Creating and Working with Spoken Language Corpora in EXMARaLDA

Thomas Schmidt

Spoken language corpora—as used in conversation analytic research, language acquisition studies and dialectology—pose a number of challenges that are rarely addressed by corpus linguistic methodology and technology. This paper starts by giving an overview of the most important methodological issues distinguishing spoken language corpus work from the work with written data. It then shows what technological challenges these methodological issues entail and demonstrates how they are dealt with in the architecture and tools of the EXMARaLDA system.

1. Introduction

At first glance, it may seem inappropriate to include a contribution on spoken language corpora in a colloquium on "Lesser Used Languages and Computer Linguistics" —after all, spoken language as such is certainly not a 'little used' kind of language. However, spoken language shares a fate with 'smaller' languages insofar as corpus and computational linguistics as fields of study have a definite bias, not only towards major, standardised languages, but also towards written language in general. Thus, large reference corpora usually consist entirely or to a great part of written texts, the greater part of corpus linguistic literature deals with phenomena of written language, and the technology for constructing and analysing language corpora is also much further developed and established for the written medium. The reasons for this may be of a theoretical nature—certainly, some research questions are best approached by looking at written data. Yet the prevailing cause for the dominance of written language in corpus linguistic seems to me to be a pragmatic one: whereas large amount of written text are easily available for integration into a corpus, spoken data has to be tediously recorded and transcribed; both processes involve difficult methodological challenges, and it is

thus much harder to arrive at a reasonably large amount of spoken language material to address a certain research question than it is for written data.

However, it is also unquestionable that certain linguistic phenomena cannot be studied by looking at written language data alone. Child language acquisition, dialectal variation and, of course, the structure of face-to-face interaction (as studied for example by conversation analysis) are all cases in point.

This paper starts by discussing some of the methodological challenges involved in constructing spoken language corpora, comparing them to the corpus construction workflow for written data. It then proceeds to demonstrate how these challenges are addressed in the EXMARaLDA system.

2. Some Methodological Challenges for Spoken Language Corpora

2. 1 Primary and secondary data, modelling

It is common in corpus linguistics to draw a distinction between primary and secondary data. When we deal with written language, 'primary data' usually denotes the original text, as published and intended to be read by its audience—for example, a printed book or an electronic (e.g. PDF) document—while 'secondary data' means a derived representation of this text as included in the corpus (e.g. a simple text file or a document with TEI markup).[1] Getting from the primary to the secondary data almost always involves some kind of simplification, abstraction, interpretation and, possibly, purposeful modification. Thus, for most written language corpora, text layout and formatting of the original document are not represented (i.e. abstracted over) in its derived version, non-textual elements (pictures, diagrams, etc.) are also left out, and some normalisation (e.g. undoing hyphenation at the end of a line) is carried out. All these modifications are (or at least should be) justified with respect to the research question the corpus is meant to address. We can therefore regard this step from primary to secondary data as a kind of scientific modelling, because it has the three general properties that, for instance, (Stachowiak 1973) uses to characterise a scientific model:[2]

1 In other contexts, though, 'primary data' means what is called 'secondary data' here, and 'secondary data' means analytic information ('annotation') added to this data.

2 See Schmidt (2005a, b) for a comprehensive discussion of the modelling aspect in the work with linguistic data.

- the representation property (*Stellvertretermerkmal*) states that, in the process of scientific study, the model takes the place of the actual thing to be studied;
- the abstraction property (*Verkürzungsmerkmal*) states that a model is always a simplified representation of the thing modelled; and,
- the pragmatic property (*Pragmatisches Merkmal*) states that this simplification is motivated by a certain purpose.

When it comes to spoken language data, the first thing to notice is that we are dealing here not with one but two distinct steps from the original linguistic fact to its representation in a corpus. In analogy to the written language case, we should call the original interaction the primary data. However, this data is ephemeral—it is not available for systematic study unless it is made more permanent through the process of recording. One or more audio or video recordings of the interaction would thus have to be called the secondary data. In transcription, this secondary data is then transferred into a written representation, which, consistently, should then be denoted 'tertiary data'.

Without a doubt, both these steps—from interaction to recording and from recording to transcription—involve a fair amount of modelling. For instance:

- The choice of recording type (audio or video) has to be motivated by the intended research to be carried out. Since audio recording blinds out all visual aspects of interaction, the resulting corpus can only be used to study the audible aspects. The same can be said for certain parameters of the recording (e.g. how many cameras or microphones to use, where to put them).
- Transcription itself has always been characterised as a "selective process reflecting theoretical goals and definitions" (Ochs 1979: 44). It is hardly controversial that the process of transferring spoken language to the written medium can only be done on the basis of a theoretically-motivated decision about which aspects of the recording to include and which to leave out. The great number of existing transcription systems (e.g. HIAT, Rehbein et al. 2004; GAT, Selting et al. 1998; CHAT, MacWhinney 2001) and the different principles for transcript layout (e.g. musical score vs. line notation) testify to this.
- In contrast to written language as "the language of distance" (Koch & Oesterreicher 1995), spoken language is embedded in a specific situational context. Understanding, interpreting and analysing spoken language therefore also depends on the availability of information about the speech situation, such as time and place of

the interaction, things that happened before the interaction started, the spatial constellation of speakers, and so forth. As with transcription itself, there is no independent external criterion for deciding which of this information to include and which to leave out. Again, the modelling of such metadata thus relies on theoretical considerations. The same holds for sociographic metadata about speakers, such as age, social status, language competence and so forth.

While the construction of both written and spoken language corpora thus involves a modelling step, this step can be said to be much more pronounced, that is, requiring more abstraction and theory-guided interpretation, for the latter type. For written language, at least the character table for the alphabet of a given language as well as the list of words defined by an established dictionary can be regarded as a common ground for all corpus modelling—meaning that, for a modern and standardised language, the mapping of these entities from original to corpus representation is usually *not* interpretative. For spoken language, there is no such common ground. A transcription of a spoken language recording is therefore a much less stable basis of analysis than, say, an ASCII text representation of a newspaper article. Hence, when working with a spoken language corpus, researchers will value the possibility to verify, and possibly revise, the transcriber's decisions by listening to the original recording.

2. 2 Data structures

When representing language data in the digital medium, a choice has to be made about general properties of data structures, that is, salient structural relations between entities that must be encoded in a file or database. Certainly, hierarchic inclusion (e.g. a paragraph being made up of sentences, which, in turn, are made up of words) and sequential ordering (e.g. the words in a sentence following one another), are two of the most important such relations in linguistic description. In fact, it has been argued in the famous OHCO thesis (De Rose et al. 1990: 6) that these two relations are sufficient to characterise the structure of a written text, or, in the author's words, that "text is an ordered hierarchy of content objects". Although this thesis has variously been refuted as being too strong (also by the authors themselves), OHCO remains the dominant modelling paradigm of many approaches to encoding corpus data, most notably the TEI guidelines. In these approaches, then, the primary data structure is a tree grouping smaller linguistic entities into larger ones, and all non-tree-like structures or overlapping hierarchies (e.g. the paragraph vs. page division of a text) are treated,

if at all, as exceptions to the rule. As the success of TEI encoded corpora shows, this has proven a practicable way of handling written language data.

Spoken language, however, as it unfolds over time, exhibits many non-sequential, non-tree-like relations on the lowest structural level: speakers' utterances may overlap, verbal behaviour is accompanied by simultaneous gestures or facial expressions, and the verbalisations of one speaker are themselves made up of different aspects (e.g. lexical words and suprasegmental characteristics like modulation, voice quality), which may need to be described in 'parallel' structures. In spoken language, the exceptions to the OHCO assumption thus become the rule. While it is still possible to use OHCO-based paradigms to encode such data (see, for instance, Schmidt 2005c), any system claiming to be adequate for spoken language representation has to pay due attention to a consistent and practicable method for also encoding non-hierarchic structures. Bird/Liberman's (2001) annotation graph formalism—arguably one of the most influential proposals in the field in the last ten years—therefore radically emphasizes the temporal aspect of spoken language, suggesting using as the primary data structure an acyclic graph whose nodes can be anchored to a timeline.

2. 3 Size and balance, speed and efficiency

The size of a corpus determines to a great deal the empirical findings that may be derived from it. Any quantitative or statistical analysis of empirical data requires a critical mass so that regularities in the sample can be generalised to the population the sample represents. And even for purely qualitative analyses, only a sufficiently large corpus allows the researcher to judge the value of an individual example, because it is only in comparison to a reasonably big number of other examples that its uniqueness or 'prototypicalness' can be plausibly evaluated. Likewise, the concept of balance, that is, the property of a corpus to represent certain parameters (like genre, geographic origin etc.) in adequate, non-skewed proportions when compared to the entirety of linguistic facts, is a very important criterion for empirical investigations.

For written language corpora, both the problems of size and balance have been addressed in a satisfactory fashion. For example, the German reference corpus of the IDS[3] consists of no less than 3.6 billion words, and the fact that the WWW makes enormous amounts of text readily available for electronic search shows that there is, in principle, no upper limit to the size of a written language corpus. Convincing

3 http://www.ids-mannheim.de/kl/projekte/korpora/

concepts for balanced corpus stratification have been applied, for instance, in the English BNC corpus[4] or the German DWDS corpus (Geyken 2009), both resources counting over 100 million words.

The situation is much different for spoken language corpora. One of the largest such resources, the Spoken Dutch Corpus (CGN, Oostdijk & Broeder 2003), although it counts an impressive 8 million transcribed words, is still one or two orders of magnitude smaller than the above-mentioned resources. Most other published spoken language corpora do not exceed the million-word boundary.

Yet, from a theoretical perspective it might be argued that spoken language corpora, in order to enable the same kind of generalisations, should actually be *larger* than their written counterparts. Since spoken language is less standardised and occurs in a much wider variety of circumstances, a really 'balanced' corpus would have to take into account a very large number of speaker and interaction types. For example, a reference corpus of spoken German would have to cover in comparable proportions dialectal, register and topic variation across such diverse interactions types as telephone calls, TV debates, service encounters, informal talk, classroom discourse and political speeches. Aside from the fact that no widely accepted model for such stratification exists (whereas, for example, Biber's [1993] ideas can be considered a kind of standard approach to written language stratification), practical reasons make it virtually impossible to construct spoken language corpora in the 100-million-word dimension.

This is so because spoken language corpus construction requires manual work for many steps that can be automated (or at least semi-automated) for written language data. In (primary) data acquisition, written texts can be harvested from the Web or otherwise be provided in electronic format, whereas spoken interaction has to be recorded in the field. In secondary (or tertiary) data creation, semi-automatic methods like HTML cleanup or Optical Character Recognition are only applicable to written language documents, whereas spoken language transcriptions have to undergo the extremely time-consuming process of manual transcription. Thus, in fields like conversation analysis with its fine-grained and detailed transcription procedure, it is not uncommon to estimate 100 hours of transcriber's time for one hour of recorded interaction. Given these drastic differences in the time and effort required to construct corpora, the speed and efficiency of corpus tools becomes a paramount concern when the size of a spoken language corpus is considered relevant in any way.

4 http://www.natcorp.ox.ac.uk/

2.4 Summary

Summarising the preceding sections, a spoken language corpus, when compared to a written language corpus, poses three methodological challenges:

- Its base is more *instable* insofar as the modelling step between original data and corpus data is much more pronounced—far-reaching theoretical decisions have to be taken very early in the corpus construction process. The corpus data may therefore require corrections during analysis; in any case, a close link between recording (secondary data), transcription (tertiary data) and contextual information (metadata) is methodologically desirable
- Its base is *complex* insofar as it involves parallel relations on the lowest structural level. Standard OHCO processing is therefore not sufficient for spoken language corpora, a more complex data model is required
- Since time-consuming manual methods prevail in the construction of spoken language corpora, tools and workflows must be optimised for speed and efficiency in order to attain adequate corpus sizes.

Translating this into requirements for corpus technology for spoken language corpora, it can be said that such technology must:

- be theory aware;
- keep a close link between recording, transcription and meta-data;
- use a data model which can naturally represent parallel temporal relationships; and,
- consider questions of speed and efficiency.

The following section will demonstrate how these requirements are met in the EXMARaLDA system.

3. EXMARaLDA

EXMARaLDA (Extensible Markup Language for Discourse Analysis) is a system of data models, data formats and software tools for the construction and analysis of spoken language corpora. It has been under development since 2000 in a project at the Special Research Centre on Multilingualism at the University of Hamburg.

EXMARaLDA is a data-centric system, that is, it is designed and implemented around a central data model (rather than, say, a specific workflow or a specific piece of software). In accordance with the above-mentioned considerations, the data model is optimised for the representation of structural relations occurring in spoken language. It is based on Bird and Liberman's (2001) idea of annotation graphs, allowing an intuitive encoding of temporally or otherwise parallel structures. Moreover, EXMARaLDA draws a distinction between temporal and linguistic entities of description. Since the former are much less dependent on a specific theoretic approach, the system can provide a set of operations applicable across theories, imposing theory-dependent structure only where it is necessary.[5]

To ensure maximal reusability of data, EXMARaLDA uses open standards like XML and Unicode in its data formats, and it provides interfaces to the most important other systems (like Praat, ELAN, CHAT, TEI) as well as to standard desktop software (Microsoft Word, Internet Browsers) for optimal interoperability. Software tools are programmed in JAVA so that they can be used on all major operating systems (Windows, Linux, Mac).

EXMARaLDA's main application areas are discourse and conversation analysis, first and second language acquisition studies, and dialectology. In addition to that, the system has also been used for multimodal analyses, phonological or phonetic studies and for the annotation of written data.

3.1 Transcription: Partitur-Editor

EXMARaLDA's transcription tool is the Partitur-Editor (Figure 1), a tool for entering and editing transcriptions in musical score notation. During transcription or in a separate step, the transcribed text can be linked to the underlying audio or video file by setting appropriate timestamps in the transcription's timeline. The interface is

5 In systems like CHAT (MacWhinney 2001) on the other hand, theory-dependent concepts like utterances are so central to the system's functionality that it is impossible to encode theory-independent structure separately.

based on the temporal structure relations alone so that the editor can be used independently of a specific theoretical approach.[6] After transcription has been completed, a separate processing step—'segmentation'—is used to calculate the linguistic structure, based on the regularities of established transcription systems (currently, HIAT, GAT, CHAT and DIDA are supported).

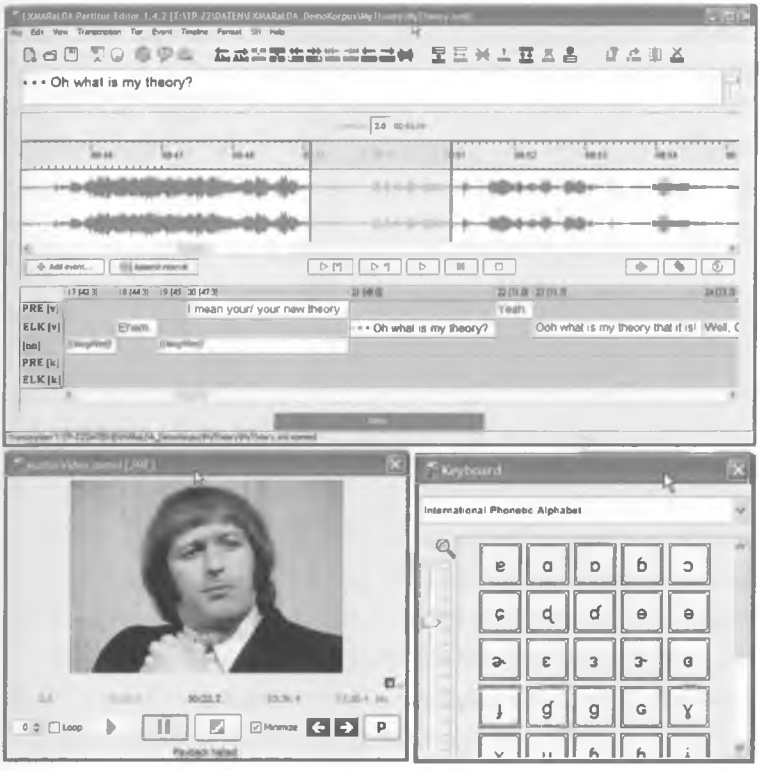

Figure 1: User interface of the EXMARaLDA Partitur-Editor

6 While the Partitur-Editor thus fulfils several of the above-stated requirements—it allows to keep a close link between transcription and recording, and it is 'theory-aware'—it should be noted that some of its flexibility is paid for in terms of speed and efficiency of transcribing. After testing the editor in a number of corpus construction scenarios, we found that its main drawback in this respect is its non-optimal use of screen real estate: like other multi-layer-tools (e.g. Praat and ELAN), the horizontal, musical-score-like layout of the interface means that transcribers can only look at small stretches of transcription text at a time and thus get a much less text-like experience than they normally have in the vertically organised layout of a standard word processor. The FOLKER transcription tool (http://agd.ids-mannheim.de/html/folker.shtml), developed on the basis of EXMARaLDA for the FOLK corpus of the IDS Mannheim, provides the user with the possibility to switch between a horizontally (musical score) and two (segment and contribution list) vertically-organised views. First experiments show that transcription can indeed be sped up considerably in this way.

3. 2 Metadata: Corpus Manager

In order to be able to deal with the various metadata requirements for spoken language corpora mentioned above, EXMARaLDA provides a separate tool, the "Corpus Manager" (CoMa), for bundling larger sets transcriptions into a corpus and describing its components through appropriate metadata sets. Coma models a corpus as a set of communications, each of which can consist of one or several recordings and corresponding transcriptions (Figure 2). Speakers are kept in a separate list and are assigned to communications in an n:m relation. In that way, it becomes possible to represent one speaker's participation in several communications as well as the fact that one communication usually involves more than one speaker, with the effect that unnecessary duplication of metadata is avoided.

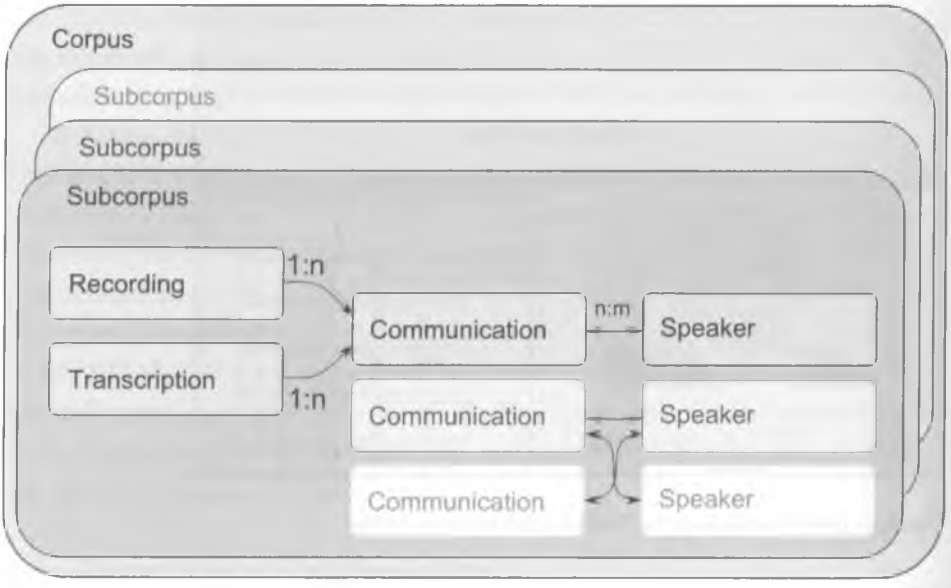

Figure 2: CoMa data model

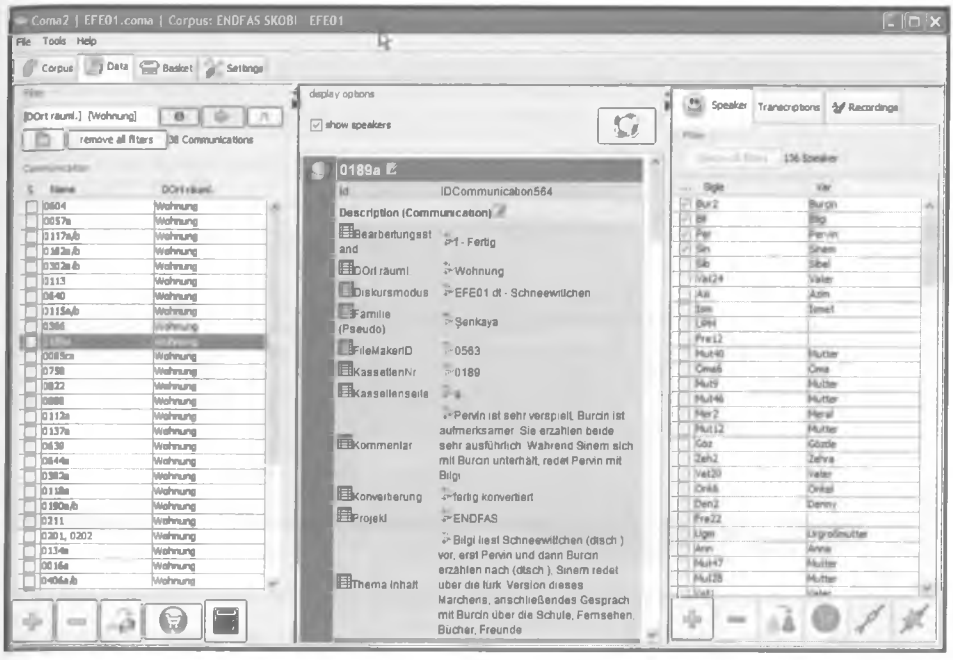

Figure 3: User interface of the Corpus Manager

The Corpus Manager presents communications, speakers, recordings and transcriptions in a graphical user interface (Figure 3), allowing the user to enter metadata for each of them, either in the form of freely definable attribute-value pairs, or as one of several pre-defined data types (e.g. location, language). A filtering mechanism can be applied to select specific communications or speakers on the basis of the meta-data (e.g. all communications taking place in Turkey, only speakers younger than 20 years), and to extract a subcorpus for that selection.

3.3 Query: EXAKT

For querying and analysing corpora, EXMARaLDA provides the "EXMARaLDA Analysis and Concordance Tool" (EXAKT). The basic functionality of EXAKT is modelled after the classical corpus analysis instrument: a KWIC (keyword in context) concordancer. After having loaded a corpus compiled in the Corpus Manager, users can enter a search expression. Several types of search expressions are offered, the most common of which is a regular expression, that is, a pattern specifying a string or a set of strings (e.g. "\b[A-Za-z]+(ing|ed)\b" for all words ending in 'ing' or 'ed').

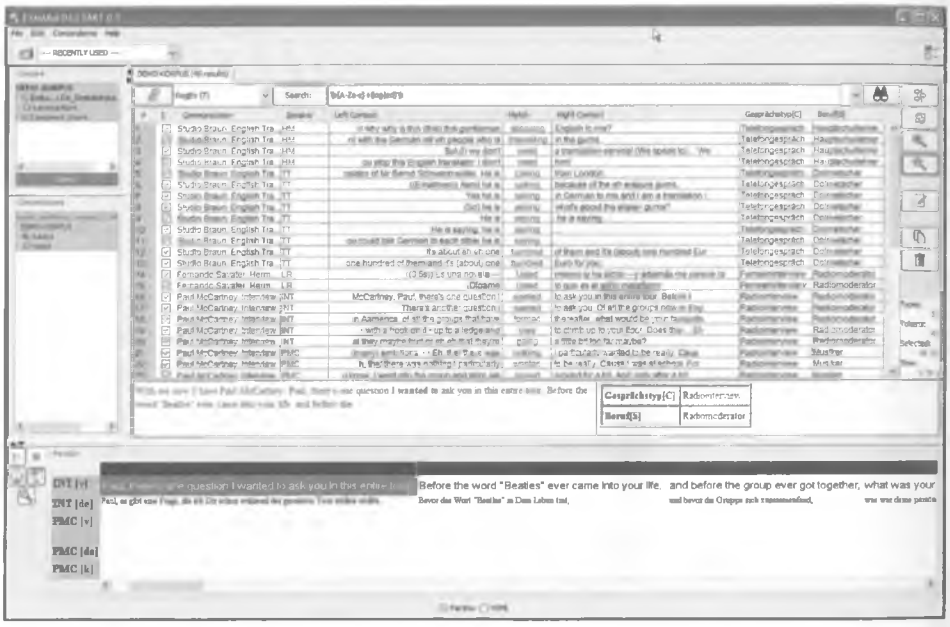

Figure 4: User interface of EXAKT

As Figure 4 demonstrates, the result of such a query is first presented as a keyword in context concordance, consisting of the matched expression itself with its immediately preceding and following context, typically the words uttered by the same speaker right before and after the word(s) matched by the search expression. As in other concordancing tools, this result can then be sorted by the left or right context column in order to facilitate the discovery of context regularities.

As discussed above, however, the analysis of spoken language data often requires additional context data, such as information about the place and time of the interaction or about a speaker's biography. For additional interactional context, EXAKT offers the possibility to display the corresponding part of a full musical score transcription (or the full transcription in some other layout) by double-clicking on any search result. Similarly, the corresponding part of the audio or video recording can be played back. In order to access meta-data about communications and speakers (as entered in CoMa), users can select arbitrary attributes to be displayed in additional columns of the KWIC table.

Spoken language research is often of an explorative nature, that is, researchers do not approach the data with an *a priori* hypothesis in mind, but rather derive their

hypotheses through a step-by-step interaction with the data. EXAKT supports such a 'corpus-driven' (rather than just 'corpus-based') approach by allowing a stepwise filtering, manual annotation and selection, and combination of search results.

4. Conclusion

The first part of this paper has discussed a few requirements corpus technology must fulfil in order to be used for the creation and the work with spoken language corpora. In summary, these requirements are:

- theory-awareness, that is, the recognition of the fact that spoken language corpora, more than written language corpora, are theory-dependent models of linguistic facts;
- a data model which adequately deals with the special structural relations occurring in spoken language; and,
- a dedicated approach to supporting quick and efficient data creation.

The EXMARaLDA system demonstrated in the second part of this paper attempts to meet these requirements by:

- using a time-based, rather than a hierarchy-based data model;
- separating theory-dependent from theory-independent constructs in the transcription interface and data model;
- supporting several widely used transcription systems as the embodiment of different theoretical approaches to spoken language;
- keeping a close link between recordings, transcriptions and metadata; and,
- paying attention to speed and efficiency in the transcription process.

Hopefully, EXMARaLDA can thus make a contribution to prevent spoken language from remaining a 'lesser' studied type of language in corpus linguistics.

References

Biber, D. (1993). "Representativeness in Corpus Design", *Linguistic and Literary Computing,* 8(4), 243–257.

De Rose, S. / Durand, D. / Mylonas, E. / Renear, A. (1990). "What is Text, Really?", *Journal of Computing in Higher Education,* 1(2), 3–26.

Geyken, A. (2009). "The DWDS corpus: A reference corpus for the German language of the 20th century" in Fellbaum, C. (ed.) (2009*). Idioms and Collocations: Corpus-based Linguistic, Lexicographic Studies*. Continuum Press.

Koch, P. / Oesterreicher W. (1985). "Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgebrauch", *Romanistisches Jahrbuch,* 36 S, 15–43.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* Mahwah, NJ: Lawrence Erlbaum.

Ochs, E. (1979). "Transcription as Theory" in Ochs E. / Schieffelin B. B. (eds.) (1979*). Developmental Pragmatics*. New York, San Francisco, London: Academic Press, 43–72.

Oostdijk, N. / Broeder D. (2003). "The Spoken Dutch Corpus and Its Exploitation Environment" in *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora* (LINC-03), April, 14, 2003. Budapest, Hungary.

Rehbein, J. / Schmidt, T. / Meyer, B. / Watzke, F. / Herkenrath, A. (2004). "Handbuch für das computergestützte Transkribieren nach HIAT" in *Arbeiten zur Mehrsprachigkeit*, Folge B, 561ff.

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

Schmidt, T. (2005b). "Modellbildung und Modellierungsparadigmen in der computergestützten Korpusanalyse" in Fisseni, B. / Schmitz, H. / Schröder, B. / Wagner, P. (eds.) (2005*) Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*. Beiträge zur GLDV-Tagung 2005 in Bonn. Sprache, Sprechen und Computer – Computer Studies in Language and Speech 8. Frankfurt a. M.: Peter Lang.

Schmidt, T. (2005c). "Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech" in *Arbeiten zur Mehrsprachigkeit*, Folge B, 62.

Selting, M. / Auer, P. / Barden, B. / Bergmann, J. / Couper-Kuhlen, E. / Günthner, S. / Meier, C. / Quasthoff, U. / Schlobinski, P. / Uhmann, S. (1998). "Gesprächsanalytisches Transkriptionssystem (GAT)", *Linguistische Berichte,* 173, 91–122.

Stachowiak, H. (1973). *Allgemeine Modelltheorie.* Wien u. a.: Springer.