EXMARaLDA

Transcription, Annotation and Analysis of Spoken Language

EXMARaLDA

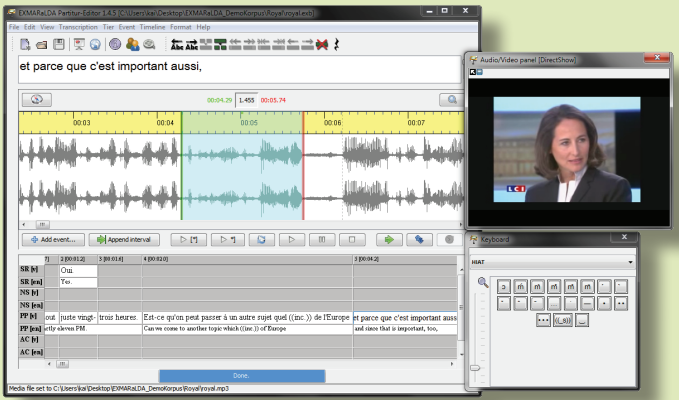
EXMARaLDA is an acronym of “Extensible Markup Language for Discourse Annotation”. It is a system of concepts, data formats and tools for the computer assisted transcription and annotation of spoken language, and for the construction and analysis of spoken language corpora.

EXMARaLDA is developed in the project “Computer assisted methods for the creation and analysis of multilingual data” at the Collaborative Research Center “Multilingualism” (Sonderforschungsbereich “Mehrsprachigkeit” - SFB 538) at the University of Hamburg. All components of the system are freely available to users outside the SFB.

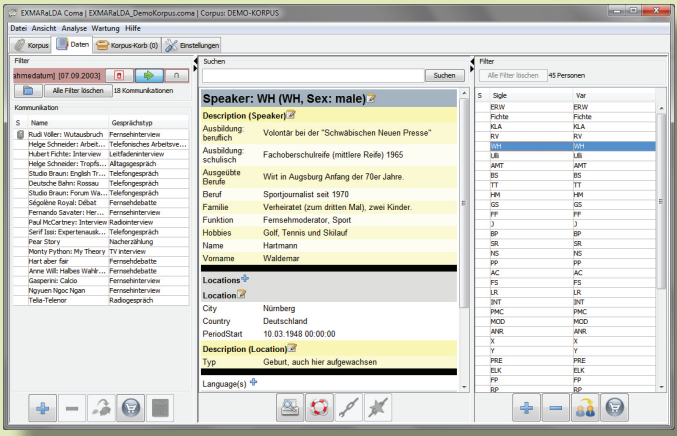
Java based tools

All software tools for creating and working with EXMARaLDA data are JAVA applications. This makes them suitable for all currently used operating systems (Windows, Macintosh, Linux, Unix).

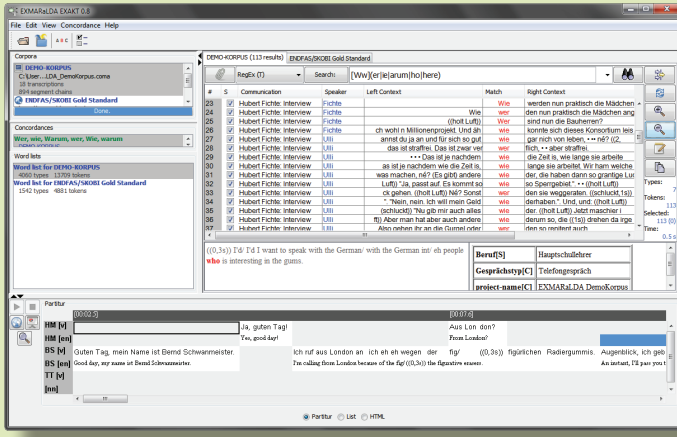
The EXMARaLDA **Partitur Editor** is a tool for inputting, editing and outputting transcriptions in partitur (musical score) notation. In addition, the editor has functions for data exchange with other systems and for segmentation of transcription data according to several transcription conventions.



The EXMARaLDA **Corpus-Manager (Coma)** is designed to assemble transcripts created with the EXMARaLDA Partiture-Editor with their corresponding recordings into corpora and enrich them with metadata. Metadata can be about speakers, communications (settings), recordings and the actual transcripts.



The EXMARaLDA query tool **EXAKT ("EXMARaLDA Analysis and Concordancing Tool")** is a tool for searching transcribed and annotated phenomena in an EXMARaLDA corpus.



Departing from a KWIC concordance, researchers can contextualize query results, correlate them with metadata, and categorize or quantify the findings.

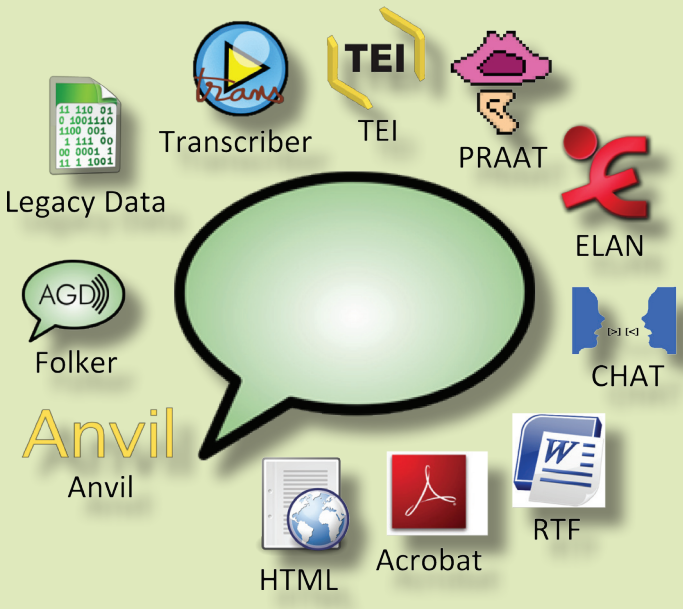
XML based data formats

All EXMARaLDA data are stored in Unicode compliant XML files. The use of this W3C standard ensures flexible usability and long-term archivability of the data.

```
<tier id="TIE0" speaker="SPK0" category="sup" type="a">
    <event start="T1" end="T3">louder </event>
    <event start="T38" end="T39">louder </event>
    <event start="T44" end="T46">in a low voice </event>
    <event start="T50" end="T52">louder </event>
</tier>
```

Interoperability

The EXMARaLDA concept is based on the annotation graph framework (Bird/Liberman 2001) and thus aims at a maximal exchangeability and reusability of transcription data. Hence, it is possible to create and edit EXMARaLDA data not only with the system's own tools, but also with other popular software.



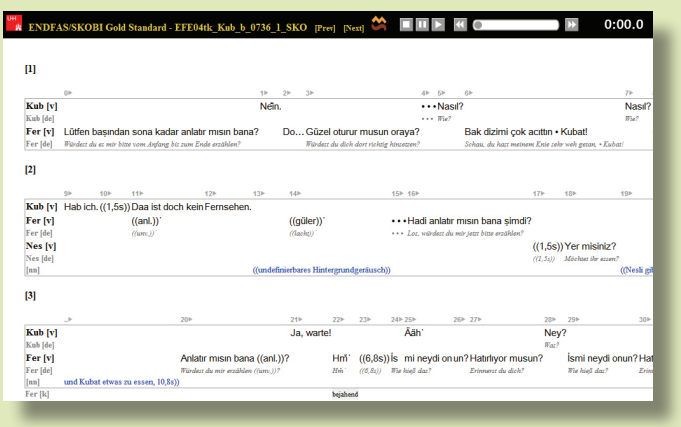
Furthermore, EXMARaLDA data can be transformed into a number of widely used presentation formats and supports several important transcription systems (HIAT, GAT, CHAT, DIDA).

EXMARaLDA Corpora

EXMARaLDA is used by discourse and conversation analysts, dialectologists, phonologists etc. to compile spoken language corpora for various purposes.

Among the corpora currently being constructed with EXMARaLDA are the METU Corpus of Spoken Turkish
[http://std.metu.edu.tr]
and the corpus of the SiN Project documenting regional variation in Northern Germany
[http://sin.sign-lang.uni-hamburg.de]

At the Research Center on Multilingualism, completed project corpora are made available to the research community via the EXMARaLDA website:
[http://corpora.exmaralda.org]



For example, registration with the corpus owners will give you access to the following resources (see the website for a complete list):

- DiK: A corpus of interpreted doctor-patient communication (German, Turkish and Portuguese).
- DUFDE and BIPODE: longitudinal studies of bilingual language acquisition in children (French-German and Portuguese-German)
- SKOBI/ENDFAS: a corpus documenting language use in Turkish/German bilingual children.

More corpora will be added as the project progresses.

Contact

Address

Research Centre 538: Multilingualism
Research Projekt Z2
Max-Brauer-Allee 60
22765 Hamburg
Germany

Phone: (+49) 40 42838 6425

E-Mail: contact@exmaralda.org

Web: www.exmaralda.org

Team Members

Hanna Hedeland
Timm Lehmberg
Thomas Schmidt
Kai Wörner

The Research Center 538 “Multilingualism” is funded by the German Science Foundation and the University of Hamburg.

Deutsche
Forschungsgemeinschaft

DFG

Universität Hamburg

June 2009