Multilingual Corpora at the Hamburg Centre for Language Corpora

Hanna Hedeland, Timm Lehmberg, Thomas Schmidt, Kai Wörner
Hamburger Zentrum für Sprachkorpora (HZSK)
Max Brauer-Allee 60
D-22765 Hamburg
E-mail: hanna.hedeland@uni-hamburg.de, timm.lehmberg@uni-hamburg.de, thomas.schmidt@uni-hamburg.de, kai.wörner@uni-hamburg.de

Abstract

We give an overview of the content and the technical background of a number of corpora which were developed in various projects of the Research Centre on Multilingualism (SFB 538) between 1999 and 2011 and which are now made available to the scientific community via the Hamburg Centre for Language Corpora.

Keywords: corpora, spoken language, multilingualism, digital infrastructures

1. Introduction

In this paper, we give an overview of the content and the technical background of a number of corpora which were developed in various projects of the Research Centre on Multilingualism (SFB 538) between 1999 and 2011 and which are now made available to the scientific community via the Hamburg Centre for Language Corpora.

Between 1999 and 2011, the Research Centre on Multilingualism (SFB 538) brought together researchers investigating various aspects of multilingualism focussing either on the language development of multilingual individuals, on communication in multilingual societies, or on diachronic change of languages in multilingual settings. Without exception, the projects of the Centre worked empirically, basing their analyses on corpora of spoken or written language. Over the years, an extensive and diverse data collection was thus built up consisting of language acquisition and attrition corpora, interpreting corpora, parallel (translation) corpora, corpora with a sociolinguistic design and historical corpora.

Since corpus creation, management and analysis were thus crucial to the work of the Research Centre, a project was set up in June 2000 with the aim of designing and implementing methods for the computer-assisted processing of multilingual language data. One major result of that project is EXMARaLDA, a system for setting up and analysing spoken language corpora (Schmidt & Wörner, 2009, Schmidt et al., this volume). The focus of this paper will be on the spoken language corpora of the Research Centre which were either created or curated with the help of EXMARaLDA.

2. Overview of corpora

As the list of resources in the appendix shows, altogether 31 resources constructed at the SFB 538 were transferred to the inventory of the Hamburg Centre for Language Corpora. 27 of these are spoken language corpora, 3 are corpora of modern written language, and one is a corpus of historical written language. More specifically, we are dealing with the following resource types:

- Language acquisition corpora which document the acquisition of two first languages or a second language. Most of these corpora are longitudinal studies of child language in different bilingual language combinations (German-French, German-Portuguese, German-Spanish, German-Turkish), but other corpus designs (e.g. cross-sectional studies) and other speaker types (e.g. adult learners or monolingual children) are also present.
- Language attrition corpora which document the development of a "weaker" language in adult bilinguals. Three different language combinations

(German-Polish, German-Italian, German-French) are involved.

- Interpreting corpora which document consecutive and simultaneous interpreting involving trained and ad-hoc interpreters for different language combinations (German-Portuguese, German-Turkish, German-Russian, German-Polish, German-Romanian) and in different settings (doctor-patient communication and expert discussion).
- Corpora with a sociolinguistic corpus design whose data are stratified according to biographic characteristics (e.g. age) of the speakers and/or their regional provenance. This comprises a corpus documenting Faroese-Danish bilingualism on the Faroese Islands and a corpus documenting the use of Catalan in different districts of Barcelona.
- Parallel and comparable corpora in which originals and translations of texts are aligned or which consist of original texts from specific genres in different languages.

The entirety of spoken language resources amounts to approximately 5500 transcriptions with approximately 5.5 million transcribed words (not counting secondary annotations).

3. Data model

The spoken language corpora, while sharing the common theme of multilingualism, are still highly heterogeneous with respect to many parameters. As far as their content is concerned, they do not only cover a spectrum of fourteen different languages, but also greatly differ with respect to the recorded discourse types (e.g. interviews, free conversation, expert discussion, classroom discourse, semi-controlled settings, and institutional discourse). Even more variation is to be found with respect to the research interests pursued with the help of the corpora and, consequently, the methodology used to record, transcribe and annotate the data. To begin with, either only audio or both video and audio data are recorded, depending on whether or not non-verbal behavior plays a role for analysis (as is the case, for example, for data of young children). As some projects focused their research on syntactic aspects of language, while others where interested in phonological properties or discourse structures, different systems where applied in transcribing (e.g. orthographic vs. phonetic transcription or complete vs. selective transcription) and annotating (e.g. prosodic annotations, annotation of code switches) the data.

The challenge in representing the corpora on a common technical basis was thus to find a degree of abstraction which, on the one hand, allows operations common to all resources (such as time alignment of transcription and media) to be carried out efficiently on a unified structure, but, on the other hand, also makes it possible to apply theory or resource specific functions (such as segmentation according to a specific model) to the data. A data model based on annotation graphs (Bird & Liberman, 2001), but supplemented with additional semantic specifications and structural constraints, turned out to be suitable for this task (Schmidt, 2005).

4. Data curation

The construction of a non-negligible part of the resources had been completed or started before EXMARaLDA was available as a working system. A number of legacy software tools (syncWriter, HIAT-DOS, LAPSUS, WordBase) was used for the construction of these corpora resulting in data for which there was hardly a chance of sustainable maintenance. The resources therefore had to be converted to EXMARaLDA in a laborious process described in detail in Schmidt & Bennöhr (2007).

From about 2003 onwards, all projects used EXMARaLDA or other compatible tools (e.g. Praat) for corpus construction. Although these resources were much easier to process once they had been completed, there was still a considerable amount of data curation to be done before they could be published. This involved various completeness and consistency checks on the transcription and annotation data and the construction of valid metadata descriptions for all parts of a resource.

5. Data dissemination

Completed resources are made available to interested users via the WWW[1] through several methods:

- A hypermedia representation of transcriptions, annotations, recording and metadata allows users to browse corpora online (see figure 1).

[1] http://www.corpora.uni-hamburg.de

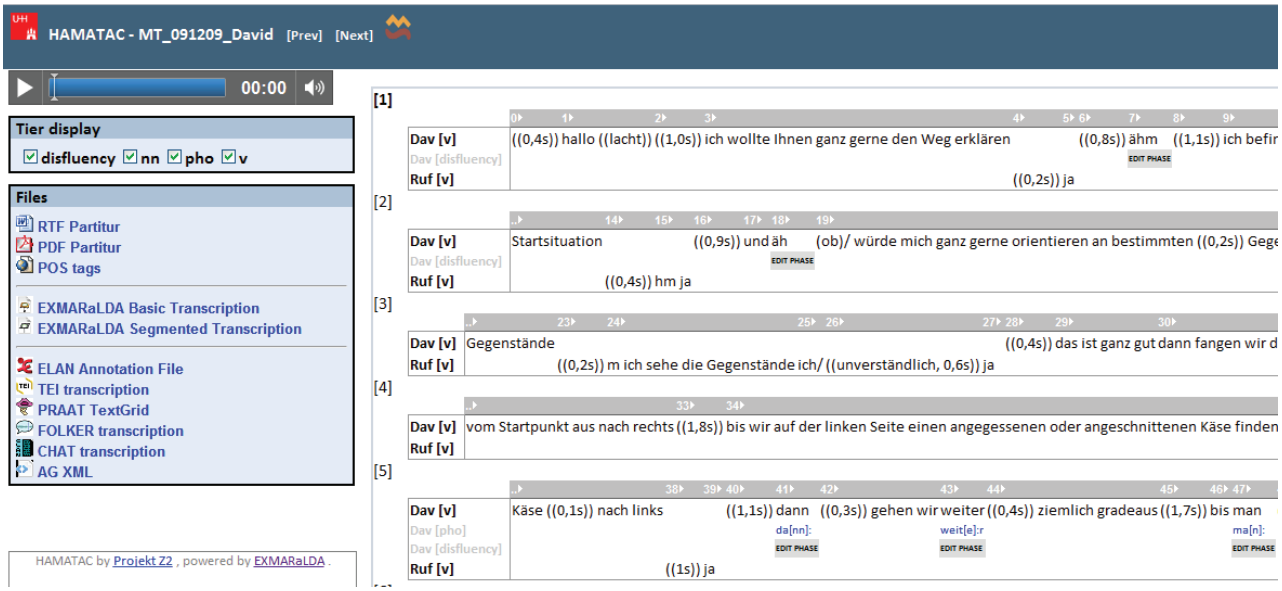

Figure 1: Hypermedia representation of a transcription from the Hamburg Map Task Corpus (HAMATAC)

- Resources can be downloaded in the EXMARaLDA format and then edited and queried with the system's tools (Partitur-Editor for editing transcriptions, Coma for editing and querying metadata, EXAKT for querying transcription and annotation data).
- Queries via EXAKT can also be carried out on remote data, i.e. without downloading the resource first, or through a web interface, i.e. without the need to install local software first.
- A number of export formats are offered for each annotation file making it possible to edit or query the data also with non-EXMARaLDA tools. Most importantly, most data are also available in the CHAT format of the CHILDES system, as ELAN annotation files, as Praat TextGrids and as TEI files.

Access to all corpora is password protected. The process for obtaining a password varies from resource to resource, but always requires the data owner's consent. Due to privacy protection issues, a part of the spoken resources can only be made accessible in the form of transcriptions, not audio or video recordings.

6. Future plans

In order to cater for the long term archiving and availability of the data beyond the finite funding period of the Research Centre, in January 2011 the Hamburg Centre for Language Corpora (HZSK, http://www.corpora.uni-hamburg.de) was set up. This institution is intended to provide a permanent basis not only for the corpora and tools referred to in this paper, but also for further resources existing or under construction at the University of Hamburg. The HZSK is part of the CLARIN-D network and will, in the years to come, integrate its resources into this infrastructure by providing protocols for metadata harvesting, assigning PIDs to resources, allowing for single-sign-on mechanisms and implementing interfaces as defined by CLARIN for access to metadata and annotations.

7. References

Bird, S., Liberman, M. (2001): A formal framework for linguistic annotation. In: Speech Communication (33), pp. 23-60.

Schmidt, T. (2005): Computergestützte Transkription - Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln. Frankfurt a. M.: Peter Lang.

Schmidt, T., Bennöhr, J. (2008): Rescuing Legacy Data. In: Language Documentation and Conservation (2), pp. 109-129.

Schmidt, T., Wörner, K. (2009): EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. In: Pragmatics 19(4), pp. 565-582.

Appendix: List of resources

Corpus name Project / Data Owner Type	Short description	Language(s)	Size
Spoken resources			
HABLA (Hamburg Adult Bilingual LAnguage) E11 / Tanja Kupisch spoken/audio/exmaralda	Audio recordings of semi-spontaneous interviews (elicited grammaticality judgments and production data are collected from the same speakers)	deu, fra, ita	169 communications 127 speakers 737797 transcribed words 169 transcriptions
DUFDE (Deutscher und Französischer doppelter Erstspracherwerb) E2 / Jürgen Meisel spoken/video/exmaralda	Video recordings (longitudinal study) of seven French-German bilingual children aged between 1 year;6 months and 6 years;11 months (+some later recordings).	deu, fra	562 communications 14 speakers ca. 1000000 transcribed words 849 transcriptions
BIPODE (Bilingualer Portugiesisch-Deutscher Erstpracherwerb) E2 / Jürgen Meisel spoken/video/exmaralda	Video recordings (longitudinal study) of three Portuguese-German bilingual children aged between 1 year;6 months and 5 years;6 months.	deu, por	250 communications 48 speakers ca. 250000 transcribed words 227 transcriptions
CHILD-L2 E2 / Jürgen Meisel spoken/video/exmaralda	Video recordings of children which start acquiring French or German as a second language at the age of three or four years.	deu, fra	181 communications 69 speakers 376114 transcribed words 181 transcriptions
ZISA (Zweitspracherwerb Italienischer und Spanischer Arbeiter) E2 / Jürgen Meisel spoken/audio/exmaralda	Recordings of adult L2-German-learners	deu	101 communications 5 speakers 119667 transcribed words 100 transcriptions
BUSDE (Baskischer und Spanischer doppelter Erstspracherwerb) E2 / Jürgen Meisel spoken/video/other	Longitudinal language aqcuisition study on bilingual Basque-Spanish children	eus, spa	*unknown*
PAIDUS (Parameterfixierung im Deutschen und Spanischen) E3 / Conxita Lleó spoken/audio/exmaralda	Audio recordings of monolingual children.	deu, spa	253 communications 66 speakers 166976 transcribed words 253 transcriptions
PHONBLA Longitudinalstudie Hamburg E3 / Conxita Lleó spoken/audio+video/exmaralda	Longitudinal data of Spanish/German bilingual children	deu, spa	413 communications 61 speakers 303792 transcribed words 413 transcriptions
PHONBLA Querschnittsstudie Madrid E3 / Conxita Lleó spoken/audio+video/exmaralda	Cross sectional study of bilingual German-Spanish L1 acquisition	deu, spa	113 communications 34 speakers 56722 transcribed words 113 transcriptions
PEDSES (Phonologie-Erwerb Deutsch-Spanisch als Erste Sprachen) E3 / Conxita Lleó spoken/audio/exmaralda	Longitudinal data of Spanish/German bilingual children	deu, spa	127 communications 21 speakers 101292 transcribed words 127 transcriptions
PHON-CL2 E3 / Conxita Lleó spoken/audio/exmaralda	Recordings of German subjects/children who have learned (or are learning) Spanish after the age of two	deu, spa	26 communications 22 speakers 17412 transcribed words 26 transcriptions
PHONMAS E3 / Conxita Lleó spoken/audio/exmaralda	Recordings of monolingual Spanish children (as comparable data for Madrid-PhonBLA)	spa	49 communications 4 speakers 3067 transcribed words 49 transcriptions
TÜ_DE-cL2-Korpus E4 / Monika Rothweiler spoken/video/exmaralda	Video recordings of (spontaneous and elicited language) of eight bilingual children with Turkish as their first language	deu	112 communications 19 speakers 348292 transcribed words 112 transcriptions
TÜ_DE-L1-Korpus E4 / Monika Rothweiler spoken/audio/exmaralda	Video recordings of (spontaneous and elicited language) of twelve bilingual children with Turkish as their first language	tur	12 communications 22 speakers 13 transcriptions

Rehbein-ENDFAS/Rehbein-SKOBI-Korpus E5 / Jochen Rehbein spoken/audio/exmaralda	Audio recordings of evocative field experiments with Turkish and German monolingual and Turkish/German bilingual children.	deu, tur	1017 communications 523 speakers 289012 transcribed words 836 transcriptions
ENDFAS/SKOBI Gold Standard E5 / Jochen Rehbein spoken/audio/exmaralda	Audio recordings of Turkish and German monolingual and Turkish/German bilingual children. Demo Excerpt from the larger Rehbein-ENDFAS/Rehbein-SKOBI-Korpus	deu, tur	3 communications 8 speakers 4862 transcribed words 3 transcriptions
Catalan in a bilingual context H6 / Conxita Lleó spoken/audio/exmaralda	Prompted, read and spontaneous speech data of Catalan speakers from Barcelona, stratified according to district and age of speakers	cat	225 communications 234 speakers 187967 transcribed words 875 transcriptions
Hamburg Corpus of Polish in Germany H8 / Bernhard Brehmer spoken/audio/exmaralda	Audio recordings of bilingual (Polish and German) and monolingual (Polish) adults (16-46 years). Recordings of semi-spontaneous data (3 topics) and renarration of a picture story (from 'Vater und Sohn')	pol	354 communications 94 speakers ca. 350000 transcribed words 358 transcriptions
Hamburg Corpus of Argentinean Spanish (HaCASpa) H9 / Christoph Gabriel spoken/audio/exmaralda	Recordings of spontaneous speech and laboratory data of speakers of Porteño Spanish in Argentina (read speech, story retelling, read question-answer pairs, intonation questionnaires, free interviews); 7 experiments altogether.	spa	259 communications 63 speakers 141321 transcribed words 261 transcriptions
Dolmetschen im Krankenhaus K2 / Kristin Bührig Bernd Meyer spoken/audio/exmaralda	Monolingual and interpreted doctor-patient communication in hospitals	deu, por, tur	91 communications 189 speakers 165689 transcribed words 92 transcriptions
SkandSemiko (Skandinavische Semikommunikation) K5 / Kurt Braunmüller spoken/audio/exmaralda	Radio recordings, recordings of group discussions and classroom discourse with speakers of two or more Scandinavian languages (Swedish, Danish, Norwegian) interacting.	dan, nor, swe	162 communications 515 speakers 269945 transcribed words 74 transcriptions
CoSi (Consecutive and Simultaneous Interpreting) K6 / Bernd Meyer spoken/audio+video/exmaralda	Recordings of simultaneously and consecutively interpreted lectures	deu, por	3 communications 8 speakers 35432 transcribed words 5 transcriptions
FADAC Hamburg (Faroese Danish Corpus Hamburg) K8 / Kurt Braunmüller spoken/audio/exmaralda	Recordings of semi-structured interviews in Faroese and Danish with bilingual speakers living on the Faroe Islands.	dan, fao	92 communications 82 speakers 440194 transcribed words 92 transcriptions
ALCEBLA T4 / Conxita Lleó spoken/audio/exmaralda	Recordings of Spanish-German bilingual children living in Germany and attending the Spanish complementary school at the first level	deu, spa	66 communications 23 speakers 36717 transcribed words 66 transcriptions
Simuliertes Dolmetschen im Krankenhaus T5 / Kristin Bührig, Bernd Meyer spoken/audio+video/exmaralda	Simulations of interpreted doctor-patient communication.	deu, pol, ron, rus	4 communications 12 speakers 4018 transcribed words 4 transcriptions
EXMARaLDA Demo Corpus Z2 / Hamburger Zentrum für Sprachkorpora spoken/audio+video/exmaralda	A selection of short audio and video recordings in different languages for demonstration of the EXMARaLDA system	deu, eng, fra, ita, nor, pol, spa, swe, tur, vie	19 communications 50 speakers 11659 transcribed words 19 transcriptions
Hamburg Map Task Corpus Z2 / Hamburger Zentrum für Sprachkorpora spoken/audio/exmaralda	Audio recordings of map tasks with advanced learners of German	deu	24 communications 26 speakers 24409 transcribed words 24 transcriptions

Written resources			
HaCOSSA (Hamburg Corpus of Old Swedish with Syntactic Annotations) H3 / Kurt Braunmüller written/tei	Bible translations, religious and secular prose, law texts, non-fiction literature (geographical, theological, historic, natural science), diploma.	dan, deu, isl, lat, nob, swe	35 texts
Covert translation: popular science K4 / Juliane House written/tei	Translation corpora of original texts with translations and comparable texts from the genre popular scientific prose	deu, eng	114 texts 500446 words
Covert Translation: business communication (old) K4 / Juliane House written/tei	Translation corpora of original texts with translations and comparable texts from the genre external business communication	deu, eng	119 texts 169154 words
Covert Translation: business communication (new) K4 / Juliane House written/tei	Translation corpora of original texts with translations and comparable texts from the genre external business communication	deu, eng	198 texts