LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 7 (2013), pp. 31-40
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/4571

EXMARaLDA

Developed at the Research Center on Multilingualism, University of Hamburg, Germany

Reviewed by Cordula Meißner, *University of Leipzig,*
and Adriana Slavcheva, *University of Leipzig*

1. INTRODUCTION. EXMARaLDA is a system for creating, managing and analyzing spoken language corpora (Schmidt & Wörner 2009, Schmidt et al. 2011), developed between 2000 and 2011at the Research Centre on Multilingualism (SFB 538) at the University of Hamburg. It is now maintained at the Hamburg Center for Speech Corpora (HZSK)[1], and since November 2011, also in cooperation with the Archive for Spoken German (AGD) at the Institute for the German Language (IDS) in Mannheim. It comprises tools for transcribing spoken language (Partitur-Editor), managing metadata (Corpus Manager), and querying spoken language corpora (EXAKT). The software components are freely available and operate on all platforms (Windows, Linux, Macintosh). EXMARaLDA forms the basis for 23 multilingual corpora of spoken language at the Hamburg Center for Speech Corpora. Its primary scope of application covers discourse and conversation analysis, first and second language acquisition studies, and dialectology (cf. Schmidt 2009: 158). This paper reviews the software from the perspective of its application in the GeWiss project, one of several larger corpus projects that have used EXMARaLDA.[2] As a starting point, the review will introduce the software requirements of the project, and their role in choosing the EXMARaLDA package for the creation of the GeWiss Corpus. As we worked with all three components of the software, the review will then deal in turn with the Partitur-Editor (version 1.5.1), the Corpus Manager (version 1.9), and EXAKT (version 1.1). In conclusion, some remarks concerning support and compatibility of the software will be made.

2. BACKGROUND: THE GEWISS PROJECT – SOFTWARE REQUIREMENTS. The aim of the GeWiss project[3] was to create a comparable corpus of spoken academic language, and further, one that would be searchable for scientific and pedagogical purposes via the Internet after free registration (cf. Fandrych, Meißner & Slavcheva 2012). The languages comprised in the corpus are German, English and Polish, and academic speech events where recorded by partners in three countries: Germany, The United Kingdom, and Poland. As the corpus was created from the pedagogic perspective of German as a Foreign Language it includes data from native speakers of all three languages, as well as data from

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

[2] Other corpora that have been created using EXMARaLDA are the METU Corpus of Spoken Turkish, the SiN Corpus of Northern Germany Language Variation, and the KgSR Corpus of Spoken Language in the Ruhrgebiet (cf. Schmidt et al. 2011: 253).

[3] GeWiss was funded by the Volkswagen Foundation. For more information about the project see https://gewiss.uni-leipzig.de/.

E-ISSN 1934-5275

second language learners of German. The GeWiss Corpus is aimed as an empirical basis for research in the fields of applied linguistics and language pedagogy.

The speech events recorded for the corpus were audio taped, and to be usable for analysis had first to be transcribed. For conversation analytic analysis, it is crucial to display parallel and overlapping speech in the transcript, and so we desired a tool offering musical-score-like notation, where every speaker is transcribed on a separate tier. In spoken language research, information about the context in which a speech event is embedded is important for the analysis. To equip later users with as much context information about the recorded events as possible, detailed metadata had to be collected about the setting of the speech event, the language use (e.g. the degree of spontaneity of the speech, and the occurrence rate of alternations with other languages), and the speakers themselves. Furthermore, additional materials (such as slides, handouts, etc.) were to be collected and archived in order to allow for as comprehensive an analysis as possible. These additional materials were to be linked to the speech event they were used in, and thus to organize and systematically archive this information sufficient metadata management software was required. Ideally, the corpus would also be searchable via the Internet, so a query tool was required that would fit together both with the software used to manage the metadata, as well as with that used for the transcription. Above all, the corpus was to be created for long-term use. It was therefore important to use software based on open standards and is compatible with existing and emerging standards for digital archiving. The whole package of EXMARaLDA tools seemed to meet our software requirements best, and thus we chose to apply it to the GeWiss project.

3. USING EXMARALDA. In the following sections we will describe our experience using EXMARaLDA in our work on the GeWiss Corpus. In particular, we will examine how the three subcomponents of the software (Partitur-Editor, Coma, and EXAKT) were applied to creating the corpora, and highlight some useful features and workflows that we discovered in the process.

3.1. THE EXMARALDA PARTITUR-EDITOR. The main component of the EXMARaLDA system, the Partitur-Editor, is a professional tool for transcription and annotation of verbal interactions from digital audio and/or video recordings. It uses musical score (German: Partitur) notation[4], which offers a wide variety of features applicable for complex linguistic analysis work. Since there are virtually no limitations on the number of speakers, the Partitur-Editor is perfectly suitable for the transcription of complex interactions, such as classroom interactions, discussions, and talk shows. In addition, the layout of the musical score notation allows for precise visualization of the temporal sequence and simultaneity between different speakers, as well as between the verbal and non-verbal behavior of any particular speaker. This is an important advantage, especially for multiparty speech events,

[4] In linguistics, the musical score notation describes a special way of representing spoken language in written form in order to allow for a visualization of its temporal sequence and simultaneity. Just like in a musical score, each instrumental or vocal part is represented on a separate staff in vertical alignment. Different speakers, different modalities (verbal and non-verbal behavior) etc. are written on different lines of a linguistic score. For detailed information cf. Ehlich (1992).

as they often naturally contain many instances of overlap. Furthermore, this concept can be extended to linguistic annotation of specific phenomena in a corresponding annotation tier. The Partitur-Editor is available in the German, English, French, Swedish, Turkish, Czech and Spanish languages.

The EXMARaLDA Partitur-Editor is a transcription and annotation tool for complex linguistic work, and so there are several main concepts and features to be explained in order to begin a project. The initial setup of a new file includes the entering of meta-information about the transcription (**Transcription > Meta Information**), the linking of the recording (**Transcription > Recordings**), the creation of a speaker table with all speakers involved (**Transcription > Speaker Table**), and finally the adding of tiers for every speaker (**Tier > Add Tier**). It is recommended to keep the meta-information about the transcriptions and the speakers consistent, since this information can be used later in Coma or EXAKT to create a corpus from a set of EXMARaLDA transcriptions that already have metadata stored in their transcription heads.

Once the transcription has been set up, the transcribing of monologic sequences can be performed easily by using the **Append Interval** button, fine-tuning the end point of the applied interval, and finally entering the transcription into the tier of the appropriate speaker. Moreover, the **Add Event** button allows for transcribing complex sequences with overlaps. In Figure 1 there is an overview of the main window of the graphical user interface in the Partitur-Editor.

When completed, the transcription can be formatted, checked for structural and segmentation errors, and outputted in several data formats. A word list of the transcription can be generated, the annotated time can be calculated, and an EXAKT search for specific linguistic phenomena within the actual transcription can be performed (cf. 3.3).

Even though the EXMARaLDA Partitur-Editor offers numerous features for complex analysis work with spoken data, it is reasonably easy to get familiar with the tool and to use it in practice. In the GeWiss project for example, our student assistants were able to handle the basic functions of the Partitur-Editor after a 4-hour training session, and became fluent users in just a few weeks. However, for familiarization with the Partitur-Editor in the absence of an EXMARaLDA training to attend, we recommend the video tutorial on the developer's website.

The EXMARaLDA Partitur-Editor allows users to choose between different player types (**Edit > Preferences > Media**) – DirectShow-Player, JMF-Player, JDS-Player, BAS-Audio-Player, Quicktime-Player and ELAN-Quicktime-Player – and supports a large number of media file formats, e.g. WAV, MP3, AIF, MPG, MPV, AVI, DIVX, MP4, WMV and OGG. However, for the waveform display an uncompressed WAV-file is needed.

The Partitur-Editor is Unicode compliant, so transcriptions can be made with different writing systems (e.g. in the GeWiss project, we used it for Latin alphabet transcriptions of German, English and Polish data, as well as for Cyrillic alphabet transcriptions of Bulgarian data). In addition, the keyboard panel lets the user enter symbols not included on a standard keyboard (e.g. special alphabets, IPA symbols, special transcription signs, etc.). The tool is also independent of the transcription conventions and has functions for segmentation of transcription data according to HIAT, DIDA, GAT, CHAT and IPA.

The EXMARaLDA Partitur-Editor is also designed to be interoperable with other systems, like TASX, Praat, ELAN, CHAT, FOLKER and TEI. Beyond this, the transcriptions

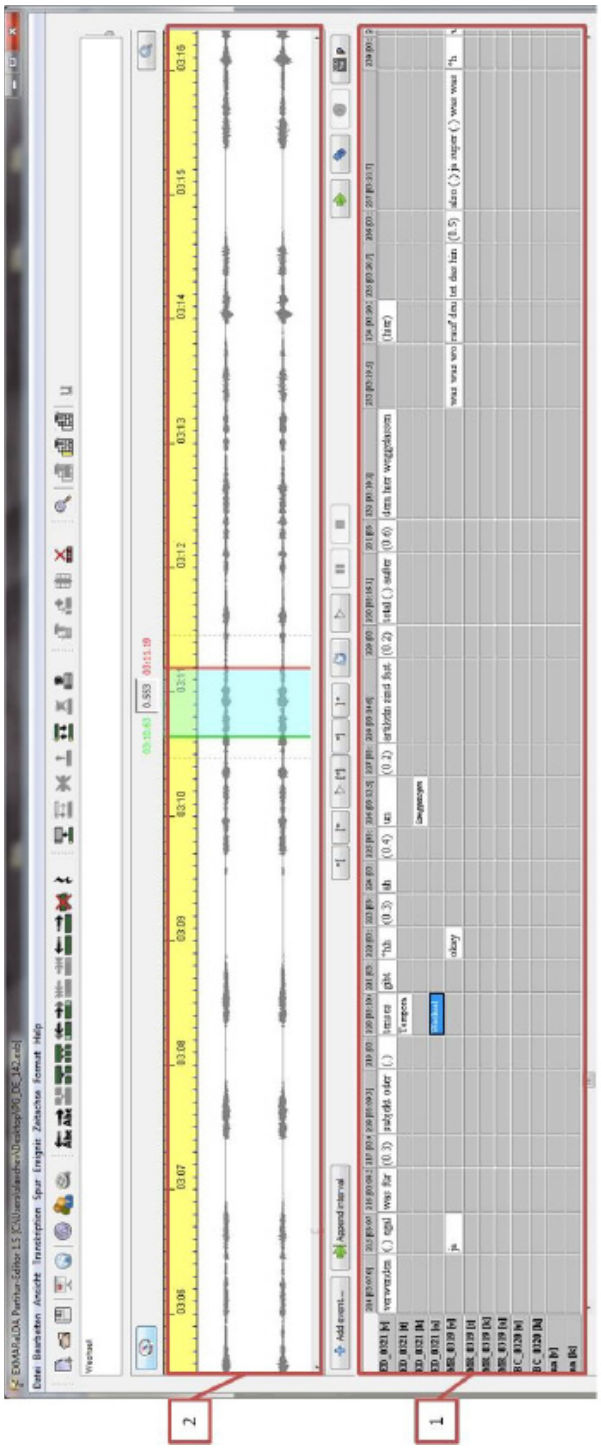

FIGURE 1. Main window of the user interface of the EXMARaLDA Partitur-Editor.
[1] partitur with different tier types for transcription and annotation; [2] waveform panel

can be outputted in several data formats, like HTML, XML, RTF, and TXT, and thus be further processed on other machines. However, since in the GeWiss project we did not use any of these programs, we cannot report on the handling of such import and export functionalities.

3.2. THE EXMARALDA CORPUS MANAGER (COMA). The Corpus Manager is a tool for archiving and organizing data into a corpus, including primary data such as recordings, as well as secondary data such as transcripts or annotations. Coma allows one to bundle all kinds of gathered data pertaining to a speech event and to integrate them into a corpus. Furthermore, the software allows one to define sub-corpora within the whole corpus and analyze them via the query tool EXAKT (cf. 3.3). In addition, Coma also offers some features for corpus maintenance.

Using the Corpus Manager the user structures the data into two broad categories: data belonging to so called communications, i.e. the individual speech events, on the one side, and metadata belonging to speakers on the other. These types of metadata are kept separate, because there is often a many-to-many relationship between speakers and speech events. Speakers participating in a communication are simply linked to that communication, they do not need to be duplicated.[5] However, things can be even more complex: In the GeWiss project it occurred that the same speaker took part in different speech events, but each time in a different role. For example, a speaker being the presenter of a conference paper in one recording could be the examiner or the lecturer in other recorded speech events. To deal with these cases, we entered all roles a speaker appeared in as part of his unitary, archived speaker metadata in Coma.

For each communication to be archived, the software offers the pre-defined metadata slots - location, language, and setting - with their specific metadata forms.[6] The category language, for example, was used in the GeWiss project to archive metadata regarding the basic language of the speech event (i.e. the language used most of the time), occurring alternations in other languages, and the degree of spontaneity (freely spoken vs. read or learned by heart). Aside from the pre-defined forms for some categories, metadata can be entered in freely definable attribute-value pairs, such as for example 'topic of the presentation - language change in the 18th century'. To facilitate the manual input of similar metadata for different speech events, Coma allows you to save previous attribute-value pairs entered once as templates and to reuse them when adding metadata for the next speech event. This feature proved to be very useful in the GeWiss project, where metadata for 380 speech events and 522 different speakers had to be entered in Coma. In addition to entering general metadata, Coma allows you to attach recordings (audio and/or video),

[5] Coma data are stored in XML files, whose top level structure - a list of communications and a list of speakers with an n:m assignment between them - can be mapped directly to a relational data structure (Thomas Schmidt, personal communication, January 8, 2013).

[6] For languages, the standardized format of the ISO language codes (ISO 639-1) is employed, and time data within the location slot is archived in a standardized format as well. At the level of the corpus as a whole, metadata slots based on Dublin Core or OLAC can be inserted automatically. In general, Coma allows for the use of standardized metadata vocabularies, but it does not enforce it (Thomas Schmidt, personal communication, January 8, 2013).

transcriptions, and additional files via hyperlink along with additional metadata for these materials as well. In the GeWiss-Project, we used all of these attachment options, because for every speech event we had at least an audio recording, often also a video recording, a transcription, and several PDF files of additional material such as slides or handouts used in the presentations. Coma offered a simple solution to systematically organize all those materials. The screenshot in Figure 2 shows the graphical interface of the Corpus Manager.

Beyond the management of corpus data, Coma can also be used to define sub-corpora for further analysis. The software includes a filter function, by which you can select communications according to specific metadata values, and the selected sub-corpus can then be analyzed directly with the EXAKT tool linked to Coma (**Analysis > Search Corpus Using EXAKT**). There are also options for creating a word list (**Analysis > Create Wordlist**) and corpus statistics, counting the segments and words per speaker and communication (**Analysis > Generate Corpus Statistics**).

Finally, Coma offers some functions for checking the consistency of the metadata and the attached transcripts. You can, for example, check for typing alternations in the names of attributes. This proved to be extremely useful in the final checking phase of the GeWiss project (**Maintenance > Harmonize Description Keys**).

3.3. THE EXMARALDA QUERY TOOL EXAKT. The third component of the EXMARaLDA system – the EXMARaLDA query tool EXAKT – allows corpus-driven (quantitative and qualitative) analyses of transcribed spoken language data.

Simply put, EXAKT is a concordancer that permits users to search for a particular expression, and outputs all instances of the expression with the immediately preceding and following context (cf. Figure 3). In order to support further corpus-driven analyses, different kinds of metadata can be displayed, the search results can be filtered (e.g. according to the metadata), and additional analyses can be added manually to the KWIC table. A special advantage for the work with spoken data is the possibility to display more interactional context (as encoded in the musical score notation of the transcription) by double-clicking on any search result, and listening to the corresponding part of the transcription-aligned recording. Finally, all search results and additional analyses can be saved or exported to other applications (e.g. Excel) for further statistical analyses. For transcriptions, which have been segmented for words, EXAKT also can generate a word list of the corpus.

Besides using EXAKT as a desktop application for analyzing data stored on the local computer, the tool can also be run as a Java Applet for corpus releases with differing functionalities.[7]

EXAKT can handle transcriptions created with the most common transcription tools – not only with EXMARaLDA, but also ELAN, CHAT, Transcriber, and FOLKER. However, since the transcriptions in the GeWiss project were created with the EXMARaLDA Partitur-Editor, we tested the described functionalities of EXAKT only for the EXMARaLDA corpora.

[7] An EXAKT Java Applet has been developed for the online use of the EXMARaLDA corpus HAMATAC (cf. http://www1.uni-hamburg.de/exmaralda/files/z2-hamatac/public/index.html).

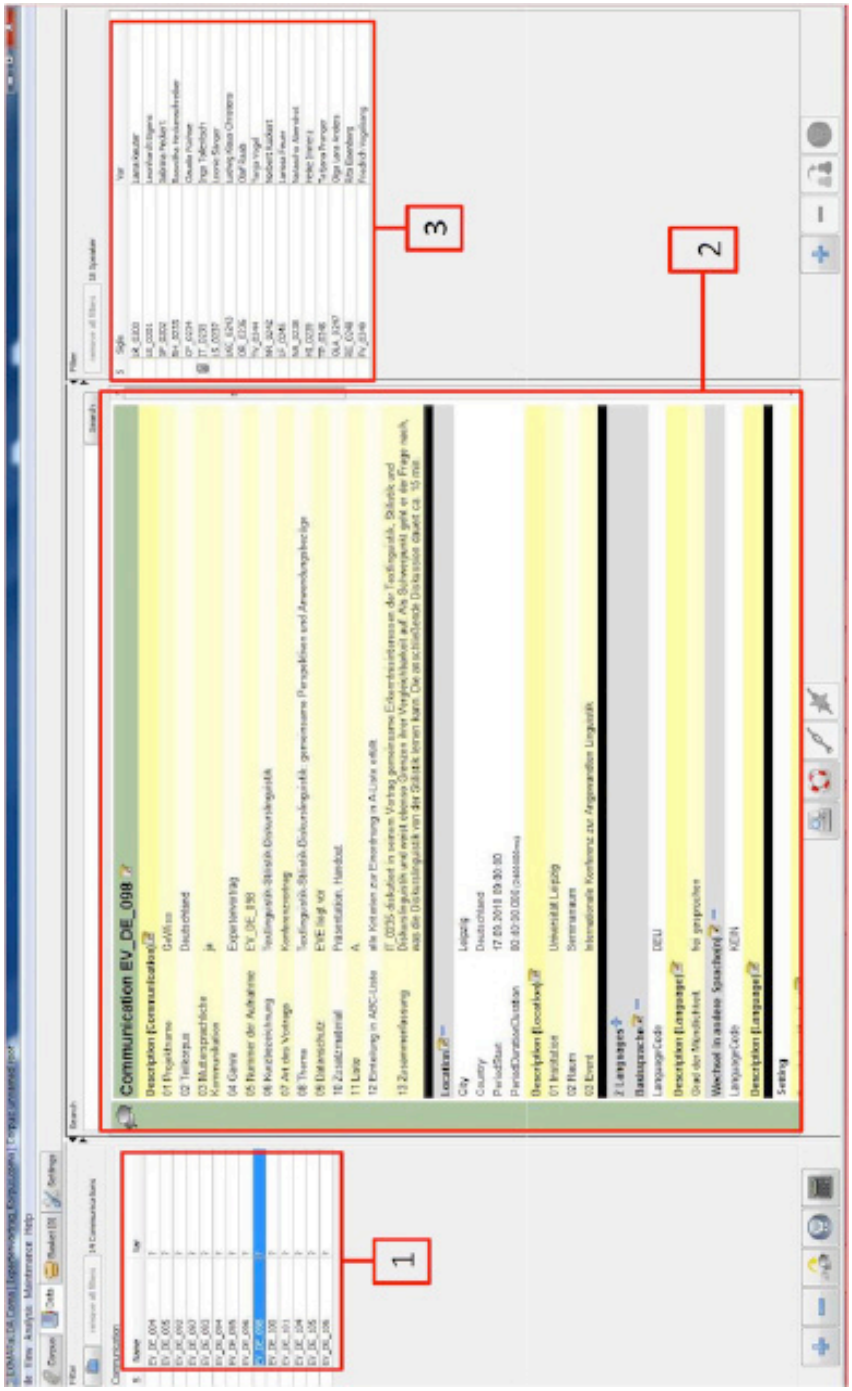

FIGURE 2. The graphical interface of the Corpus Manager.
[1] list of the communications of the corpus; [2] metadata, attached recordings and files belonging to a selected communication/speech event;
[3] list of all speakers of the corpus; speakers participating in the selected communication are marked by a paper clip.

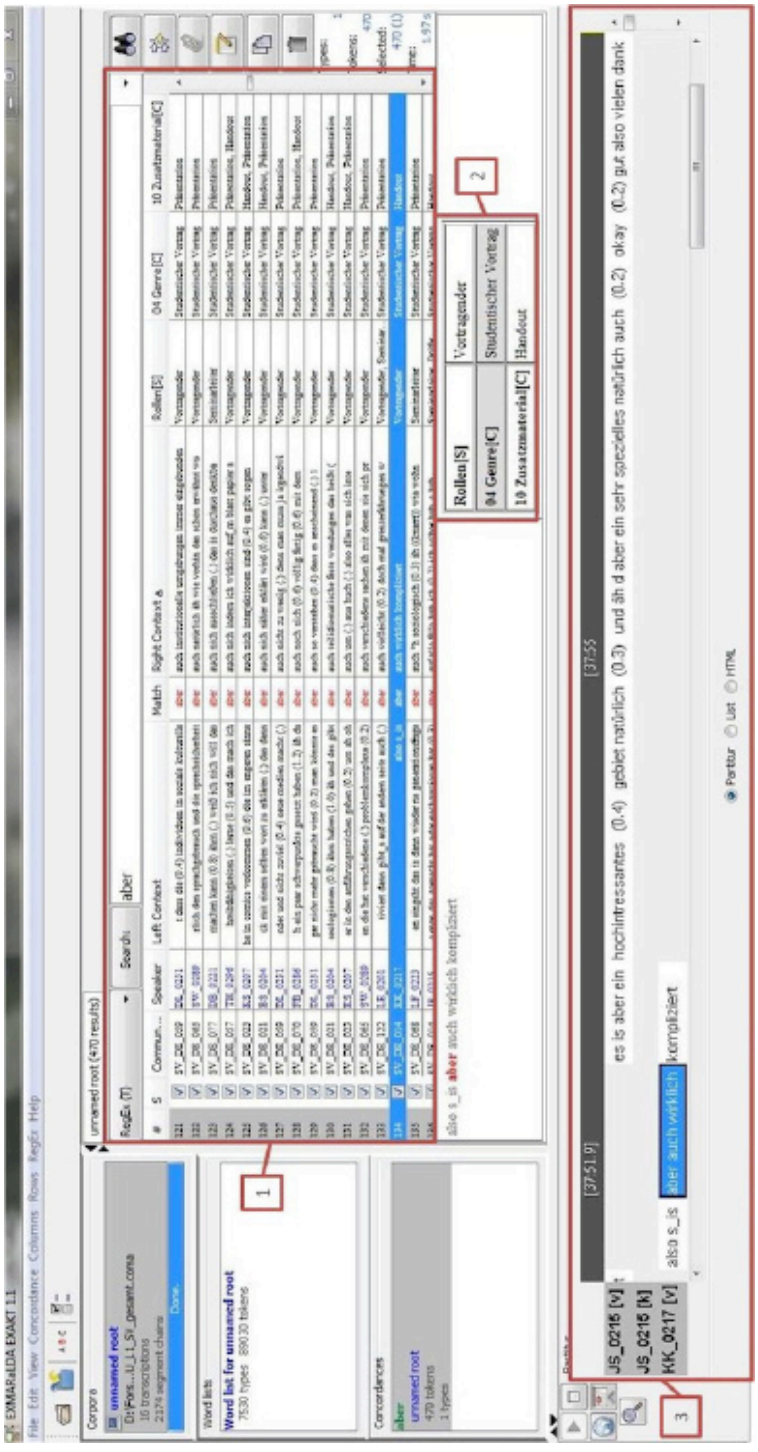

FIGURE 3. The graphical user interface of EXAKT.
[1] KWIC concordance; [2] additional metadata displayed; [3] interactional context displayed

4. SUPPORT. The support for EXMARaLDA users is substantial. All of the EXMARaLDA tools are very well documented on the developer's website, where there are detailed manuals for every tool, a quick start guide, an excellent video tutorial for the Partitur-Editor, and numerous additional tutorials. In addition, specific issues can also be discussed with the developers and the whole EXMARaLDA community on the EXMARaLDA mailing list, in both German and in English. Furthermore, the developers aim to fix bugs very quickly and are keen to include new features based on user requests.

5. COMPATIBILITY AND MAINTENANCE OF EXMARALDA. Sustainability and interoperability were the two key features of EXMARaLDA that motivated us to use the system for the GeWiss corpus. The former is guaranteed by the use of open standards like XML and Unicode for the data, and the latter by providing the ability to interface with most other common software, including ELAN, CHAT, Praat, and TEI. Moreover, due to the fact, that the software tools are programmed in JAVA, they can be used on all major operating systems (Windows, Mac, Linux).

As for the sustainability of the software, the funding period for the Research Centre on Multilingualism at the University of Hamburg, where EXMARaLDA was developed, ended in June 2011. However, the maintenance of EXMARaLDA is still secure, as EXMARaLDA has been integrated into the language resources of the evolving CLARIN-D infrastructure[8].

6. SUMMARY OF EXMARALDA. As outlined above, the creation of the GeWiss Corpus involved several key demands. In particular, the abilities to package and hyperlink volumes of external files, to create and edit complex musical score notations, and to append and search multilayered metadata proved key. EXMARaLDA was able to meet our content creation and management needs; our final thoughts and software specifications are detailed below.

Primary function:	Transcribing and annotating spoken language data, creating, managing and analyzing spoken language corpora
Secondary function:	Further annotation of language data (e.g. POS-tagging)
Pros:	EXMARaLDA is a software package that provides all tools needed to create and work with speech corpora.
Cons:	Some training is needed to use the tools efficiently.
Platforms:	Windows, Linux, Macintosh
Open Source:	Yes
Proprietary:	No
Reviewed version:	Partitur-Editor 1.5.1, Coma 1.9, EXAKT 1.1
Application size:	41 MB (complete installation file for Windows)
Available from:	http://www.exmaralda.org/downloads.html

[8] http://de.clarin.eu/de/

References

Ehlich, Konrad 1993. HIAT: A Transcription System of Discourse Data. In Jane A. Edwards & Martin D. Lampert (ed.), *Talking Data. Transcription and Coding in Discourse Research*, 123-148. Hillsdale: Erlbaum.

Fandrych, Christian, Cordula Meißner & Adriana Slavcheva 2012. The GeWiss Corpus: Comparing Spoken Academic German, English and Polish. In Thomas Schmidt & Kai Wörner (ed.), *Multilingual Corpora and Multilingual Corpus Analysis*, 319-337. Amsterdam: Benjamins.

Schmidt, Thomas 2009. Creating and Working with Spoken Language Corpora in EXMARaLDA. *Lesser Used Languages and Computer Linguistics* (LULCL) 2. 151-164.

Schmidt, Thomas & Kai Wörner 2009. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* 19. 565-582.

Schmidt, Thomas, Kai Wörner, Hanna Hedeland & Timm Lehmberg 2011. New and future developments in EXMARaLDA. *German Society for Computational Linguistics and Language Technology* (GSCL) 96. 253-256.

Cordula Meißner
cordula.meissner@uni-leipzig.de

Adriana Slavcheva
slavcheva@uni-leipzig.de