Ubiquitous multilingual corpus management in computational fieldwork

Dafydd Gibbon

Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Postfach 100131
D–33501 Bielefeld
Germany
gibbon@spectrum.uni-bielefeld.de

Abstract

The present application addresses the issue of portability in the context of linguistic fieldwork, both in the sense of platform interoperability and in the sense of ultra-mobility. A three-level networked architecture, the UbiCorpus model, for information gathering in the field is described: (1) a Resource Archive server layer, (2) a Data Processing application layer, and (3) a new Corpus Pilot layer designed to support specific fieldwork sessions under adverse conditions, for on-site questionnaire presentation and metadata editing.

1. Goals

In linguistic fieldwork,[1] conceptually the initial stage in any language documentation procedure, the issue of portability is important in two senses: first, the sense of platform interoperability and second, in the sense of ultra-mobility. This issue is addressed by the present application. A three-level networked architecture, the UbiCorpus model, for information gathering in the field is described: (1) a Resource Archive server layer, typically non-mobile, and distributed; (2) a Data Processing application layer, typically a local laptop or desktop; and (3) a new Corpus Pilot layer, designed to support specific fieldwork sessions under adverse conditions with questionnaire presentation and metadata editing, and typically, it is suggested, implemented on a handheld PDA. The UbiCorpus model is based on extensive fieldwork experience, mainly in West Africa. The Corpus Pilot layer is described in detail.

Owing to severe financial and platform resource limitations in practical linguistic fieldwork situations, the general development strategy is to use available freeware or open source components as far as possible, and to augment these with custom applications which are distributed as freeware for initial testing, and subsequently published as as open source software.

Figure 1: Questionnaire-based interview on Anyi syntax with Kouamé Ama Bié by Sophie Salffner & Sandrine Adouakou in Adaou, Ivory Coast (equipment: field laryngograph, DAT, Palm, pen & paper).

2. Requirements specification

Relatively recently, issues of corpus standards and resources as developed in the field of speech technology (Gibbon et al., 1997; Gibbon et al., 2000; Bird and Liberman, 2001) have been extended to fieldwork corpora in linguistics, ethnography, and related sciences, and specific issues such as the role of metadata in resource archiving and reusability have come to the fore, adding to the complexity of the documentation task facing the fieldworker. The present application area is computational support for this fieldwork documentation task within an integrated fieldwork resource environment. This concern is on the one hand more comprehensive than the currently popular issues of annotation-based data enhancement and web-based resource dissemination, and on the other hand orthogonal to these expensive technologies in that an effective but inexpensive practical new "low end high tech" technique for grass roots applications in geographically inaccessible areas is introduced.

From the perspective of field linguistics, language documentation traditionally consists in the main of field notes, an outline of the situation of the language, transcriptions, and generally including a sketch grammar consisting of basic phonology, morphology, and grammar, together with a lexicon containing glosses and examples and perhaps a thesaurus. The prompt materials for eliciting this kind of documentation are mainly systematic linguistic and ethnographic questionnaires, and the media for production of the documentation are generally office-oriented software such

[1]Grateful acknowledgements to Sandrine Adouakou, Firmin Ahoua, Doris Bleiching, Bruce Connell, Eddi Gbery, Ulrike Gut, Ben Hell, Sophie Salffner, Thorsten Trippel and Eno-Abasi Urua for discussion of problems addressed in this contribution.

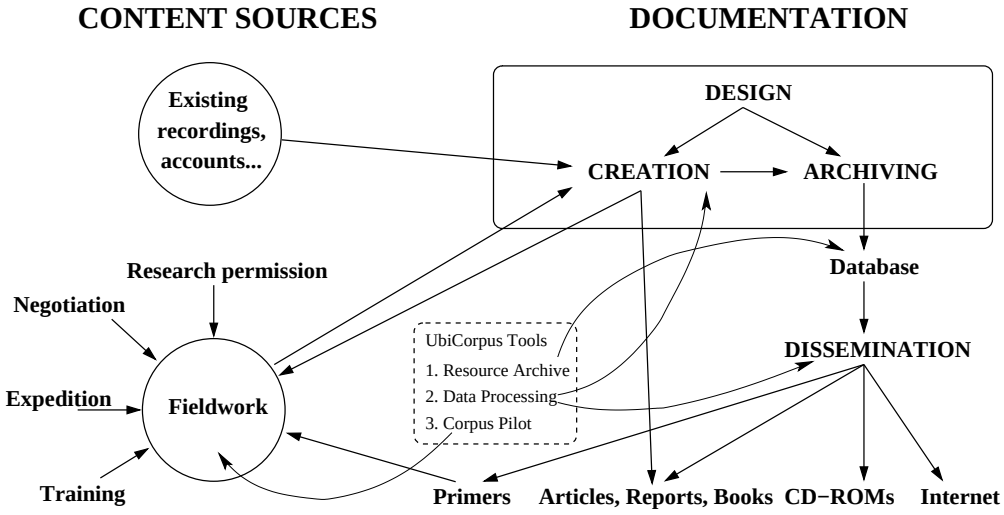

Figure 2: Language documentation logistics model.

as word processors (MS-Word etc.), DBMS (Access, FilemakerPro etc.), and spreadsheets (Excel, etc., also used for database entry). The guiding objectives of this concept of documentation are applications in the production of translations, terminologies, and alphabetisation materials.

The UbiCorpus model is designed to support this kind of fieldwork in the following main respects:

1. questionnaire presentation (either by database or in free format, as a plain text editor or with special formatting and rendering, for example by means of an IPA font),

2. transcription (either plain ASCII such as X-SAMPA, or in an IPA font),

3. metadata input.

One of the main advantages of the model is that when implemented on a modern palmtop device it provides a convenient, efficient and — important for many applications — inconspicuous method for the frequently neglected task of systematic on-site metadata logging.

However, the scope of the model is more general, and supports both the documentation of spoken language corpora in general, and further corpus processing in the form of the development of structured computational lexica (van Eynde and Gibbon, 2000) and computationally supported grammar testing. The UbiCorpus model is embedded in a comprehensive documentation model which covers not only the fieldwork activity itself, but the environment of preparation, archiving and application in which fieldwork is embedded (cf. Figure 2).

The first general operational requirement for the UbiCorpus model is portability. In the present context the term is systematically ambiguous:

- interoperability of applications on different OS and hardware platforms,

- compatibility of data formats through import and export filters for functionally equivalent or interfaced applications,

- ubiquity, i.e. time and place independent mobile deployment.

In the present context, the primary focus is on ubiquity, with interoperability and compatibility seen from this perspective.

Computational support for certain aspects of linguistic fieldwork has been available for many years, both for laptop-based data entry and initial analysis on the move or in isolated areas, and for desktop-based detailed descriptive work and document production (with increasing overlap between laptop and desktop functionalities). Software applications have been characteristically in the following areas:

- Lexical databases, either using general office DBMS such as FileMakerPro and MS-Access, or custom lexicon project software such as SIL's Shoebox; the latter also includes lexical support for textual glossing.

- Publication support such as DB export functions, fonts.

- Phonetic software, for signal analysis (e.g. general signal editors such as CoolEdit, or SIL's CECIL and signal analysis packages, or Praat) and for the symbol-signal time alignment (labelling) of digital recordings (e.g. Praat, Transcriber).

- Computational linguistic software for basic phonological, morphological and syntactic processing.

Some of this functionality (lexical databases, document production, computational linguistic processing) overlaps with the new Corpus Pilot layer, but this layer has the following characteristic additional fieldwork corpus acquisition functionality (Gibbon et al., 1997; Gibbon et al., 2000):

Pre-recording phase: planning of the overall corpus structure and contens, in particular design of corpus recording sessions, including the preparation of scenario descriptions, interview strategies, questionnaires, data prompts (for instance with prompt randomisation),

Recording phase: conduct of corpus recording sessions, including session management with the logging of metadata in a metadata editor and database, questionnaire consultation and data prompt presentation;

Post-recording phase: provision of recorded and logged data for archiving and processing, including metadata export, transcription, lexicon development, systematic sketch grammar support and document production.

3. Design: modules, interfaces

The language documentation model within which the UbiCorpus model is deployed is visualised in Figure 2; the documentation model was developed for project work in West Africa. The two components of the model with which the UbiCorpus tools are concerned are the *Creation* and *Archiving* component, and the *Fieldwork* information source. The latter is directly associated with the Corpus Pilot layer described below. The UbiCorpus model itself is visualised in Figure 3.

The three layers of the UbiCorpus model are characterised as follows:

Resource Archive (RA) layer

The bottom layer represents the archive database and the access and media dissemination functions associated with it. On the declarative side, a number of current language resource and documentation proposals may be assigned to the Resource Archive layer: a single resource database such as a corpus or a lexicon, a multiple resource database such as a browsable corpus or concordance system, a web portal constituting a large and systematic resource world, or an entire dissemination agency. On the procedural side, the Resource Archive layer provides search functions of various kinds, from standard browsing strategies to intelligent search and concordance construction, with token renderings of resources in any suitable media, whether entire corpora or lexica.

Data Processing (DP) layer

This is the layer which is familiar to the "ordinary working linguist". The data include paper fieldwork logbooks, transcriptions, sketch grammars and card index lexica; word processor and database versions of these; analog and digital audio and video recordings; time aligned digital annotations of recordings, and concordance or browsing software based on annotations; metadata catalogues for all of these Data Processing layer data types. Procedurally, the platforms and applications used at the Data Processing layer are very varied, though there is a tendency to go for platform independence and standardised data interchange formats. By using modern laptops, both the Resource Archive and Data Processing layers can be integrated into a single mobile environment.

Corpus Pilot (CP) layer

The top layer of the model represents the functionality which needs to be available in an actual fieldwork situation. This functionality can be very varied, and much — especially free format interviews and film recording — lies outside the range of systematic computational support. However, the following on-site support features can easily be covered:

1. metadata editor and database,
2. participant database for interviewee, interviewer etc.,
3. structured or free format questionnaire presentation.

Interfaces

The interfaces between these three layers, and modules within these layers, are defined mainly on the basis of generic ASCII formats, including XML annotated text, CSV database tables, and RTF formatted documents (including IPA font information). For the interface between a palmtop implementation of the Corpus Pilot layer and the Data Processing layer, conversion scripts are provided as required, in order to export palmtop database and text formats into the generic ASCII formats. Data transfer at the implementation level is via the usual synchronisation functions provided with handheld devices, or via scp, http, and ftp procotols for laptops, desktops and server.

4. Implementation: hybrid applications

Resource Archive (RA) layer

The server archive provides web portal access for the local and global linguistic communities, CD-ROM access for the local linguistic community, and analogue selections (in general, tape cassette, print media) for practical applications in the local user community. Currently, the leading models for the Resource Archive level are provided by the LDC and ELRA dissemination agencies; the E-MELD project is developing a general model for best practice in resource collation, and a meta–portal for flexible access to language resources. The local server currently used for initial database collation contains a number of specific search functionalities for corpus analysis, in particular an audio concordance (Gibbon and Trippel, 2002).

Data Processing (DP) layer

The classical environment for fieldwork data processing is a laptop, often a Mac, but also very frequently an Intel based device configured alternatively with Linux or MS based portable standard software. The kinds of application typically used are for basic corpus processing: Transcriber and Praat for transcription and annotation; Shoebox for lexical database development; MS Office or StarOffice for word processor, database and spreadsheet applications. These may be augmented with custom applications in Java (cf. the TASX engine (Milde and Gut, 2001)) and Perl (PAX audio concordance).

Corpus Pilot (CP) layer

The Corpus Pilot layer is implemented as custom-developed Palm compatible PDA applications. The rationale behind the use of the PalmOS based handhelds, as opposed to the use of a laptop, is based on the following considerations:

1. extremely inexpensive (in relation to other computational equipment),

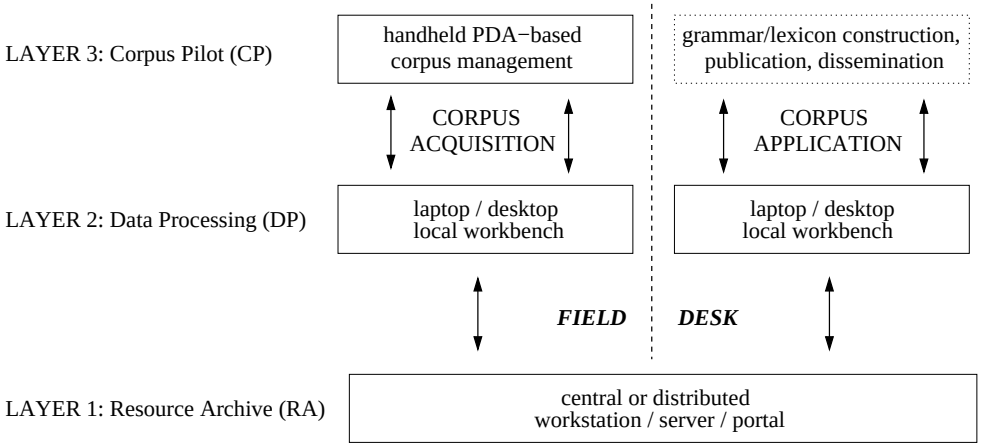

Figure 3: The three layer UbiCorpus model.

2. ultra-lightweight (lighter than other standard portable fieldwork equipment such as field laryngograph, DAT recorder),

3. long operating cycle (with normal use, around 3 weeks on 2 AAA batteries or one charge), depending on model,

4. fast and highly ergonomic in use,

5. small and unobtrusive in the interview situation,

6. an integrated environment with other PDA functionalities such as calendar, diary, address and other databases, other custom applications in C and Scheme.

Networking

The three levels are networked by standard techniques: server-to-applications in general via TCP/IP-based protocols and mobile or landline telephone. The applications-to-acquisition via dedicated sychronisation software of the kind typically used to link handheld PDAs to desktop installations.

Use in the field

The satisfaction of these criteria points towards a high level of suitability for use in extreme fieldwork situations without power supplies, for instance in isolated outdoor locations (forest, village, etc.).

The functionality which has been included in the Corpus Pilot layer so far covers the following:

- Metadata editor and database for audio/video recordings, photos, paper notes, artefact cataloguing. This application is based on a widely used PalmOS DBMS application, HanDBase, which provides a wide range of input support facilities (popups, date picker, free format notes, etc.), as well as cross-table linking.

- Questionnaire administration. In general, free text format has been used for questionnaire administration, and responses have been recorded for later out-of-field processing. For some questionnaire types (e.g. demographic information), the HanDBase DBMS is used.

- Lexicon development tools. Three applications are used for lexical database input (excluding freely formatted notes):

  1. an Excel-compatible spreadsheet, QuickSheet, which permits export in either CSV or Excel format (Excel is widely used in field linguistics as a convenient input tool for lexical databases, because of the ease with which databases may be constructed and restructed, and because it has many database-like functions, as well as built-in arithmetic functions if required for corpus work),
  2. the HanDBase DBMS which is also used for the metadata editor database,
  3. an implementation of the DATR lexicon knowledge representation language in LispMe, a Scheme implementation for the PalmOS platform (this application is a more Data Processing layer oriented tool, but is included in the Corpus Pilot layer implementation suite for convenience).

- Transcription support. In general, transcription in X-SAMPA (Gibbon et al., 2000) is used, but if required, IPA fonts may be used with the WordSmith word processor for PalmOS devices; RTF import and export facilities are available.

- Statistics package for initial evaluations. This is also a more Data Processing layer application, but integrated into the Corpus Pilot layer; functions include all the measures used in basic experimental and corpus work (including random sorting, mean, median, standard deviation, standard error, as well as standard pairwise comparison measures).

- Context-free parser package for basic grammar development. This is another Data Processing layer application, which is integrated into the Corpus Pilot layer because of the convenience of the LispMe Scheme application in which the parser suite is implemented.

The metadata application has been selected for detailed description, because it is most immediately relevant to the issue of language resources.

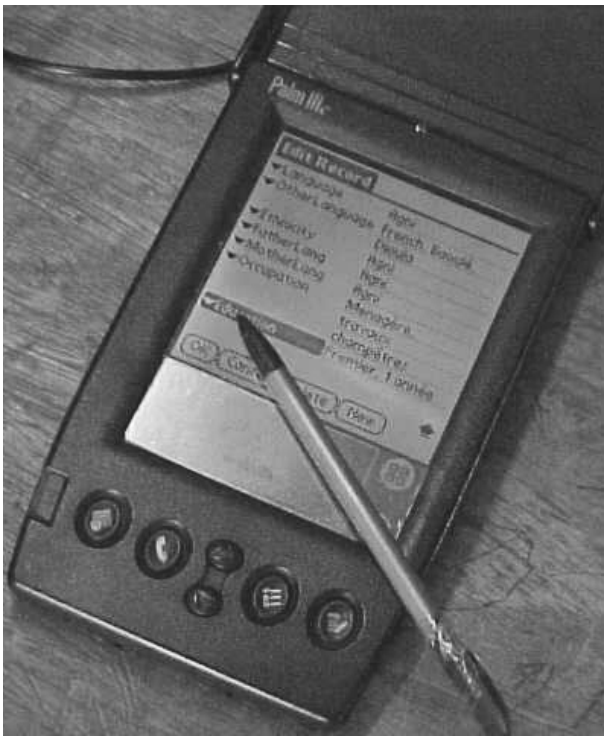

Figure 4: Palmtop metadata editor.

5. Metadata editor and database application

A metadata editor for audio/video recordings, photos, paper notes, artefact cataloguing was designed, based on a standard PalmOS relational database shell (HanDBase). The metadata editor provides a fast and inconspicuous input method for structured metadata for recordings and other field documentation, based on current work on metadata in the ISLE, E-MELD projects, and in the pilot phase of the DOBES project.

For the work in hand, standardised metadata specifications, such as the Dublin Core and IMDI sets, were taken into account. However, new resource types such as those which are characteristic of linguistic fieldwork demonstrate that the standards are still very much under development, since some of the standard metadata types are not relevant for the fieldwork data, and the fieldwork data types contain information not usually specified in metadata sets, but which are common in the characterisation of spoken language resource databases (Gibbon et al., 1997). In respect of the fieldwork resource type, it appears that it cannot be expected that a truly universal — or at least consensual — set of corpus metadata specifications will be developed in the near future, or perhaps at all, at a significant level of granularity. It may be possible to constrain the attribute list, though the existence of many different fieldwork questionnaire types belies this. However, the values of the attributes are in general unpredictable, entailing not only free string types but possibly unpredictable rendering types (e.g. different alphabets; scanned signatures of approval).

Indeed, it may be noted in passing that the expectation of fully standardising the entire metadata specification tends to reveal singularly little awareness of the potential of machine learning and text mining procedures for handling generalisation tasks of this kind. It may be predicted that such procedures will be applied in future not only to extensive resource data sets but also to increasingly extensive sets of metadata.

Table 1: Fieldwork metadata specifications.

Attribute	**Type**
RecordID:	string
LANGname(s):	popup: Agni,Agni; Ega
SILcode:	popup: ANY; DIE
Affiliation:	string
Lect:	string
Country:	popup: Côte d'Ivoire
ISO:	popup: CI
Continent:	popup: Africa; AmericaCentral; AmericaNorth; AmericaSouth; Asia; Australasia; Europe
LangNote:	longstring
SESSION:	popup: FieldIndoor; FieldOutdoor; Interview; Laboratory
SessionDate:	pick
SessionTime:	pick
SessionLocale:	string
Domain:	popup: Phonetics; Phonology; Morphology; Lexicon; Syntax; Text; Discourse; Gesture; Music; Situation
Genre:	Artefacts; Ceremony; Dialogue; Experiment-Perception; ExperimentProduction; History; Interview; Joke/riddle; Narrative; Questionnaire; Task
Part/Sex/Age:	string
Interviewers:	string
Recordist:	string
Media:	popup: Airflow; AnalogAudio; AnalogAV; AnalogStill; AnalogVideo; DigitalVideo; DigitalAudio; DigitalAV; DigitalStill; DigitalVideo; Laryngograph; Memory; Paper
Equipment:	longstring
SessionNote:	longstring

In consequence, the metadata specifications used in the UbiCorpus applications are deliberately opportunistic, in the sense that they are task-specific and freely extensible. A selection of attributes and values for the current fieldwork application are shown in Table 1. Metadata attributes concerned with the Resource Archive layer of archiving and property rights are omitted.

For current purposes, databases are exported in the attribute-value format shown below and converted into the TASX reference XML format (Milde and Gut, 2001). A specific example of the application of the metadata editor in the fieldwork session pictured in Figure 1 is shown in the exported record shown in Table 2.

The metadata editor and database application has been tested extensively in fieldwork on West African languages, and has proved to be an indispensable productivity tool, especially in difficult situations where very limited time is available.

6. Conclusion

Architectures using the first two levels, e.g. a server configuration and a laptop for use in the field, are very com-

Table 2: Fieldwork metadata example.

Attribute	Value
RecordID:	Agni2002a
LANGname(s):	Agni, Anyi
SILcode:	ANY
Affiliation:	Kwa/Tano
Lect:	Indni
Country:	Côte d'Ivoire
ISO:	CI
Continent:	Africa
LangNote:	
SESSION:	FieldIndoor
SessionDate:	11.3.02
SessionTime:	8:57
SessionLocale:	Adaou
Domain:	Syntax
Genre:	Questionnaire
Part/Sex/Age:	Kouamé Ama Bié f 35
Interviewers:	Adouakou
Recordist:	Salffner, Gibbon
Media:	Laryngograph
Equipment:	1) Audio: 2 channels, l laryngograph, r Sennheiser studio mike 2) Stills: Sony digital 3) Video: Panasonic digital (illustration of techniques)
SessionNote:	Adouakou phrases repeat

mon. However, in many situations the laptop concept is unsuitable because of heavy power requirements which are not available in many fieldwork locations. For these applications, the PalmOS based family constitutes the platform of choice because of minimal size and power requirements, permitting several weeks use on one charge or small battery. Although the PalmOS platform is obviously unsuitable for signal processing applications (such as time-aligned annotation) it is well-suited for logging, transcription and reference purposes.

The power of PDA miniature computing platforms as useful components of laboratory and office environments is often underestimated, and we demonstrate that a number of applications for which even a laptop is clumsy or unsuited for the developing field of computational ethnolinguistic fieldwork may be elegantly provided on the Palm PDA platform. The addition of a foldable keyboard further enhances the text handling capacity of the devices.

In the medium term, it will be possible to integrate the hybrid applications at the Corpus Pilot, Data Processing and Resource Archive levels into a corpus management environment which not only permits seamless dataflow and workflow, a goal already achieved, but also into a non-technical user-friendly prototype which may serve as the basis of a fieldwork management product implementation.

The UbiCorpus architecture has been used as the basic specification for different kinds of language documentation work in a variety of different projects. The Resource Archive layer was originally designed and implemented for web–based lexical database development in the VerbMobil project (Wahlster, 2000), funded by the German Federal Ministry of Education and Research (BMBF). The concept has been further developed theoretically and practically in connection with the projects *Theorie und Design multimodaler Lexika* funded by the German Research Council (DFG), *Enzyklopädie der Sprachen der Elfenbeinküste* funded by the German Academic Exchange Service (DAAD) and *Ega: a documentation model for an endangered Ivorian language* in the pilot phase of the DOBES funding programme of the Volkswagen Foundation.

In its local implementation, the current Resource Archive layer version also includes support for telecooperation and web-teaching. The Data Processing layer includes numerous applications which cannot be specified here. The Corpus Pilot layer as described in the present contribution has been informally but extensively field tested at a number of fieldwork locations, most recently in the framework of DAAD funded doctoral thesis work. It is planned to apply the field testing criteria defined in (Gibbon et al., 2000) to an extended implementation of the components of UbiCorpus model.

7. References

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, (33 (1,2)):23–60.

Dafydd Gibbon and Thorsten Trippel. 2002. Annotation driven concordancing: the pax toolkit. In *Proceedings of LREC 2002*. LREC.

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

Dafydd Gibbon, Inge Mertins, and Roger K. Moore, editors. 2000. *Handbook of Multimodal and Spoken Dialogue Systems, Resources, Terminology and Product Evaluation*. Kluwer Academic Publishers, Boston/Dordrecht/London.

Jan-Torsten Milde and Ulrike Gut. 2001. The TASX-engine: an XML-based corpus database for time aligned language data. In *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia. University of Pennsylvania.

Frank van Eynde and Dafydd Gibbon. 2000. *Lexicon Development for Speech and Language Processing*. Kluwer Academic Publishers, Dordrecht.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Verlag.