enhancing speech corpus resources with multiple lexical tag layers. *Proceedings of LREC 2000, Athens.*