New and future developments in EXMARaLDA

Thomas Schmidt, Kai Wörner, Hanna Hedeland, Timm Lehmberg

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

Abstract

We present some recent and planned future developments in EXMARaLDA, a system for creating, managing, analysing and publishing spoken language corpora. The new functionality concerns the areas of transcription and annotation, corpus management, query mechanisms, interoperability and corpus deployment. Future work is planned in the areas of automatic annotation, standardisation and workflow management.

Keywords: annotation tools, corpora, spoken language, digital infrastructure

1. Introduction

EXMARaLDA[1] (Schmidt & Wörner, 2009) is a system for creating, managing, analysing and publishing spoken language corpora. It was developed at the Research Centre on Multilingualism (SFB 538) between 2000 and 2011. EXMARaLDA is based on a data model for time-aligned multi-layer annotations of audio or video data, following the general idea of the annotation graph framework (Bird & Liberman, 2001). It uses open standards (XML, Unicode) for data storage, is largely compatible with many other widely used media annotation tools (e.g. ELAN, Transcriber, CLAN) and can be used with all major operating systems (Windows, Macintosh, Linux). The principal software components of the system are a transcription editor (Partitur-Editor), a corpus management tool (Corpus Manager) and a KWIC concordancing tool (EXAKT).

EXMARaLDA has been used to construct the corpus collection of the Research Centre on Multilingualism comprising 23 multilingual corpora of spoken language (see Hedeland et al., this volume). It is also used for several larger corpus projects outside Hamburg such as the METU corpus of Spoken Turkish[2] (Middle Eastern Technical University Ankara, see Ruhi et al., this volume), the GEWISS corpus of spoken academic discourse[3] (Universities of Leipzig, Wroclaw and Aston), the Corpus of Northern German Language Variation[4] (SiN – Universities of Hamburg, Bielefeld, Frankfurt/O., Münster, Kiel and Potsdam) and the Corpus of Spoken Language in the Ruhrgebiet[5] (KgSR, University of Bochum).

This paper focuses on new functionality added or improved during the last two years and sketches some plans for the future development of the system.

2. New and improved functionality

2.1. Transcription and annotation

The Partitur-Editor now provides additional support for time alignment of transcription and audio and/or video in the form of a time-based visualisation of the media signal. Navigation in this visualization is synchronised with navigation in the transcript, and the visualization can be used to specify the temporal extent of new annotations and to modify the start and end points of existing annotations. This has turned out a way to significantly improve transcription speed and accuracy.

[1] http://www.exmaralda.org

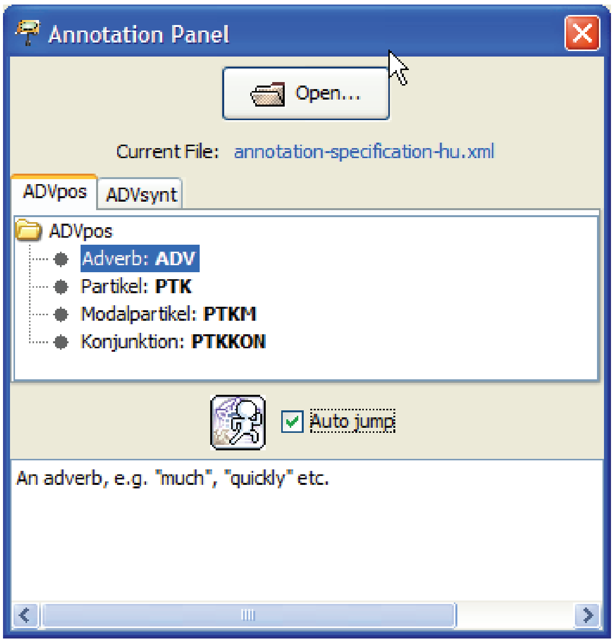
[2] http://std.metu.edu.tr/

[3] https://gewiss.uni-leipzig.de/de/

[4] http://sin.sign-lang.uni-hamburg.de/drupal/

[5] http://www.ruhr-uni-bochum.de/kgsr/

Similarly, systematic manual annotation with (closed) tag sets is now supported through a configurable annotation panel which allows the user to define one or several hierarchical tag sets, assign tags to keyboard shortcuts and link them to specific labels of annotation layers. It is also possible to specify dependencies between different tag sets so that the user is offered only those tags which are applicable in a certain context. Annotation speed and consistency can thus be improved considerably.



Current File: annotation-specification-hu.xml

- ADVpos
  - Adverb: **ADV**
  - Partikel: **PTK**
  - Modalpartikel: **PTKM**
  - Konjunktion: **PTKKON**

☑ Auto jump

An adverb, e.g. "much", "quickly" etc.

Figure 1: Annotation Panel in the Partitur-Editor

For large scale standoff annotation of corpora, a separate tool – Sextant (Standoff EXMARaLDA Transcription Annotation Tool, Wörner, 2010) – was added to the system's tool suite. Sextant can be used to efficiently add standoff tags from closed tag sets to a segmented EXMARaLDA transcription. Annotations are stored as TEI conformant feature structures which point into transcriptions via ID references. For further processing, the standoff annotation can also be integrated into the main file.

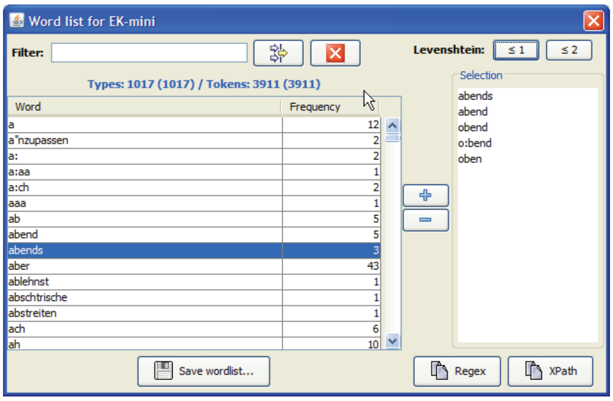
2.2. Corpus management

The Corpus Manager was supplemented with a set of operations to aid in the maintenance of transcriptions, recordings and metadata. This includes functionality for checking the structural consistency (e.g. temporal integrity of time-alignment, correct assignment of annotations to primary layers etc.), the validity of transcriptions with respect to a given transcription convention, as well as the completeness and consistency of metadata descriptions. Furthermore, a set of analysis functions operating on a corpus as a whole was added. Users can now generate and manipulate global type/token and frequency lists for a given corpus, perform global search and replace routines or generate corpus statistics according to different parameters. These new features are intended to facilitate both corpus construction and corpus use.

2.3. Query mechanisms

For the query tool EXAKT, several new features were added to support the user in formulating complex queries to a corpus.

A Levenshtein calculation was made available which selects from a given list of words all entries which are sufficiently similar to a form selected by the user. This can help to minimize the risk that (potentially unpredictable) variants – as are common in spoken language corpora – are accidentally overlooked in queries.



Types: 1017 (1017) / Tokens: 3911 (3911)

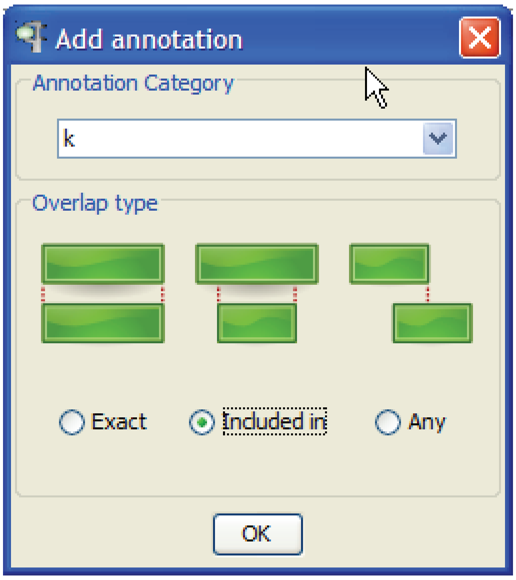
Word	Frequency
a	12
a"nzupassen	2
a:	2
a:aa	1
a:ch	2
aaa	1
ab	5
abend	5
abends	3
aber	43
ablehnst	1
abschtrische	1
abstreiten	1
ach	6
ah	10

Selection:
abends
abend
obend
o:bend
oben

Figure 2: Word list with Levenshtein functionality in EXAKT

A regular expression library can now be used to store and retrieve common queries. This is meant mainly as a help for those users who are not experts in the design of formal queries.

Through an extension of the original KWIC functionality, EXAKT is now also able to carry out queries across two or more annotation layers. This is achieved by adding one or more so called annotation columns in which annotation data from a specified annotation level overlapping with the existing search results are added to the concordance. The type of overlap between annotations can be specified as illustrated in figure 3.



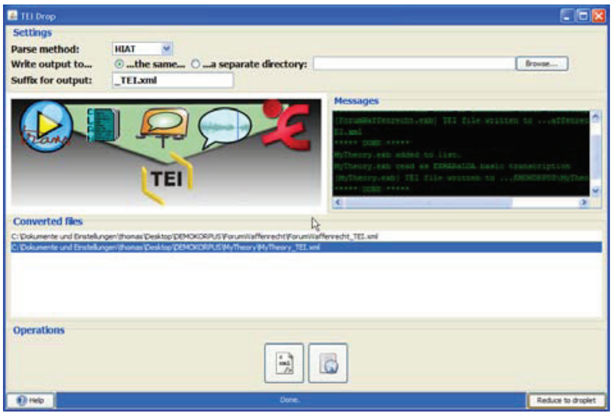
Annotation Category: k

Overlap type: ○ Exact ● Included in ○ Any

Figure 3: Specifiying the overlaptype for a multilevel search in EXAKT

2.4. Interoperability

Much work has been invested to further improve and optimise EXMARaLDA's compatibility with other widely used transcription and annotation tools. Wizards for importing entire corpora from Transcriber, FOLKER, CLAN and ELAN were integrated into EXAKT thereby considerably extending the tool's area of application. Moreover, a proposal for a spoken language transcription standard based on the P5 version of the TEI guidelines was formulated (Schmidt, 2011), and a droplet application (TEI-Drop) was added to the EXMARaLDA toolset which enables users to easily transform Transcriber, FOLKER, CLAN, ELAN or EXMARaLDA files into this TEI conformant format.



Parse method: HIAT

Write output to... ● ...the same... ○ ...a separate directory:

Suffix for output: _TEI.xml

```
[ForumWaffenrecht.exb] TEI file written to ...affenrec
EI.xml
***** DONE *****
MyTheory.exb added to list.
MyTheory.exb read as EXMARaLDA basic transcription
[MyTheory.exb] TEI file written to ...DEMOKORPUS\MyTheo
***** DONE *****
```

C:\Dokumente und Einstellungen\thomas\Desktop\DEMOKORPUS\ForumWaffenrecht\ForumWaffenrecht_TEI.xml
C:\Dokumente und Einstellungen\thomas\Desktop\DEMOKORPUS\MyTheory\MyTheory_TEI.xml

Figure 4: Screenshot of TEI-Drop

2.5. Corpus deployment

Completed EXMARaLDA corpora can now also be made available (i.e. queried) via a relational database with EXAKT. Compared to the deployment in the form of individual XML files which are then queried either locally or via http with EXAKT, this method not only facilitates data access, but also considerably improves query performance (by a factor of about 10 for smaller corpora, probably more for larger corpora) and allows for a more fine-grained access management. Furthermore, making data available in this way is also a prerequisite for integrating EXMARaLDA data into evolving distributed infrastructures like CLARIN.

With the general availability of HTML5, methods for visualizing corpus data for web presentations could also be simplified and improved considerably. The integration of transcription text and underlying audio or video recording now no longer depends on Flash technology, but can be efficiently realised with standard browser technology.

3. Future work

With the end of the maximum funding period of the Research Centre on Multilingualism in June 2011, EXMARaLDA's original context of development has also ceased to exist. Although the system is now in a stable state and should remain usable for quite some time with some minimal maintenance work, we still see much potential for future development in at least three areas.

3.1. Automatic annotation

Additional manual and automatic annotation methods are required in order to make spoken language corpora more useful for corpus linguistic research. We have consequently started to explore the application of methods developed for written language, such as automatic part-of-speech-tagging or lemmatisation to EXMARaLDA corpora.

First tests were carried out on the Hamburg Map Task Corpus (HAMATAC, Hedeland & Schmidt, 2012) with TreeTagger (Schmid, 1995), which was integrated via the TT4J interface (Eckart de Castilho et al., 2009) into EXMARaLDA. HAMATAC was POS-tagged and lemmatised with the default German parameter file, trained on written newspaper texts. The data were first tokenized using EXMARaLDA's segmentation functionality which segments and distinguishes words, punctuation, pauses and non-phonological segments. Only words and punctuation were fed as input into the tagger in the sequence in which they occur in the transcription. The tagging results were saved as EXMARaLDA standoff annotation files which can be further processed in the Sextant tool (see above). A student assistant was instructed to manually check and correct all POS tags. An evaluation shows that roughly 80% of POS tags were assigned correctly. The error rate is thus considerably higher than for the best results which can be obtained on written texts (about 97% correct tags). By far the most tagging errors, however, occurred with word forms which are specific to spoken language, such as hesitation markers ("äh", "ähm"), interjections and incomplete forms (cut-off words). Since especially the former are highly frequent but very limited in form (three forms *äh*, *ähm* and *hm* account for about half of the tagging errors), we expect a retraining of the TreeTagger parameter file on the corrected data to lead to a much lower error rate.

3.2. Standardisation

Further work in standardisation of data models, metadata descriptions, file formats and transcription conventions is needed in order to integrate spoken language data on equal footing with written data into the language resource landscape. EXMARaLDA as one of the most interoperable systems of its kind already provides a solid basis for developing and establishing such standards. Future work in this area should attempt to consolidate this basis with more general approaches like the guidelines of the Text Encoding Initiative, standardisation efforts within the ISO framework and emerging standards for digital infrastructures.

3.3. Workflow management

As we survey, train and support users in constructing and analysing spoken language corpora with EXMARaLDA, we observe how important it is to organise the tools' functionalities into an efficient workflow. Right now, the EXMARaLDA tools operate in a standalone fashion on local file systems, leaving many important aspects of the workflow (e.g. version control, consistency checking etc.) in the users' responsibility. A tight integration of the tools with a repository solution may make it much easier, especially for larger projects, to organise their workflows and construct and publish their corpora in a maximally efficient and effective manner. We plan to explore this possibility further in the follow-up projects at the Hamburg Centre for Language Corpora (HZSK).[6]

4. Acknowledgements

Work on EXMARaLDA was funded by the University of Hamburg and by grants from the Deutsche Forschungsgemeinschaft (DFG).

5. References

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

Eckart de Castilho, R., Holtz, M., Teich, E. (2009): Computational support for corpus analysis work flows: The case of integrating automatic and manual annotations. In: Lingustic Processing Pipelines Workshop at GSCL 2009 - Book of Abstracts (electronic proceedings), October 2009.

Hedeland, H., Schmidt, T. (2012): Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German. To appear in: Schmidt, T., Wörner, K.: Multilingual Corpora and Multilingual Corpus Analysis. To appear as part of the series 'Hamburg Studies in Multilingualism' (HSM). Amsterdam: Benjamins.

Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. March 1995.

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

Schmidt, T. (2011): A TEI-based approach to standardising spoken language transcription. In: Journal of the Text Encoding Initiative (1).

Wörner, K. (2010): Werkzeuge zur flachen Annotation von Transkriptionen gesprochener Sprache. PhD Thesis, Universität Bielefeld, http://bieson.ub.uni-bielefeld.de/volltexte/2010/1669/.

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