LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 2, No. 1 (June 2008), pp. 109-129
http://nflrc.hawaii.edu/ldc/

Rescuing Legacy Data

Thomas Schmidt and Jasmine Bennöhr

University of Hamburg

This paper discusses issues that arise in the transformation of electronic language data from outdated to modern, sustainable formats. We first describe the problem and then present four different cases in which corpora of spoken language were converted from legacy formats to an XML-based representation. For each of the four cases, we describe the conversion workflow and discuss the difficulties that we had to overcome. Based on this experience, we formulate some more general observations about transforming legacy data and conclude with a set of best practice recommendations for a more sustainable handling of language corpora.

1. INTRODUCTION. Starting in the late 1970s or early 1980s, more and more linguists discovered the computer as an aid in constructing and analysing large collections of empirical data. At that time, little specialized software and no widely-accepted standards for building such corpora were available, so that solutions were typically self-tailored to the specific task and hardware environment at hand. Moreover, the idea that data could be shared and exchanged over networks had not yet gained ground; compatibility issues among different software tools, operating systems, and hardware configurations were therefore not considered as important as they are today. The data resources that were constructed under these circumstances are what are called today "legacy data" – data that would be valuable for current and future research, but the (re)use of which is made difficult or impossible through their dependence on an outdated piece of software or operating system. More often than not, a lack of documentation (of corpus content, of data formats, or of other specifics of the data structure) further aggravates these problems. In the long run, it thus seems certain that legacy data will either be lost or must be transformed into a "more modern" form. Thanks to recent developments in corpus technology and to increasingly fruitful standardization efforts, it is relatively clear what such a modern form should ideally look like. The best practice recommendations by Bird and Simons (2003) give a comprehensive characterization of requirements which need to be met to make a corpus "portable," i.e., easy to reuse and archive. In our understanding, the most important of these requirements are:

- the use of open standards for the digital encoding of corpus data,
- a content-oriented, rather than presentation-oriented, markup of textual data, and
- a careful and comprehensive compilation of corpus metadata.

This paper describes work carried out at the Research Centre on Multilingualism at the University of Hamburg.[1] With the ultimate aim of creating an archive of multilingual

[1] The work described in this paper was carried out with the help of a research grant from the German Research Council (DFG). The authors are part of a project that is concerned with technological and methodological issues in corpus construction and use. The actual data that we write about come

E-ISSN 1934-5275

language data, several corpora of legacy data, which in one or several ways failed to meet the above-mentioned requirements, were transformed into the EXMARaLDA format, an XML-based format for the construction and analysis of spoken language corpora, which *does* meet them (see Schmidt 2005 or Schmidt and Wörner 2007).[2] In the following section, we describe the original form of each of these legacy corpora and the problems this form posed for data reuse. We then outline the process that we applied in order to transform the corpora to EXMARaLDA. We think that the details of this transformation process are of interest to a larger audience insofar as they exemplify what kind of problems may arise in a legacy data rescue effort and how laborious the solutions are.[3] That is, we do not claim to describe a general method for this task, nor do we believe that one exists. What steps need to be taken depends very much on the individual characteristics of the legacy data in question and consequently can only be decided on a case-by-case basis. However, we found that some more general observations can be made which, on the one hand, can help other owners of legacy data to determine how easy or difficult transforming their data might be,[4] and which, on the other hand, illustrate (once more) how important it is to follow certain rules of best practice in the construction of corpora. These observations are summarised in sections 3 and 4.

2. LEGACY DATA CONVERSION: FOUR EXAMPLES.

2.1 THE DUFDE, BUSDE AND BIPODE CORPORA. The DUFDE, BUSDE, and BIPODE corpora were created mainly for the investigation of grammar acquisition of bilingual children, and differ with respect to size and to the languages involved. For each child, monthly

not from this project, but from other projects at the Research Centre on Multilingualism or their predecessor projects. We would therefore like to thank the researchers who conducted these projects and did the original construction of the corpora for their cooperation: Jürgen Meisel (the DUFDE, BUSDE and BIPODE corpora), Conxita Lléo (the PAIDUS corpus), Jochen Rehbein (the ENDFAS/SKOBI corpus), and Kurt Braunmüller (the Scandinavian corpus). We are also grateful to a large number of research and student assistants, without whose help we could not have done the work described in this paper: Matthias Bonnesen and Madeleine Turcaud (the DUFDE corpus); Imme Kuchenbrandt (the PAIDUS corpus); Annette Herkenrath, Tülay Selçuk, Nurkan Darıcalı, Hatice Yıldırım, Tuba Özcan, Seçil Yusun, Eylem Şentürk, Ezel Babur, Nesrin Esen, Rasim Aksoy, Seçil Icelliler, and Kimberly Aidin (the ENDFAS/SKOBI corpus); Ludger Zeevaert, Hanna Hedeland, and Frank Stinner (the Scandinavian corpus). We would also like to thank two anonymous reviewers for their valuable comments and Sönke Häseler for proofreading.

[2] EXMARaLDA's data model follows the general approach of the annotation graph framework (Bird and Liberman 2001).

[3] It should be noted that, although legacy data conversion is done in many places, there is, to our knowledge, not a single publication which describes this process.

[4] While the techniques we describe here have been applied only to corpora of widely used languages, we think that the issues we encountered are of equal, or even greater, importance to the portability of endangered languages data (and hence to the readers of this journal).

recordings were made in both languages, generally covering a period of four years up to the age of five.

- The DUFDE (*Deutsch und Französisch: Doppelter Erstspracherwerb*) corpus contains more than 1,000 transcripts and the corresponding video material of interviews with seven bilingual children between 1981 and 1991 in German and French (see Köppe 1994).
- The BUSDE (*Baskisch und Spanisch: Doppelter Erstspracherwerb*) corpus aims at the study of simultaneous language acquisition of Basque and Spanish. Three children provided the basis for 350 recordings altogether between 1994 and 2002.
- The BIPODE (*Bilingual Portugiesisch-Deutsch*) corpus comprises a total of 450 interviews with three bilingual children in Portuguese and German and was compiled between 1997 and 2004.

Each transcript includes a table with several columns (see figure 1 below). The most important of these are the name (information about the transcription, including a short version of the child's name, the transcription number, and the language), the sentence number,[5] the sentence (the child's utterance), the context (the utterance of the interviewer), and a comment (a description of the situation).

name	sentence no.	sentence	context	comment
Fr18D	0	Eingabe: Jérôme R. April 2003	Interviewerin: Gisela	
Fr18D	2	die hier - passt (x)	passt das da hin?_	sitzt mit G auf dem Boden
Fr18D	3	(x) hier [pe:se] (=Pilse)	passt hier?_	nimmt aus dem Deckel des Puzzles ein Teil
Fr18D	4	(ho-he-hör) (=wo gehört?) (das denn-)	pilse?_	
Fr18D	6	ho he-hör da de hin? (=wo gehört das denn hin?)		Puzzle-Teil
Fr18D	7	da?	_oh du kannst das ja schon alles	legt das Puzzle-Teil an die richtige Stelle

FIGURE 1: Typical Transcript from DUFDE

Additionally, the DUFDE transcripts contain a column for the age of the child at the time of recording. The transcripts of the other corpora contain up to 160 additional columns for coding data. However, since coding of a specific phenomenon was usually done only for a small portion of the data, most of these additional cells are empty.

[5] Here we can observe an instance of "tag-abuse" which was not documented but consistent and easy to decipher: In this project, the sentence number "0" was obviously used to designate metadata within transcriptions.

For each child we have metadata (see figure 2 for an example) containing information such as the number of the recording, whether the recordings are stored on old or on new video tapes, the date (of the recording), the age (of the child at the time of the recording), a comment, whether and where paper transcripts exist, whether they have been digitized, and if so, where to find the files.

***name of child*,** date of birth *xx.xx.xx* last modification**: 03.01.2006**

VMP: Von-Melle-Park 6, 5th, 6th or 11th floor
MBA: Max-Brauer-Allee 60, Room 015/016
transcripts: K: copy; **O**: original
old: recording is on old video tapes
new: recording is on new video tapes

recording	video tapes			date	age	comment	transcripts		entry	
	no.	old	new				VMP	MBA	MBA	VMP
701		?	**VMP**	26.11.84	1;10;03	+JP 3;09;11				
702		VMP		08.01.85	1;11;14		K d/f	O d/f	d/f	
703		VMP		22.01.85	1;11;28	without sound!				

FIGURE 2: Metadata for the DUFDE corpus

Before the data were converted to the EXMARaLDA format, a great deal of analysis was necessary in order to find patterns and identify peculiarities in the data. This would not have been possible without the assistance and knowledge of people who still worked on a successor project. Only late in the conversion process did we succeed in retrieving a paper version of the project-internal transcription conventions of the DUFDE project from December 1987, which are based on those of Pienemann (1981:187 ff), the ZISA project (Meisel, Clashen, and Pieemann 1981), and Bloom and Lahey (1978:605-609). The conventions prescribe in detail how to transcribe utterances, but do not (and could not, because the transcriptions were made on paper at the time) provide any information regarding the structure of, and the dependencies between, the digital files. The original transcription conventions did not seem to be known to newer transcribers; the document had never been digitized, but was instead stored in a different building from where the transcription was done. Despite these obstacles, the transcription process seems to have changed only very little since 1987, which must be due to careful passing on of knowledge by word-of-mouth. The transcription conventions of the BIPODE and BUSDE projects still remain untraceable.

Yet, even if all the necessary information about the creation and transcription of the corpus is available, the data may still pose a number of problems. Data that are easily readable and understandable for humans, but which are inconsistent, will likely resist any

attempt at automatic processing or at least cause a considerable amount of unexpected and unwelcome additional work. As illustrated in figure 3, the problems we faced can be classified in three major categories.

1. Inconsistent naming of categories.
    a. by using translations or similar words in addition to the proper category name (*marked with an ellipsis below*),
    b. by using abbreviations in addition to the long form of the category name (*trapezoid*),
    c. through typos in the category name (*triangle*).
2. Use of different symbols for the same function, e.g., semicolon and colon used interchangeably to separate months from days in age information (*rectangle*).
3. Swapping, omitting, or inserting columns (*trapezoid*).

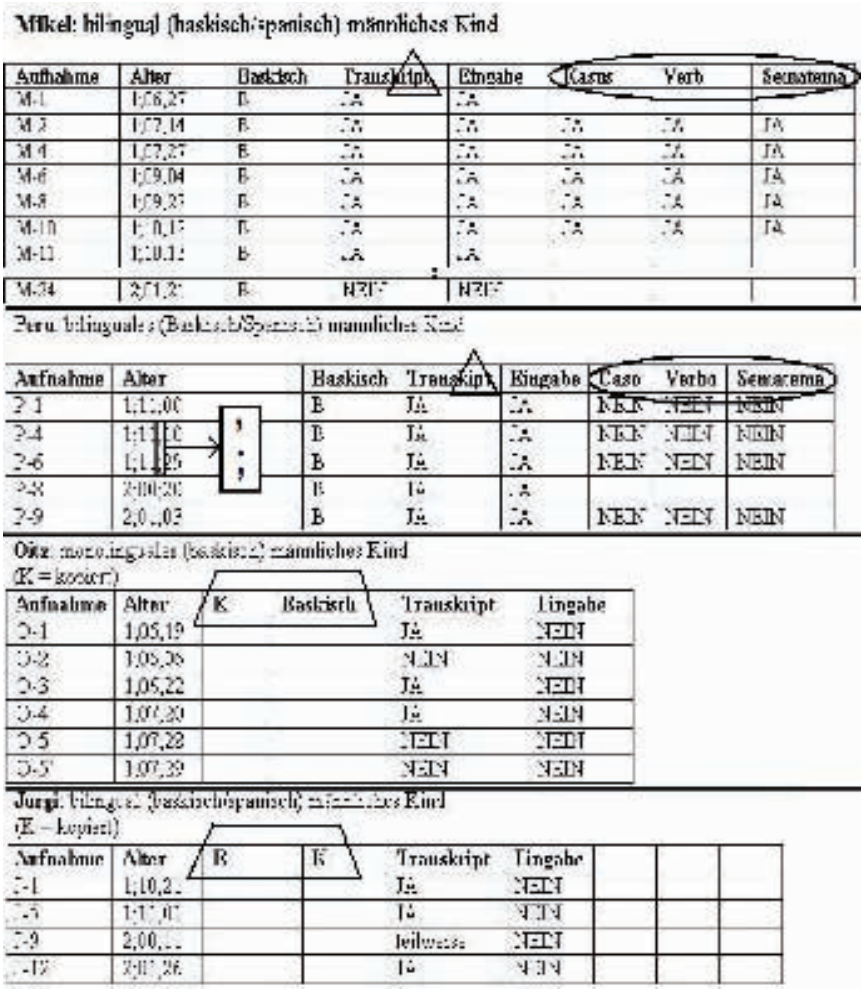

FIGURE 3: Inconsistencies in the BUSDE data

Many inconsistencies are easily taken into account and can be fixed in the conversion process if they are known. But as inconsistencies are usually introduced unintentionally into the data and are likely to be more frequent when automatic processing at a later stage is not considered at the time the data are created, a large amount of data had to be checked for irregularities. Thus, the tilde accompanying letters such as "n" in Portuguese proved to be quite a nuisance. When the recordings were transcribed, owing to technical reasons, the tilde could not be placed above the letter to which it refers, but was instead inserted either before or after the letter. This ambiguity could only be resolved manually.

Transformation steps

1. Data from the DUFDE corpus were available only in proprietary formats. Depending on when the data were recorded and transcribed, the transcriptions were stored in DBASE, EXCEL, or ACCESS files. The very problem of inaccessibility – the prevention of which was one purpose of this conversion project – already turned up during our work: The old DBASE files could not be opened with any text editor or any common spreadsheet software except Open Office. Once the problem of accessibility was solved, the transcription data were exported from the DBASE, EXCEL, or ACCESS file and split into separate transcription files as character-separated text files.

2. The metadata for each child had been stored in separate Word files, and it was not directly clear which metadata belonged to which transcript. The varying number and order of columns had to be manually corrected and the data had to be exported to a character-separated text file to create a document that could be processed automatically. Once we established that the "name" field in the transcriptions was systematically linked with the recording number of the metadata file, it was possible to automatically extract metadata and combine it with the corresponding transcription files, resulting in EXMARaLDA transcripts with meta-information directly attached to them.

3. A considerable amount of automatic post-editing was carried out on the EXMARaLDA files. Most important, it was decided to delete a large portion of the coding data, since most of it had been done only for a small part of the corpus, and knowledge about the details of the coding schemes had been lost.

4. Additional metadata about speakers and recordings were transferred from paper to a digital form and linked to the transcriptions.

5. The video recordings from the original VHS tapes are currently being digitized, using Pinnacle studio for the digitization itself and Auto Gordian Knot for video compression. The AVI wrapper format and the XVID codec are used as a target format, resulting in file sizes of roughly 200 MB for a thirty-minute recording.

6. Finally, the transcription files are synchronized with the digitized video recordings, using the EXMARaLDA Partitur-Editor.

Besides preserving precious language data for future researchers, the conversion of these three corpora yielded some potentially valuable by-products. For instance, one very positive outcome of the close cooperation with a member of the DUFDE successor project is that new transcriptions in the project will now be directly entered into EXMARaLDA. Moreover, the conversion routines we developed may be directly applicable to other data that are stored in spreadsheet format.

2.2 THE PAIDUS CORPUS. In the PAIDUS (Parameterfixierung im Deutschen und Spanischen [Parameter Fixing in German and Spanish], 1992-1995) project (Lleó et al. 1996), longitudinal data from ten German and Spanish children between ages 9 months and 3 years were recorded to investigate the fixing of phonetic and prosodic parameters in German and Spanish. Altogether, there are roughly 250 audio recordings, amounting to about 130 hours of recorded conversation.

The children's utterances in these recordings were transcribed phonetically, and an orthographic transliteration was added. Further annotations included an utterance number, an orthographic transcription of related adult participants' utterances, a characterization of the utterance type (spontaneous vs. imitation), and the utterance's syllable structure. These additional annotations were, however, not provided for every utterance. Figure 4 is a typical example of such a data set.

Phonetic	Orthographic	Number	Adult	Type	Syllable structure
[dʊ̯*ɐʊ̯.*ˈmɐ.mɐ]	Du auch, Mama.	11		spontaneous	[CG*VG.*CV.CV]

FIGURE 4: A typical data set from the PAIDUS corpus

Data were entered and stored in tables of a 4th-Dimension database (see http://www.de.4d.com/) for the Macintosh, using the WORDBASE application developed by the Max Planck Institute for Psycholinguistics in Nijmegen. In the database tables, the phonetic transcription was represented as a string in 8-bit MacRoman encoding, which would display as the desired IPA symbols when used with a certain IPA font provided by the Summer Institute of Linguistics (SIL, see http://www.sil.org/).

When the PAIDUS data were to be reused in a later project (Prosodic Constraints on Phonological and Morphological development in Bilingual First Language Acquisition, see http://www.uni-hamburg.de/fachbereiche-einrichtungen/sfb538/projekte3_e.html), a number of problems were encountered: First, there were serious stability and capacity problems of the 4th-Dimension database, leading to frequent crashes of the software. Second, the idiosyncratic structure and format of the database severely limited the possibilities of extending the existing data or exchanging them with other applications. This, in turn, made it difficult to integrate the PAIDUS data with newly acquired data. Third, because of the limited programmability of the database, certain data processing steps (like the calculation of syllable structure) had to be done manually, although they could in principle have been entirely automated. It was therefore decided to transform the PAIDUS corpus into the EXMARaLDA format. This transformation consisted of four steps:

1. The recordings on the analogue audio tapes were digitized using Audacity (http://audacity.sourceforge.net/). Since the data are intended for acoustic analysis, the best possible quality (highest sample rate, uncompressed encoding in WAV format) was chosen.

2. After some basic preprocessing, the original transcription data were exported from the database table as tabulator-separated text files. Using a conversion routine written in Java, these text files were then transformed into a simple XML-format. In this process, the original MacRoman encoding was changed into a Unicode-based UTF-8 encoding, including a mapping of the phonetic transcription symbols to the corresponding Unicode symbols so that future processing would not depend on a specialized font.

```
<?xml version="1.0" encoding="UTF-8"?>
<data-set imSpont="spontaneous" aeusserungsNr="11">
    <orthographie> Du auch, Mama.</orthographie>
    <phonKind>[dʊ*ɐʊ̯.*ˈmɛ.mɐ]</phonKind>
    <syllStructure>[CG*VG.*CV.CV]</syllStructure>
</data-set>
```

FIGURE 5: XML representation of PAIDUS data

3. An editor with a graphical user interface was written in JAVA to enable student assistants to correct errors and add missing information in individual data sets manually. Most important, this step included bringing utterances into the order in which they occurred in the transcribed conversation – a piece of information which the original database did not include.

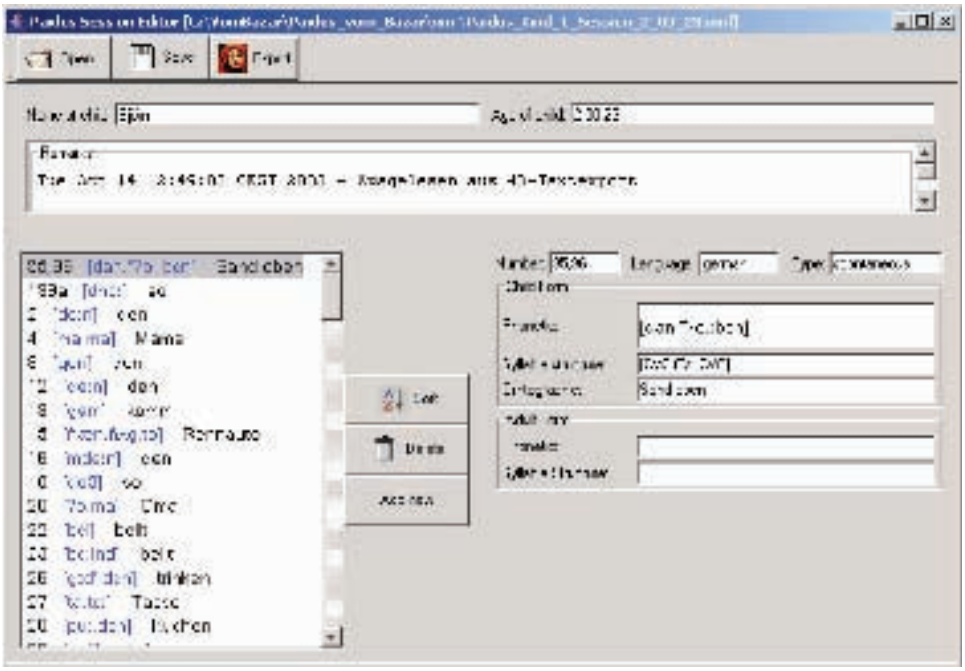

FIGURE 6: Screenshot of the PAIDUS session editor

4. Finally, these manually post-edited files were automatically converted into EXMARaLDA Basic-Transcriptions. All further processing could then be done with the help of the EXMARaLDA Partitur-Editor. This included:
    a. adding transcription metadata,
    b. adding adult participants' utterances where they were missing,
    c. synchronizing the transcription with the digitized recording, i.e., linking points in the transcription to the corresponding position in the audio file, and
    d. adding syllable structure annotations where they were missing. For this task, a finite state transducer was written which maps IPA transcription symbols to their corresponding syllable structure representations.

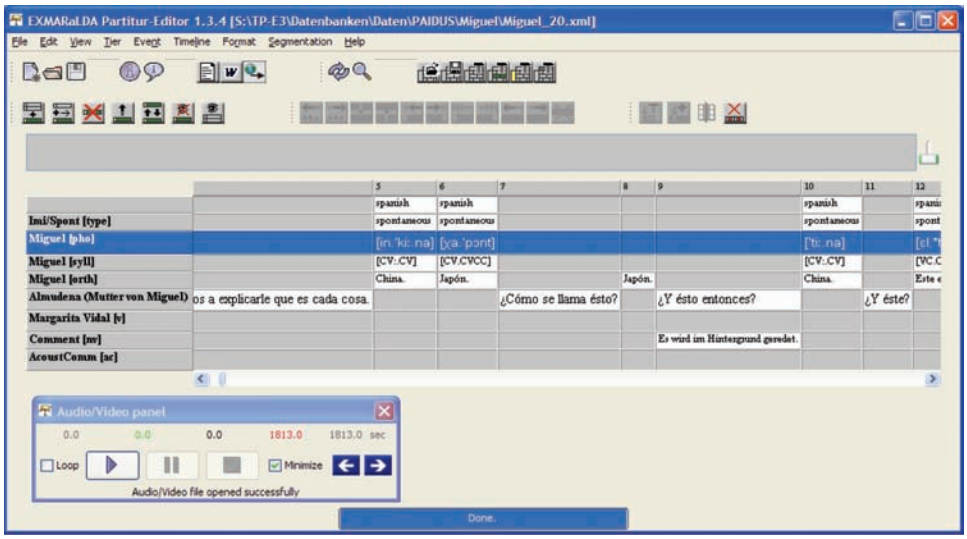

FIGURE 7: PAIDUS data in the EXMARaLDA Partitur-Editor

2.3 THE REHBEIN-ENDFAS AND REHBEIN-SKOBI CORPORA. The projects ENDFAS (Entwicklung narrativer Diskursfähigkeiten im Deutschen und Türkischen in Familie und Schule, 1989-1995) and SKOBI (Sprachliche Konnektivität bei Türkisch-Deutsch Bilingualen Kindern, 1999-2006) both collected recordings of Turkish-German bilinguals and Turkish monolinguals in different communicational settings (see, for example, Rehbein 2007, Herkenrath 2007, and Karakoç 2007). The total number of recordings amounts to about 1,000 audio tapes, corresponding to roughly 1,000 hours of recorded conversation. Large excerpts from 233 of these recordings have been transcribed, yielding a total of 834 transcriptions.

Transcription was done following a modified version of the HIAT conventions (Rehbein et al. 1993) with the help of syncWriter. SyncWriter is a software tool for Macintosh systems (OS versions 7.x to 9.x) which supports the editing of texts in interlinear notation. It was developed around 1990 and distributed commercially by a German software company (med-i-bit). Development came to a standstill sometime during the 1990s and med-i-bit stopped distributing and supporting the software around 2002.

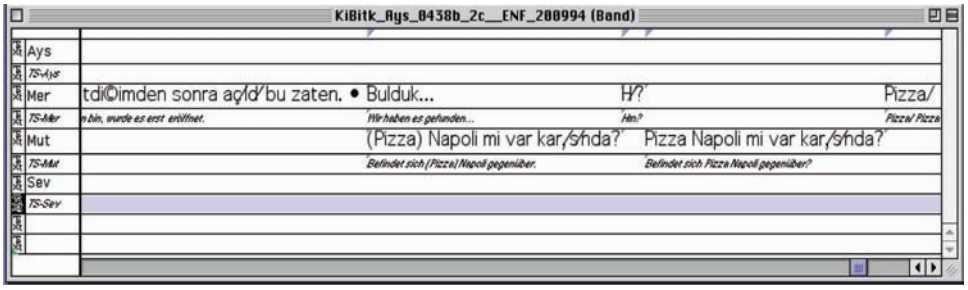

FIGURE 8: syncWriter interface

The problems that the users of the ENDFAS and SKOBI corpora were consequentially faced with are symptomatic of the legacy data issue.

- syncWriter uses a largely undocumented binary format. The software itself does not provide a lossless export, and no other software has an import facility for syncWriter data; syncWriter data can therefore only be processed by syncWriter itself;
- the syncWriter source code is not available, making a reuptake of the development or even a simple code inspection for the specifics of the data format impossible;
- syncWriter is dependent on specific versions of the Macintosh operating system. With the advent of MAC OS X and Apple's dwindling support for older MAC OS versions, it becomes more and more likely that the software will cease to be executable on future machines.
- In order to be able to represent Turkish extensions of the Latin alphabet (like *ğ* or *š*) and other special characters needed in transcription, the project used MacRoman character encoding together with a custom font (called HIAT Times) which was, however, completely undocumented and not officially available for users outside the project.

The combination of these factors made it very likely that large amounts of data created in the two projects would be completely lost to future generations of researchers. Once the hope of motivating the original syncWriter developer to participate in a "rescue effort" had been given up, a solution was found through educated guesswork, based on pieces of information gathered from people who had been associated with the syncWriter develop-

ment and from (incomplete and unpublished) bits of documentation. The solution consisted of a four step process:

1. Using the syncWriter software, some pre-processing steps were performed to make the data more consistent.
2. Based on a hint that syncWriter is AppleScript-enabled, a trial-and-error method was employed to write an AppleScript routine that reads out data from a running syncWriter instance and transforms it into a preliminary XML format.[6]
3. A conversion filter was written to transform this preliminary XML format into an EXMARaLDA Basic-Transcription. In that process, the character encoding was also changed to UTF-8 (again, some trial-and-error was involved in getting the mapping from the undocumented custom encoding right).

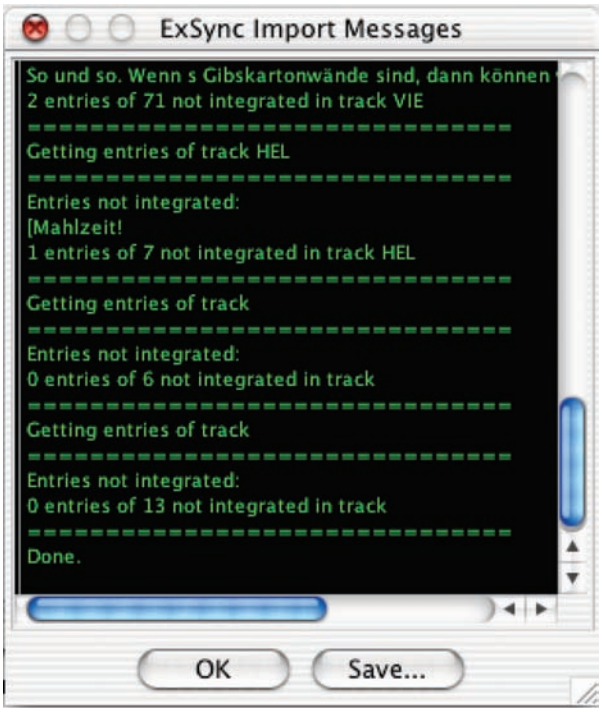

FIGURE 9: Result of the sycnWriter import

4. Using the EXMARaLDA Partitur-Editor, extensive post-editing was carried out on the transcriptions. This involved setting the end points of utterances, reconstructing accent markers from the original file (because these were lost in the conversion) and ensuring consistency with respect to the transcription conventions.

[6] The AppleScript is downloadable from http://www.exmaralda.org.

At the same time, the analogue audio tapes were digitized using Audacity. The process of aligning these digitized recordings with the transcriptions is now nearing completion.

Besides the transcriptions and the audio, there was also extensive metadata for both corpora giving biographic details about the speakers (such as age, family background, and languages) and about the communication settings (such as time and location, a summary of the discourse, etc.). These had been entered into a FileMaker database according to the HcTT (Hamburger computergestützter Transkriptionsthesaurus; see Gerhard et al. 1997) metadata schema. Two student assistants were instructed to transfer the data from that database into an XML-based format for representing corpus metadata (as defined by the EXMARaLDA Corpus-Manager, CoMa, see Schmidt and Wörner 2007) using Copy&Paste. In that process, a number of errors were corrected, and additional metadata was added that had been recorded, not in the database, but in separate text files.

2.4 THE CORPUS "SCANDINAVIAN SEMI-COMMUNICATION." For the project "Semicommunication and receptive multilingualism in Scandinavia" (1999–2005, see Zeevaert 2004 or Golinski 2007), about 90 hours of audio recordings were made of communications (group discussions, radio interviews, school lessons) in which Danish, Swedish, and Norwegian native speakers interacted, each using his or her own mother tongue and his or her receptive competence in the other Scandinavian languages. About half of these recordings were transcribed following the transcription guidelines of the HIAT system (Ehlich and Rehbein 1976) for functional-pragmatic discourse analysis.

In the first project phase, transcription was carried out using the software HIAT-DOS (Schneider 2001). HIAT-DOS is an MS-DOS-based software that supports the creation of transcriptions in partitur (musical score) notation. Alignment of transcribed text on different tiers is achieved by using an equidistant font (Courier). For printout, the whole transcription is broken up into several score areas.

FIGURE 10: HIAT-DOS interface

Working with HIAT-DOS turned out to be problematic in many respects:

- Its interface (text-based, no menu or mouse support) is not up to current standards of user-friendliness; student assistants had great difficulty in learning and using the software efficiently.
- It offers no support for Scandinavian extensions of the Latin alphabet (like *å* or *ø*); more generally, its limitation to an 8-bit-codepage makes it badly suited for multilingual transcription.
- It is severely limited with respect to the maximum number of tiers as well as the total file size; transcriptions of recordings of over ten minutes in length or with many speakers had to be split up into several files.
- Local editing often led to undesired global changes in the transcription structure which transcribers found difficult to undo.

In the second phase, it was therefore decided to abandon HIAT-DOS and use Praat (Boersma and Weenik 1996) instead. At the end of the project, the data from both phases had to be unified to yield a single corpus. Again, EXMARaLDA was chosen as the target format.

For the Praat transcriptions, the conversion was relatively easy. Praat uses a text-based format in which the transcription is structured into a number of tiers, and each tier contains portions of text linked to the recording via a start and end offset. A conversion filter was written in JAVA, which parses a Praat file, transforms it into an EXMARaLDA Basic-Transcription, and outputs the latter as an XML file. Apart from the addition of some transcription metadata (e.g., assignment of tiers to speakers), no post-processing of the converted Praat files was necessary.

For the HIAT-DOS transcriptions, the conversion process was more complex. HIAT-DOS also uses a text-based format, but this represents the layout of the musical score display rather than directly encoding the structural relations in the transcription. The conversion filter (again written in JAVA) therefore attempts to translate the graphic relations of the musical score display into temporal relations of EXMARaLDA's time-based data model. Different parameters can be set for this calculation, and their optimal configuration depends on the details of the individual transcription. A student assistant was instructed to determine by trial-and-error which parameter combination to use for a given file, to perform the conversion, and then to manually post-edit the resulting file in the EXMARaLDA Partitur-Editor.

Metadata for the corpus was almost completely non-digital. For a part of the corpus, recorded speakers had filled in questionnaires about their language skills and attitudes. For another part, almost no metadata had been recorded, but could in part be reconstructed from external sources (e.g., by looking up information about a radio announcer on the web) or from the transcribed discourse itself (i.e., by making use of people's self-introductions in a classroom). The entirety of this metadata was then entered with the help of the EXMARaLDA Corpus-Manager and linked with the corresponding transcriptions.

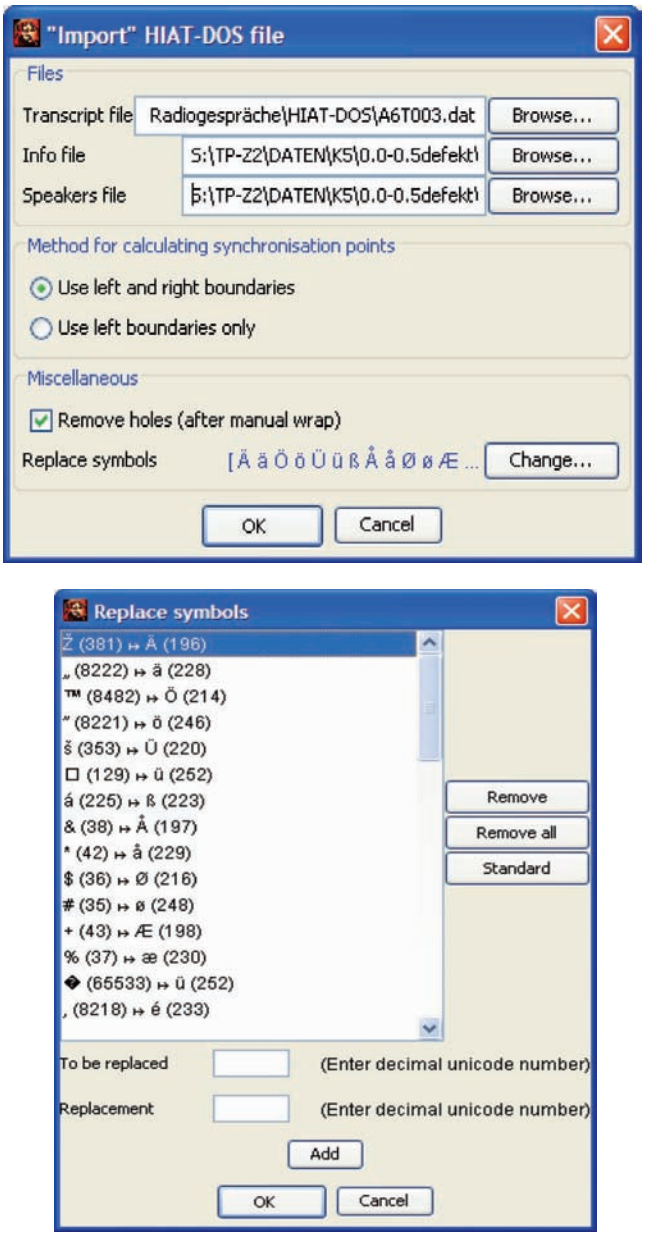

FIGURE 11: Setting parameters fort he HIAT-DOS conversion

2.5 TIME AND EFFORT. As the previous sections indicate, processing the corpora was labor intensive and required many hours of research and student assistants' work. The fol-

lowing is an attempt to estimate[7] the total amount of time invested, broken down, as far as possible, into the different processing steps.

- DUFDE data: initial analysis of the data, semi-automatic correction of inconsistencies, programming of the conversion routines, and checking the conversion results were done by a research assistant with a background in computational linguistics. The whole process took about one year with approximately 500 working hours spent on the task. Another 50 hours was invested by another research assistant for automatic post-processing of the converted data. A student assistant is now taking care of the digitization of video and metadata. With approximately 600 working hours, she has completed her work for about two-thirds of the data.
- PAIDUS data: approximately 150 hours was spent by a research assistant for initial analysis and conversion of the transcription data. Three student assistants, each with contracts for about 600 hours, did the manual post-processing and synchronization of the transcription files.
- ENDFAS/SKOBI data: initial analysis of the data, writing a conversion routine, and designing the workflow for corpus conversion and editing was done by a research assistant and took about five months or 400 working hours. Post-processing the transcriptions, digitizing the recordings, entering and editing metadata, and synchronizing transcriptions with recordings was done by a team of ten student assistants with contracts for 6,000 hours altogether. Owing to the size of the team, much time had to be invested coordinating the work (i.e., briefing and training new team members, performing quality and consistency checks, carrying out administrative tasks etc.) by a research assistant.
- Scandinavian data: initial analysis and writing the conversion routines took a research assistant about a month or 150 working hours. Post-processing of the older HIAT-DOS data was done by a student assistant in roughly 150 hours. Post-processing the Praat data, compiling corpus metadata, and adding or correcting synchronization took about 500 hours of another student assistant's work.

3. OBSERVATIONS. Despite the numerous difficulties we encountered, our efforts to rescue legacy data can be considered successful in all four cases. We can therefore state that, in principle, there seem to be no obstacles to transforming older corpora of language data into a form which is up to current standards of sustainable data handling. Considering that the legacy corpora described in the previous section cover a broad spectrum in terms of transcription and media formats, we believe that this observation should be transferable to most other legacy data. However, our experience also clearly shows that legacy data conversion is not a trivial task, and, more important, it is a task requiring a considerable

[7] These are, of course, rough estimates, because we do not use any sophisticated instruments for keeping track of working hours spent for individual work packages. However, as an anonymous reviewer of this paper rightly pointed out, giving readers at least an idea of the *dimension* of the time and effort we invested might be helpful when it comes to transferring our experience to other data rescue tasks.

amount of time and effort. The following more specific observations are an attempt to make explicit what we found to be the non-trivial parts of legacy data conversion, and where we think most time and effort have to be invested:

- As a rule, the conversion of legacy data cannot be done fully automatically. It may be necessary in some cases to manually prepare the data using the software with which they were originally created, and there is usually no way around manually checking and editing the result of the conversion. As the example of the PAIDUS data shows, it may sometimes also be useful to insert an intermediate step, in which a custom editor is used to manually correct deficiencies in a first converted version of the data, and only then to convert this corrected version to the target format.
- It is in this manual pre-, intermediate, and post-processing that most time and effort have to be invested. Writing conversion routines or customized editing tools may be labor-intensive, but it is almost negligible in comparison to such manual editing steps. If the legacy corpus exceeds a certain size, so that the manual editing has to be done by a bigger team of (usually) student assistants, we found that the organizational and administrative work necessary to initiate, coordinate, and sustain the work of such a team is also not negligible.
- Another aspect that is not to be underestimated in terms of time and effort is the (re)acquisition of more general information about the corpus. We found it very difficult, in some cases, to get a complete and definite inventory of the data files, tapes, printouts, etc. that make up the legacy corpus. Often, several versions of the same piece of data existed, and it was not always easy to determine the most valid one. It also happened that certain types of information had not been written down at all (i.e., neither digitally nor on paper). Sometimes such information had to be retrieved from a former researcher or student assistant for the project in question, who had in the meantime left the institute.
- Compared to what was necessary to transform (or (re-)create) the transcriptions and their metadata, we found that the digitization of both analogue audio and video data is a routine task. After sufficient storage space had been secured (as much as 150 GB for the video data of the DUFDE corpus) and adequate software had been found (Audacity for audio, Pinnacle Studio and Auto Gordian Knot for video), all that needed to be done was to determine the best conversion and compression parameters for the given data. The digitization itself was then more or less an assembly-line job that required no further supervision or modification.
- Beside such "external" factors, different properties of the legacy data themselves determine the overall time and effort needed to accomplish the conversion. First, undocumented file or character encodings are a serious barrier to getting the conversion process started (syncWriter is the best example of this). Second, data whose digital representation is motivated by the graphic appearance rather than the logical structure seriously complicate the task (as the experience with HIAT-DOS and, again, syncWriter data shows). Third, inconsistencies and irregularities in the legacy data structure are common and need to be taken care of early in the conversion process (cf. the DUFDE,

BUSDE and BIPODE data). Fourth, the legacy data may simply be lacking certain pieces of information that are necessary for a conversion to the target format (e.g., the information about temporal order of the utterances in the PAIDUS corpus).

The most unexpected observation, however, was to find that the closer we got to the target format, the more the extended possibilities of processing the data revealed errors, inconsistencies and gaps in the original data. These were not errors attributable to an outdated data format or insufficient technologies, but errors which resulted from an inaccurate or inconsistent application of transcription guidelines. The fact that these errors had been overlooked before (although they must have been as undesirable in the legacy data as in their converted version) can only be explained by the lack of adequate tools for checking consistency in legacy data. In any case, the correction of these errors and inconsistencies made up a considerable proportion of the overall conversion effort.

4. CONCLUSIONS. In retrospect, our legacy data rescue efforts, though successful in the end, have proven to be much more laborious and time consuming than we had originally expected. We therefore think it is legitimate to ask whether the result is worth all the time and effort that had to be invested. In theory, of course, there are always obvious and incontestable benefits of rescuing legacy data: A twenty-year-old corpus like the DUFDE corpus may today already be seen as a documentation of language use in the past, which would be impossible to recreate anew.[8] Moreover, despite the difficulties encountered, we still had to do considerably less work than would have been necessary if we had wanted to create comparable resources from scratch. In practice, however, resources are limited, and every researcher will have to weigh whether he is going to invest them in the recovery of old or in the creation of new data. In that light, it may be useful to look at the different tasks that make up the conversion effort separately and to try to formulate a recommendation for a course of action when resources are not sufficient to go through all steps of a rescue effort.

Clearly, the most valuable and indispensable part of a spoken language corpus is the primary data, i.e., the audio and video recordings of the original interactions. If these recordings are lost, the options for future researchers to reuse the corpus become severely limited because they will have to rely on the transcriptions alone without a means of judging their quality or supplementing them with their own observational annotations. In our opinion, digitization of audio and video recordings should therefore be considered the first and most important step in legacy data conversion. However, a careful compilation of metadata, i.e., of data about the interactions (e.g., time and date) and the participants involved (e.g., age and language background) is almost equally important, because only this kind of data can ensure that researchers not involved in the original construction of the corpus will be able to understand its design and conduct their own research on the resource.

[8] In a paper delivered at the Colloquium on Convergence and Divergence in Language Contact Situations, Robert E. Vann (2007) pointed out that "we have technology today to change historical linguistics of tomorrow," meaning that today's contemporary resources will become the material for historical linguists of future generations. Rescuing legacy data may be seen as a contribution to this goal.

"Primary and metadata first" is thus the principle we will use in future rescue efforts if we are uncertain whether our resources will suffice for going through all the necessary steps. Luckily, as has been described above, these are *not* the most labor-intensive steps. As a rough estimate, we would claim that they do not make up more than 30% of the overall work, all the rest being dedicated to the pre-processing, conversion, and post-editing of transcription data. If we agree (somewhat arbitrarily) that about 70% of the corpus's value is saved when the primary and metadata have been transformed to a more sustainable form, it seems that rescuing legacy data is a case in which the law of diminishing marginal returns holds.

This is not to say, however, that transcription data[9] should be excluded from the rescue effort. Considering that most of the labor in the construction of the original resource usually goes into the transcription of the recordings (in fields like conversation analysis, it is not uncommon to estimate 100 hours of transcription time for one hour of recording), it is perhaps not surprising that the transcription data also require the most work in the conversion effort. If time and money for this task are available, we will therefore continue to invest them in the future, especially because the availability of transcriptions greatly facilitates an initial access to a corpus for researchers who are not (yet) familiar with it.[10]

Finally, we would like to formulate some recommendations that we think can help to prevent the legacy data problems we are having now from being repeated in the future. A more general formulation of most of these recommendations (as well as others which we consider no less important) can also be found in Bird and Simons 2003:

- Use transcription software that produces XML-encoded files and whose file format captures the logical structure, not the layout, of the data; that is: use tools like ELAN, EXMARaLDA, or Transcriber rather than tools like HIAT-DOS or syncWriter or office software like MS Word, dBase or 4th Dimension (or any of their modern reincarnations). Alternatively, you can also use a tool whose file format is not XML

[9] We understand the term "transcription data" in a broad sense here, following Bird and Liberman's (2001:26) definition of annotation as "the provision of any symbolic description of particular portions of a pre-existing linguistic object." A finer distinction is often made between symbolic descriptions that refer directly to the recording (the "real" transcription) and symbolic descriptions that provide additional analytic information about other symbolic descriptions (called "coding" or "annotation"). The DUFDE data are a good example of a corpus in which the (primary) transcription layer makes up only a small portion of all symbolic descriptions, the remainder being (secondary) coding or annotation layers. When it comes to decisions about what part of the data merits the most attention in a rescue effort, it might be argued that transcription data are more crucial than annotation data. Since, however, the boundary between the two types of data cannot always be clearly drawn, and since some kinds of annotation (e.g., translations or morphological glosses) can be highly relevant for data reuse, we do not feel it appropriate to give a general recommendation about this point.

[10] We should point out here that matters may stand completely differently for corpora of endangered languages. We are basing our reasoning on the assumption that new transcriptions of existing recordings can be done at any time in the future, because we are dealing with widely used languages like French, Swedish, Turkish, etc. In an endangered language corpus, losing transcriptions may be just as fatal as losing the recordings themselves because there may be no one left to (re)transcribe the data.

based but can be reliably transformed to an XML-based format, e.g., Praat or CHAT. If you do so, make this transformation an integral part of your corpus construction workflow (see the 'Format' recommendations in Bird and Simons 2003).

- Digitize audio and video files and make links between transcriptions and recordings. Ideally, use a tool that allows you to make these links immediately in the transcription process (again, tools like ELAN, EXMARaLDA, and Transcriber provide this functionality, MS Word, syncWriter, etc. do not). Use open standards for the digital encoding of audio and video (again, see the "Format" recommendations in Bird and Simons 2003).
- Use a systematic and consistent approach to the encoding of metadata, i.e., data about the general circumstances in which recordings or texts were produced. Create metadata as early as possible, ideally immediately after creating the primary data. Make sure that metadata records are complete and understandable in themselves, i.e., without the help of the project members who created them. Use a standardized file format to represent your metadata (see the 'Discovery' recommendations in Bird and Simons 2003).
- Control the quality and consistency of your data in regular intervals: do not let the corpus grow to an unmanageable size before ensuring that transcribers understand and use the conventions correctly. Large corpora should be divided into smaller packages. In our experience, the risk of confusion increases heavily beyond corpus sizes of 20 hours or 40 transcription documents. Ideally, packages should therefore be smaller than this.
- Document transcription and annotation conventions as well as the design, structure, and technical realization of your corpus as early on and in as much detail as possible. Publish this documentation in a form in which it will still be accessible 50 years from now.

If these simple rules are observed, there is a good chance that the term "legacy data" will lose its negative connotations in the near future.

References

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

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3):557–582.

Bloom, Lois, and Margaret Lahey. 1978. *Language development and language disorders*. New York: Wiley.

Boersma, Paul, and David Weenik. 1996. PRAAT, a system for doing phonetics by computer, version 3.4. *Institute of Phonetic Sciences of the University of Amsterdam, Report 132*. www.praat.org

Gerhard, Ludwig, Conxita Lleó, Jochen Rehbein, and Wolff Ekkehard. 1997. Das Graduiertenkolleg Mehrsprachigkeit und Sprachkontakte (GKMS). Abschlußbericht an die DFG über die Arbeit vom 30.09.1990 bis zum 30.09.1993. Arbeiten zur Mehrsprachigkeit, Serie A (60).

Ehlich, Konrad, and Jochen Rehbein. 1976. Halbinterpretative Arbeitstranskriptionen (HIAT). *Linguistische Berichte* 45, 21–41.

Golinski, Bernadette. 2007. Kommunikationsstrategien in interskandinavischen Diskursen. *Philologia: Sprachwissenschaftliche Forschungsergebnisse 95*. Hamburg: Kovač.

Herkenrath, Annette. 2007. Discourse coordination in Turkish monolingual and Turkish-German bilingual children's talk: işte. In *Connectivity in grammar and discourse: Hamburg studies in multilingualism 5*, ed. by Jochen Rehbein, Christiane Hohenstein, and Lukas Pietsch, 291–325. Amsterdam: John Benjamins.

Karakoç, Birsel. 2007. Connectivity by means of finite elements in monolingual and bilingual Turkish discourse. In *Connectivity in grammar and discourse: Hamburg studies in multilingualism 5*, ed. by Jochen Rehbein, Christiane Hohenstein, and Lukas Pietsch, 119–227. Amsterdam: John Benjamins.

Köppe, Regina. 1994. The DUFDE project. In *Bilingual first language acquisition*, ed. by Jürgen Meisel, 15–28. Amsterdam: John Benjamins.

Lleó, Conxita, Michael Prinz, Christliebe El Mogharbel, and Pilar Larranaga. 1995. PAIDUS- Parameterfixierung im Deutschen und Spanischen. Final project report. Hamburg.

Meisel, Jürgen M., Harald Clahsen, and Manfred Pienemann. 1981. On determining developmental stages in natural second language acquisition, *SSLA* 3(2):109–135.

Pienemann, Manfred. 1981. *Der Zweitsprachenerwerb ausländischer Arbeiterkinder*. Bonn: Bouvier.

Rehbein, Jochen, Wilhelm Griesshaber, Petra Löning, M. Hartung, and Kristin Bührig. 1993. Manual für das computergestützte Transkribieren mit dem Programm syncWRITER nach dem Verfahren der Halbinterpretativen Arbeitstranskriptionen (HIAT). Unpublished ms.

Rehbein, Jochen. 2007. Matrix constructions. In *Connectivity in grammar and discourse: Hamburg studies in multilingualism 5*, ed. by Jochen Rehbein, Christiane Hohenstein, and Lukas Pietsch, 419–447. Amsterdam: John Benjamins.

SCHMIDT, THOMAS. 2005. *Computergestützte Transkription — Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln*. Frankfurt a. M.: Peter Lang.

SCHMIDT, THOMAS, and KAI WÖRNER. In press. EXMARaLDA — Creating, analysing and sharing spoken language corpora for pragmatic research. In *Corpus-based pragmatics*, ed. by Jens Allwood, (Hrsg.).

SCHNEIDER, WOLFGANG. 2001. Der Transkriptionseditor HIAT-DOS. *Gesprächsforschung* 2:29–33.

VANN, ROBERT E. 2007. Language contact, language change, spontaneous speech innovation, and why we need more digital archives of spoken language corpora from contact dialects. Paper presented at the International Colloquium on Convergence and Divergence in Language Contact Situations, October 18–20, Hamburg, Germany.

ZEEVAERT, LUDGER. 2004. Interskandinavische Kommunikation: Strategien zur Etablierung von Verständigung zwischen Skandinaviern im Diskurs. *Philologia. Sprachwissenschaftliche Forschungsergebnisse 64*. Hamburg: Kovač.

Thomas Schmidt
thomas.schmidt@uni-hamburg.de

Jasmine Bennöhr
jasminebennoehr@gmx.de