
Workshop on multilingual data, 08 July 2003
CORPUS "SCANDINAVIAN SEMICOMMUNICATION"
Thomas Schmidt, Project Zb

EXMARaLDA / Multilingual Database

Goals:

- Make data archivable
- Make data exchangeable
- Make data open for changing requirements

„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)

EXMARaLDA:

- "Interlingua" for different legacy formats (syncWriter, dBase etc.)
- XML storage of data → Longevity (Open Standard)
- Application of Annotation Graph formalism → Common formal basis for different transcription systems
- "Single Source, Multiple Target" → One data type, different input, output and analysis methods

Project K5 "Semicommunication and receptive multilingualism in contemporary Scandinavia"

Empirical basis:

- Communication between native speakers of Danish / Swedish / Norwegian using their respective mother tongue
- Recordings of
 - Radio Broadcasts (Radio Öresund)
 - Group discussions (NUAS)
 - School lessons
- Theoretical background: Discourse Analysis
- Transcription system: HIAT (Partitur notation)

Input and output method in the first phase – HIAT-DOS:

- Editor for HIAT transcription (supporting partitur notation)
- Output to RTF, printer or HTML
- Technically outdated:
 - No graphical user interface

- No mouse support
- No use of different fonts
- No media support
- Data not reusable
- Conversion to EXMARaLDA can only be done semi-automatically, requires considerable amount of manual post-editing

Input and output methods in the second phase – Praat and Partitur-Editor

- Praat: Synchronize transcription with digitized audio recording (*.wav file)
- Transcribing becomes quicker and more precise (e.g. measurement of pauses)
- Import of Praat data into the Partitur-Editor requires minimal manual post-editing
- Use the partitur editor to
 - check/verify transcription
 - add annotation tiers (e.g. for translation / code switching)
 - generate different visualizations (partitur output / list output)
 - segment the data into utterances and words
- Export to EXMARaLDA "Segmented Transcription" format
 - contains all structural information (temporal and linguistic structure)
 - basis for archiving / computer-assisted analyses (search, count etc.)

Analysis methods (to be explored)

- Corpus Browser
 - "Exploratory data analysis"
 - "Read" corpus
- Corpus Search Tools
 - Find instances for qualitative analyses
 - Count instances for quantitative analyses
 - Get context for search results
 - Correlate search results