The transcription system EXMARaLDA: an application of the annotation graph formalism as the basis of a database of multilingual spoken discourse

Thomas Schmidt
Project Z,
SFB "Mehrsprachigkeit",
University of Hamburg
Max Brauer-Allee 60
D-22765 Hamburg, Germany
Thomas.Schmidt@uni-hamburg.de

Abstract

This paper describes EXMARaLDA, a system for computer transcription of spoken discourse developed and used by the SFB "Mehrsprachigkeit" at the university of Hamburg. EXMARaLDA consists of several DTDs for XML coding of transcription data and some input and output tools for these formats. Apart from being a transcription system in its own right, EXMARaLDA also plays the role of a mediator between older existing data formats at the SFB and between these formats and a planned database of multilingual spoken discourse.

1 Introduction

The SFB "Mehrsprachigkeit" (Research Center on Multilingualism) at the University of Hamburg brings together linguists doing research on multilingualism from a variety of theoretical perspectives. The majority of the projects work with spoken language data, and the formats, tools and platforms these data come in are as diverse as the backgrounds of the research groups. There are data in many different languages, created and processed on different platforms and with different transcription conventions and transcription tools. The systems in use are for the most part outdated and technically as well as conceptionally incompatible with one another. The urgent need for an exchange format between them is therefore obvious.

The system EXMARaLDA has been developed for this purpose. As it makes use of the recently established standards XML and UNICODE and works with the concept of annotation graphs (Bird and Liberman 2001), it also places itself in the larger context of current efforts for standardizing and making exchangeable language data, such as the TALKBANK project (Bird and MacWhinney 2000), the EUDICO project (Brugman et al. 2000) or the MATE initiative (Dybkjær 2000).

This paper gives an overview over the development of EXMARaLDA. Section 2 describes the current practice at the SFB for working with spoken language data, i.e. the data formats and tools the projects are currently using. Section 3 then describes the design of EXMARaLDA and the role it is intended to play for data exchange within the SFB. Section 4 finally hints at actual and planned cooperation between EXMARaLDA and other projects.

2 Existing Data and Tools at the SFB

2.1 Data

2.1.1. Language Acquisition Data

Four projects at the SFB are concerned with different aspects of bilingual language acquisition. The data they work on partly stems from older projects and is partly being recorded and transcribed at the moment. More specifically there are:

- Data from French/German, Portuguese/German, Basque/Spanish and Italian/German bilingual children. For these data, the recordings were first transcribed on paper and then partly entered into different kinds of dBase databases (some via

LAPSUS) where some of them were also annotated syntactically.
- Data from Spanish/German bilingual children. These recordings were also first transcribed on paper and then entered into a database, 4th Dimension (a Macintosh application) in this case.
- Data from Turkish/German bilingual children. These were transcribed directly on the computer using syncWriter.

2.1.2. Other Discourse Data

Apart from the four projects working with language acquisition data, there are six other projects who use transcription of spoken discourse in their work, namely:
- Japanese and German expert discourses. These are transcribed with syncWriter.
- Doctor-patient communication mediated by non-trained interpreters in Portuguese/German and Turkish/German, also transcribed with syncWriter
- Interpreted dialogues in English/German, French/German, Japanese/German, Chinese/German and Russian/German, transcribed according to the Verbmobil conventions (Burger, 1997)
- Conversations between English native speaker and/or English L2-learners. These are transcribed with 'Simple Exmaralda', a component of the EXMARaLDA system.
- Interpreted interviews in English/Luganda, transcribed in MS WORD.
- Classroom discourse, radio broadcasts and presentation discourse, each involving two or more Scandinavian languages. These are transcribed with HIAT-DOS.

2.1.3. Written Data

The remaining three SFB projects only work with written data (historical texts in Middle High German, Latin, Greek and Old French). For reasons of time, these are currently not taken into account for the development of the database, but will be at a later point in time.

2.2 Tools

2.2.1 syncWriter

SyncWriter (Dybkjaer et al. 2001; Meyer, 2000; Rehbein et al., 1993) is a commercially distributed software for the Macintosh. It is basically an editor for interlinear text with some facilities to integrate video, audio or image data and, as such, is used in three SFB-projects for creating, editing and printing transcripts in musical score notation according to the HIAT conventions (Ehlich and Rehbein, 1976).[1] SyncWriter stores its data in a non-disclosed binary format. The program has an export facility for so called 'segment lists', i.e. text files containing a list of utterances or other segments, but these exported files do not contain all the information present in the original transcription. The possibilities to reuse syncWriter data with other applications are therefore severely limited. A DOS- or Windows-Version of the software does not exist.

2.2.2 HIAT-DOS

HIAT-DOS (Ehlich, 1992) is a software for DOS and Windows systems. It is similar to syncWriter in so far as it facilitates the creation, editing and printing of interlinear text. Unlike syncWriter, however, it directly implements some of the features of the HIAT conventions (e.g. verbal, non-verbal and intonational tiers), and is therefore more closely tied to this transcription system.

The program stores its data in text files that reflect the graphical structure of a score rather than the logical structure of the transcription that this score represents. Data reuse is therefore not excluded, but relatively difficult, because a logical structure has to be derived from its optical representation.

2.2.3 LAPSUS

LAPSUS (Crysman, 1995) is a dBase IV application designed for input and retrieval of

[1] It is, however, not restricted to this transcription system. Mainly owing to its multi media facilities, syncWriter is also extensively used for transcriptions in sign language research at the University of Hamburg and elsewhere (see also Dybkjaer et al. 2001)

child language acquisition data. It consists of an input mask for the primary data and several coding modules for annotation of the primary data. The data are stored in dBase IV database tables making a reuse relatively easy.

3 EXMARaLDA

3.1 Statement of the problem

The ultimate goal of the project in which EXMARaLDA is being developed is the construction of a database comprising all the spoken and written language data in use at the SFB and making them available for elaborate queries. As can be seen from the previous section, the formats of these data and the tools used to work with them are as diverse as the languages involved and the research interests of the projects. Altogether, there are currently about 2200 transcriptions of recordings with an average duration of 30 minutes each (many more will follow as the SFB continues its work). Fourteen different languages and nine different data formats are involved; the research interests range from phonetic over syntactic to discourse analyses. Last but not least, three different computer operating systems (Windows, MAC OS 9.x and LINUX) are in use, with a fourth one (MAC OS X) just being released. Hence, the following quote

„Particular bodies of data are created with particular needs in mind, using formats and tools tailored to those needs, based on the resources and practices of the community involved. Once created, a linguistic database may subsequently be used for a variety of unforeseen purposes, both inside and outside the community that created it."

(Bird and Liberman 2001:2)

is a nice and short description of the problems we encounter in everyday research when working with computerized language data. It is in many cases close to impossible for different projects to share or exchange their data via the computer, let alone the tools used to create, annotate or analyze them.

It is therefore not only the ultimate goal of a multilingual database that creates an urgent need for a common data format, but – at a much more basic level – the lack of possibilities to interchange language data or share tools between different projects and to adapt "foreign" data to the theoretical needs and goals of a specific project. Moreover, because tools and formats are by now several years old and have never been updated or adopted to any standard that has emerged in the meantime, a great part of the data is threatened by an eventual "data death", i.e. the data may become unusable on future operating systems.[2]

Given the diversity of the existing tools and formats and the multitude of present and potential research interests, there was hardly a possibility to place any simplifying restrictions on a candidate for a common data format. After a close look at most of the state-of-the-art proposals for transcription standards, we came to the conclusion that the concept behind the annotation graph formalism best suited our needs. The next two sections describe how we intend to make use of this concept for the construction of the multilingual database.

3.2 System architecture

Figure 1 gives an overview of the overall system architecture. The core component of the system is EXMARaLDA (EXtensible MARkup Language for Discourse Annotation). It consists of a number of XML Document Type Definitions that specify the syntax for three types of discourse transcriptions with different levels of complexity. In the first place, EXMARaLDA's role is that of an interlingua between the different "homegrown" formats, i.e. their greatest common denominator (or rather their least common multiple). However, conversion between the existing formats and EXMARaLDA is a one way street – it is not intended (and in most cases not possible) to provide conversion tools from EXMARaLDA to the existing formats. Moreover, conversion from existing formats to EXMARaLDA is in many cases not a trivial task. Especially for the interlinear syncWriter and HIAT-DOS data, only part of the conversion work can be done automatically, and a considerable amount of manual post-editing has to

[2] The best example for this are the syncWriter data that can only be read with the syncWriter software designed for MAC OS 9 or earlier. With the release of the MAC OS X operating system and because the software is no longer updated, there is reason to fear that at some point in the not too distant future, Macintosh users will be confronted with the decision to either continue using an outdated OS or "lose" their syncWriter data.

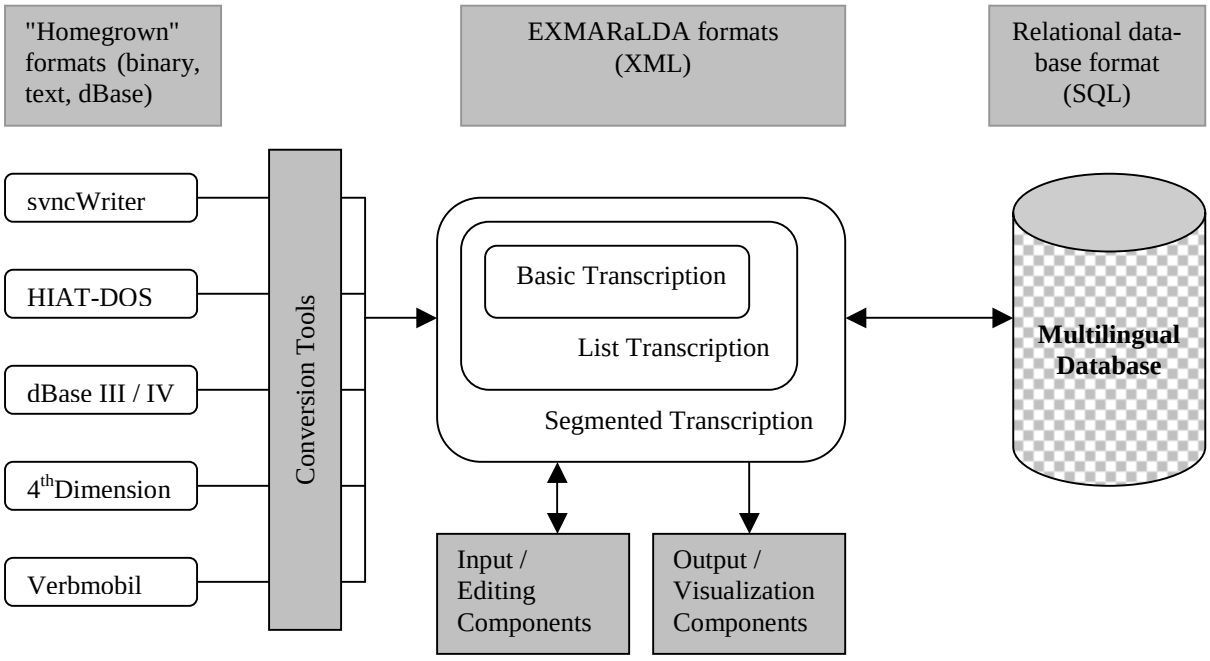

Figure 1: *System architecture*

be added. Because of this, it is in the long run necessary to provide new input/editing and ouput/visualization tools that operate directly on the EXMARaLDA formats. Some of these are currently under construction (see below). It is intended to optimize these tools to the needs of the individual projects, such that each project can continue to work with language data in its habitual way while at the same time producing data that are accessible to all other projects. Together with these input and output tools, EXMARaLDA can thus be seen as a transcription system in its own right. Finally, EXMARaLDA will also play the role of an interface between the existing data formats and the multilingual database. The latter will be an SQL-database bundling all transcriptions and making them available for elaborate and efficient querying.

3.3 Data Formats

The EXMARaLDA system provides three different DTDs for encoding discourse transcriptions as XML files. The first of these two – the basic and the list transcription – reflect the "traditional" ways of graphical organization of transcription data on a printed page, as presented, for instance, in Edwards (1992). They are the natural target formats for the conversion of the existing data, as these all follow one of the traditional layout principles of score, column or vertical notation. However, as Knowles (1995) puts it,

> "The new opportunities are not yet being fully recognized and exploited by linguists [...] Texts are still seen as objects in book format, with words running in horizontal lines from left to right. Annotations are added to these horizontal lines. But book format is an attribute not of speech, but of Western writing systems. There is no reason beyond established custom and practice to present speech in this way. On the contrary, since there are often several annotations relating to the same piece of data, book format is in many cases inappropriate. The use of book format without consideration of other possibilities is based on a confusion between the organization of the data itself, and the presentation of the data on the printed page."

transcriptions of discourse may well have a more complex structure than what is presentable on a printed page. In fact, we found that already a simple word level annotation of multi-party discourse usually requires a more complex structure that that of a basic or a list transcription. EXMARaLDA therefore defines a third format – the segmented transcription – in order to be able to describe such structures.

MAX [v]: You keep interrupting me, Tom.
MAX [nv]: --- *pointing at Tom* ----------
TOM [v]: Oh, I'm sorry for that.
TOM [nv]: --- *smiling* ----------

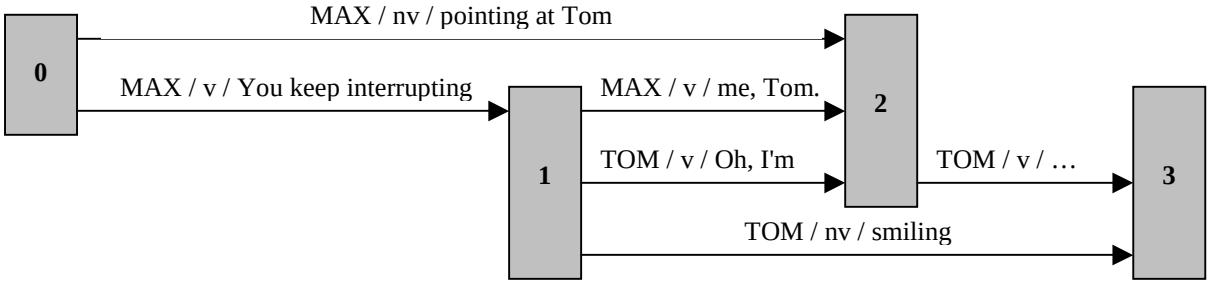

Figure 2: Score notation (*Partitur*) and underlying data structure

As the three EXMARaLDA formats constitute a system of subsets –

basic ⊂ list ⊂ segmented

– the problem of a new multitude of formats does not arise. Instead, as the system provides tools for moving between the different formats, different tools can limit themselves to one of them without having to bother with possibly unnecessary complex structures of the other.

3.3.1 Basic Transcription

EXMARaLDA's basic transcription format is the target format for most of the existing data at the SFB, especially for the interlinear syncWriter and HIAT-DOS data. As figure 2 shows, the structure underlying transcriptions in score notation is a very simple one: the transcription is organized in several tiers, and each tier contains a number of event descriptions that are all anchored to the same timeline and that do not overlap one another within a tier. A basic transcription can be visualized either in score notation or in column notation (using the terminology of Edwards 1992).

3.3.2 List Transcription

The list transcription format is the target format for transcriptions in vertical notation (Edwards 1992). In addition to the information contained in a basic transcription, i.e. events anchored to a common timeline and to a tier, it summarizes events from different tiers in speaker turns. Figure 3 illustrates this. With this additional information, a visualization in vertical notation becomes possible. Of course, any list transcription can be transformed into a basic transcription, and hence also be visualized in score or column notation.

MAX: (pointing at Tom) You keep interrupting [me, Tom.]
TOM: (smiling) [Oh, I'm] sorry for that

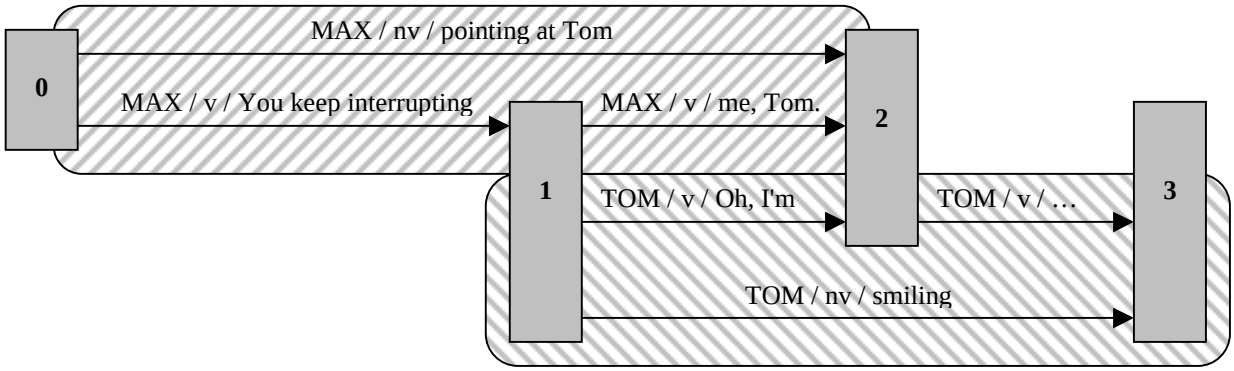

Figure 3: Vertical notation and underlying data structure

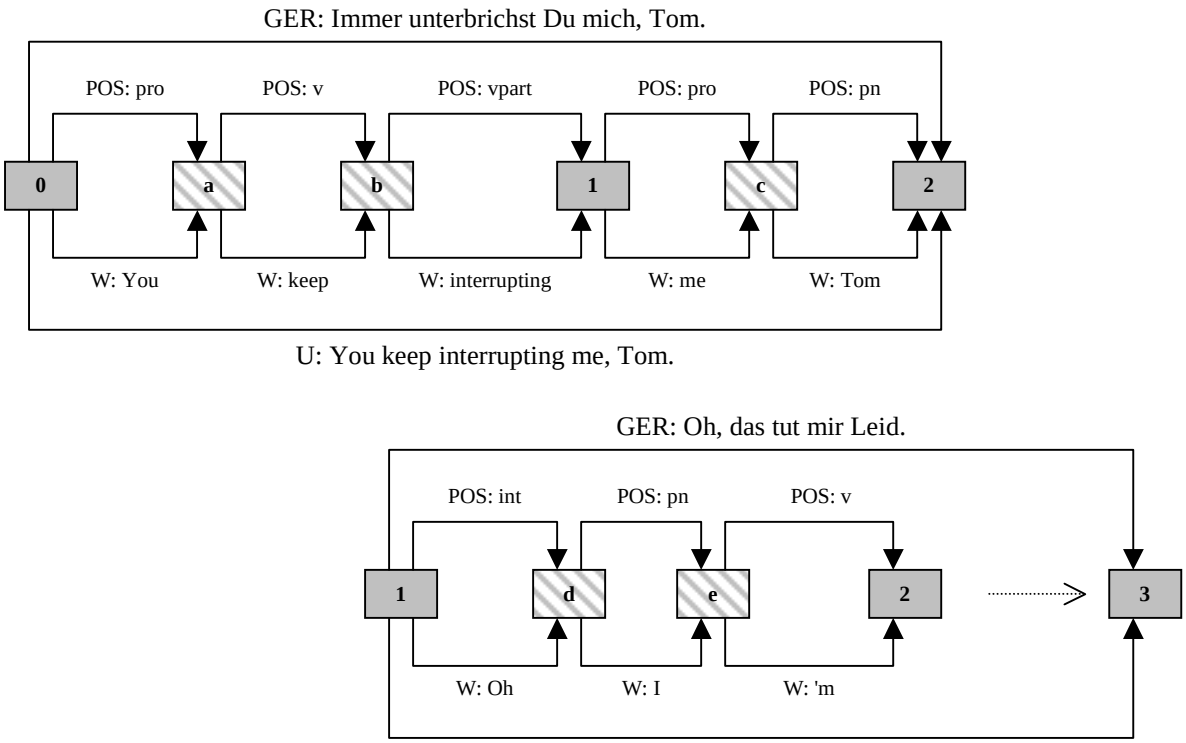

Figure 4: Structure of annotated data

3.3.3 Segmented Transcription

The primary transcription, i.e. the description of the temporal structure of actual discourse events is usually done following one of the traditional design principles. All of these require one single timeline common to all events. However, when it comes to annotating the primary data (e.g. providing translation of utterances or POS-tagging words), it is usually not the temporal, but the linguistic structure of events that is important. As figure 4 shows, this may result in new points in the timeline that cannot always be brought into a unequivocal order. The typical case arises when overlapping stretches of speech of two or more speakers contain more than one linguistic unit (as is the case with the unit "(w)ord" in figure 4). The EXMARaLDA segmented transcription format therefore allows multiple, partially intersecting timelines in order to make annotation of arbitrary units possible without having to bring these all into a strict order (which would in many cases be impossible for the transcriber, anyway).

3.4. Tools

As, for reasons explained above, platform independence of the software tools is an important criterion, it was decided to implement all tools in JAVA (Version 1.3.1.). It has turned out in the meantime that this in not a perfect solution either, because Apple will not upgrade the JVM for the "old" operating systems (i.e. OS 9.X and earlier), but no better solution seems to be available, and we count on Macintosh users upgrading their OS in the near future. All software will thus work (and has been tested) on Windows 98, Windows NT, Windows ME, Windows 2000, Windows XP, MAC OS 10.1. and some or most LINUX and UNIX versions. Once the basic tests have been completed the software will be made available via the internet. Some preliminary command line tools can already be downloaded from

http://www.rrz.uni-hamburg.de/exmaralda

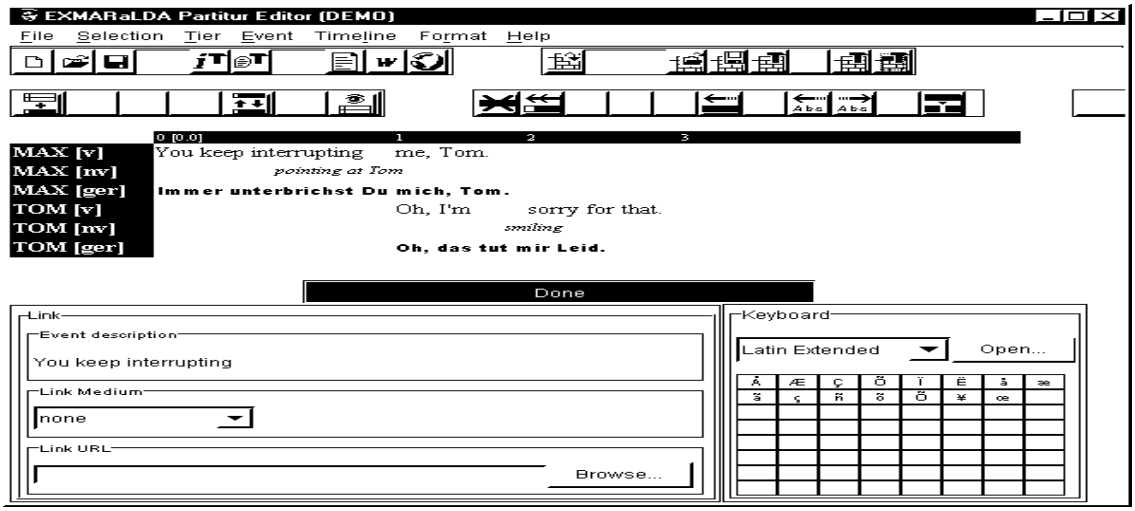

Figure 5: Screenshot of Score editor

3.4.1. Ouput/Visualization Tools

In our experience, the crucial element that decides on the user acceptance of a computer transcription format is not so much its theoretical or computational power, but rather the way in which its data is visualized. Many users are more interested in having a comfortable method of viewing and printing transcripts than in issues such as flexibility or platform independence of the data format. The (technically rather complex) method of visualizing transcriptions as interlinear text (i.e. in score notation), which many of the SFB projects are used to, is therefore an important component of the system. At present, JAVA-methods for reading in a basic transcription and putting out a visualization in the form of a score, either as a RTF- or a HTML-file, have been implemented. These can be parameterized so that different fonts, paper sizes etc. can be used. Other visualization methods (e.g. for column notation) will follow.

3.4.2. Input/Editing tools

As a first input method, an import filter for transcriptions stored in text files was implemented. These transcriptions have to follow some very simple syntactic rules specified in the "Simple EXMARaLDA" conventions. For instance, the above example from figures 2 to 4 could look as follows in a Simple EXMARaLDA file:

MAX: [pointing at Tom]
You keep interrupting <me, Tom.>1>
{Immer unterbrichst Du mich, Tom.}
TOM: [smiling]
<Oh, I'm >1> sorry for that.
{Oh, das tut mir Leid.}

With the help of the import filter, such text files can be read into an EXMARaLDA (list) transcription and then be further edited with other system components.

Whereas this is a simple and convenient method for creating a first raw version of a transcription, it is not sufficient for more complex transcriptions with many speakers and several tiers for each speaker. As a second input and editing tool, an editor has therefore been implemented that presents a basic transcription as a score and allows interactive editing of the transcription. Figure 5 shows a screenshot. The appropriate output methods have also been integrated into this tool along with a facility for linking the transcript to external media files and a virtual keyboard for entering non-standard Unicode characters, as, for instance, IPA symbols.

It is intended to construct further input and editing tools (e.g. an input tool providing a column notation view of the data, and a tool for annotating segmented transcriptions) as the project continues.

4 Cooperation

4.1 EXMARaLDA and TASX

The TASX project (see Milde, this volume) has similar goals to our project, i.e. providing a flexible format for time-aligned language data and tools for input/editing and analysis of such data. The TASX file format of level 1 is very similar to EXMARaLDA's basic transcription format. We have therefore implemented tools for conversion between these two formats, so that EXMARaLDA data can be processed with TASX tools and vice versa. We intend to further pursue this cooperation in the future and possibly integrate our tools into one common environment.

4.2 EXMARaLDA and AG

As can be seen from the short description in section 3.3 the EXMARaLDA file formats are based on the concept of annotation graphs. Constructing an export filter to the ATLAS interchange format should therefore be an easy task (it has, however, not yet been tackled). We are attentively watching the development of TALKBANK and related projects and hope to profit from their findings and solutions for the development of our multilingual database.

4.3 EXMARaLDA and Standoff Annotation

The concept of standoff annotation (Dybkjær 2000) is currently not applied to the data at the SFB, simply because the "home-grown" formats provide no appropriate linking points for such a technology. However, we believe the possibility to perform standoff annotations on the data will be crucial in future everyday work at the SFB as well as in development of the database - There must be an efficient way to separate commonly reusable annotation from user-specific annotation that is largely irrelevant to others. EXMARaLDA's segmented transcription format therefore provides ID-attributes for each specified segment, and these IDs may be used as a reference for standoff annotation.

5 Conclusion

This paper wants to give a rough overview over the EXMARaLDA system. Technical details cannot be discussed here, so the reader interested in technical issues is instead referred to the project homepage (as yet only in German, but an English version will follow soon)

http://www.rrz.uni-hamburg.de/exmaralda

or invited to contact the author.

References

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

Steven Bird and Brian MacWhinney 2000. *Talk Bank: A Multimodal Database of Communicative Interaction.* [http://www.talkbank.org/resources/talkbank.pdf]

Henning Brugman et al. 2000. *EUDICO – Annotation of MultiMedia Corpora.* LREC 2000 Workshop, Athens.

Susanne Burger 1997. *Lexicon of the Conventions for Transliteration of Spontaneous Speech – Verbmobil II.* VM-Techdoc 56. München.

Berthold Crysman 1995. *LAPSUS: A utility for the transcription of empirical data in language acquisition research.* Unpublished manuscript. Hamburg.

Laila Dybkjær 2000: *MATE Deliverable D6.2: Final Report.* [http://mate.nis.sdu.dk]

Laila Dybkjær et al. 2001. *Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data.* ISLE Natural Interactivity and Multimodality Working Group Deliverable D11.1.

Jane Edwards and Martin Lampert (ed.) 1992. *Talking Data – Transcription and Coding in Discourse Research*. Hillsdale.

Jane Edwards 1992. *Principles and Contrasting Systems of Discourse Transcription*. In: Edwards and Lampert 1992: 3 — 31.

Konrad Ehlich 1992. *HIAT - a Transcription System for Discourse Data.* In: Edwards and Lampert 1992: 123—148.

Konrad Ehlich and Jochen Rehbein 1976. *Halbinterpretative Arbeitstranskriptionen (HIAT).* In: Linguistische Berichte 45: 21— 41.

Gerry Knowles 1995. *Converting a corpus into a relational database: SEC becomes MARSEC.* In: Leech et al. Geoffrey Leech et al. (ed.). *Spoken English on Computer: Transcription, Markup and Application*: 208—219. Harlow: Longman

Bernd Meyer 2000. *The Computer-Based Transcription of Simultaneous Interpreting.* In: Birgitta Dimitrova and Kenneth Hyltenstam (ed.). *Language Processing and Simultaneous Interpreting - Interdisciplinary perspectives*: 151—158. Amsterdam.

Jochen Rehbein et al. 1993. *Manual für das computergestützte Transkribieren mit dem Programm syncWriter nach dem Verfahren der Halbinterpretativen Arbeitstranskriptionen (HIAT).* Hamburg.

Thomas Schmidt 2001. *EXMARaLDA - ein System zur Diskurstranskription auf dem Computer.* To appear in: Arbeiten zur Mehrsprachigkeit (AZM). Hamburg.

Thomas Schmidt 2001. *Gesprächstranskription auf dem Computer - das System EXMARaLDA.* To appear in: Gesprächsforschung (2). [http://www.gespraechsforschung-ozs.de]