

LREC-Conference Panel „Collaborative Commentary“, Lisbon, 27 May 2004

Database „Multilingualism“ – Perspectives for collaborative corpus construction and collaborative commentary

Thomas Schmidt

Sonderforschungsbereich 538 Mehrsprachigkeit

University of Hamburg



Background

SFB „Multilingualism“, University of Hamburg

- ❖ 13 projects organized in 3 groups (Multilingual acquisition / Multilingual communication / Historical multilingualism)
- ❖ Empirical work – corpora of written texts and corpora of transcribed recordings (video / audio), all computerized
- ❖ Roughly 2000 transcripts / 1000 hrs of transcribed speech
- ❖ “Raison d’être”: Collaboration (!)

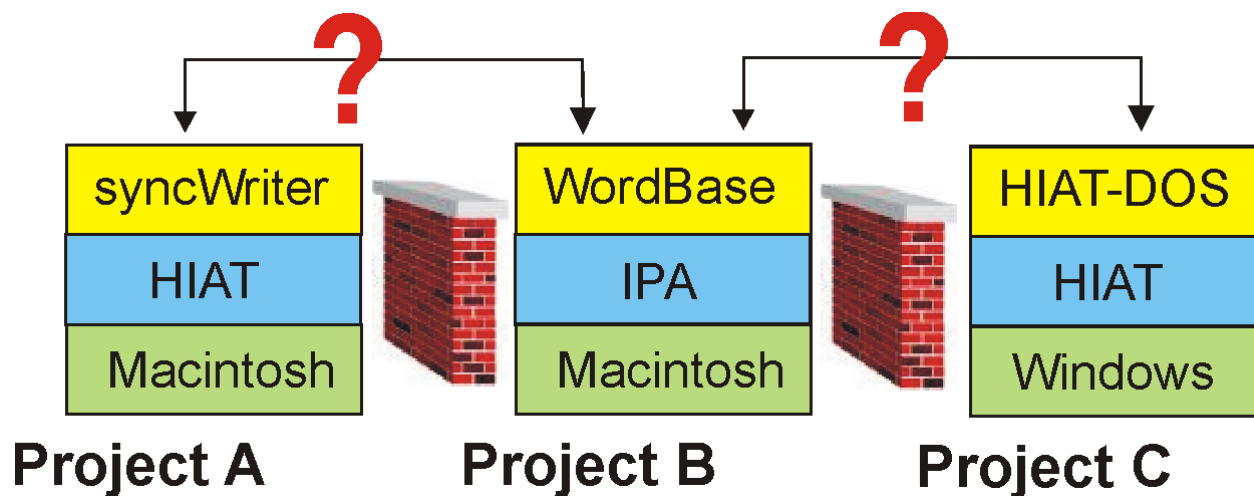
Background

Diversity of Transcription data

- ❖ Research background: Generative Grammar / Discourse Analysis / Phonetic Research
- ❖ Transcription systems: HIAT / IPA / ...
- ❖ Presentation formats: Score notation / Line notation / Column notation
- ❖ Writing systems: Latin, Greek, Cyrillic, Japanese
- ❖ Transcription software: syncWriter / WordBase / HIAT-DOS / Lapsus
- ❖ Operating systems: Windows / Macintosh / Linux

- ❖ Interrelatedness of these dimensions

Data exchange

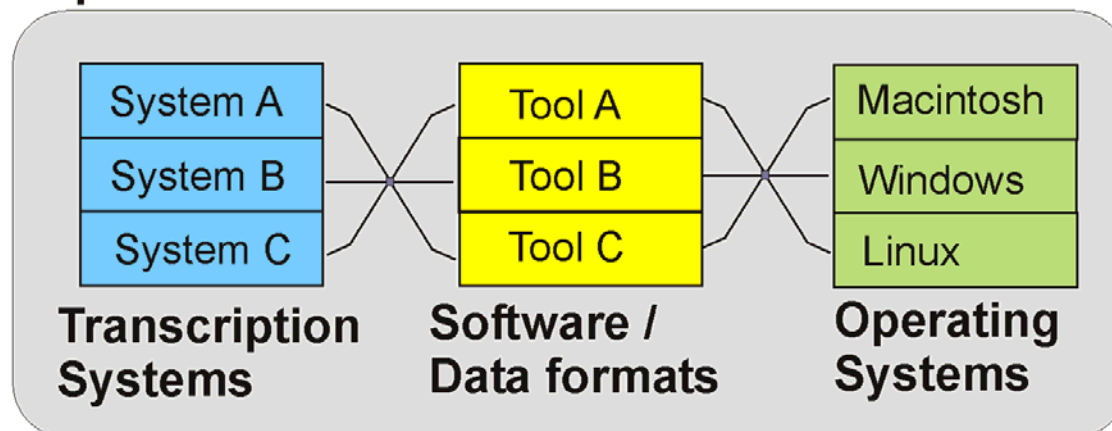


Problems:

- Use project A's data with project B's operating system?
- Use project B's tools with project C's data?
- Use project C's transcription system with project B's tools?
- Exchange corpora?
- Build larger corpora from existing ones?
- Build a common tool for all projects' data?
- Collaborative commentary?

Data exchange

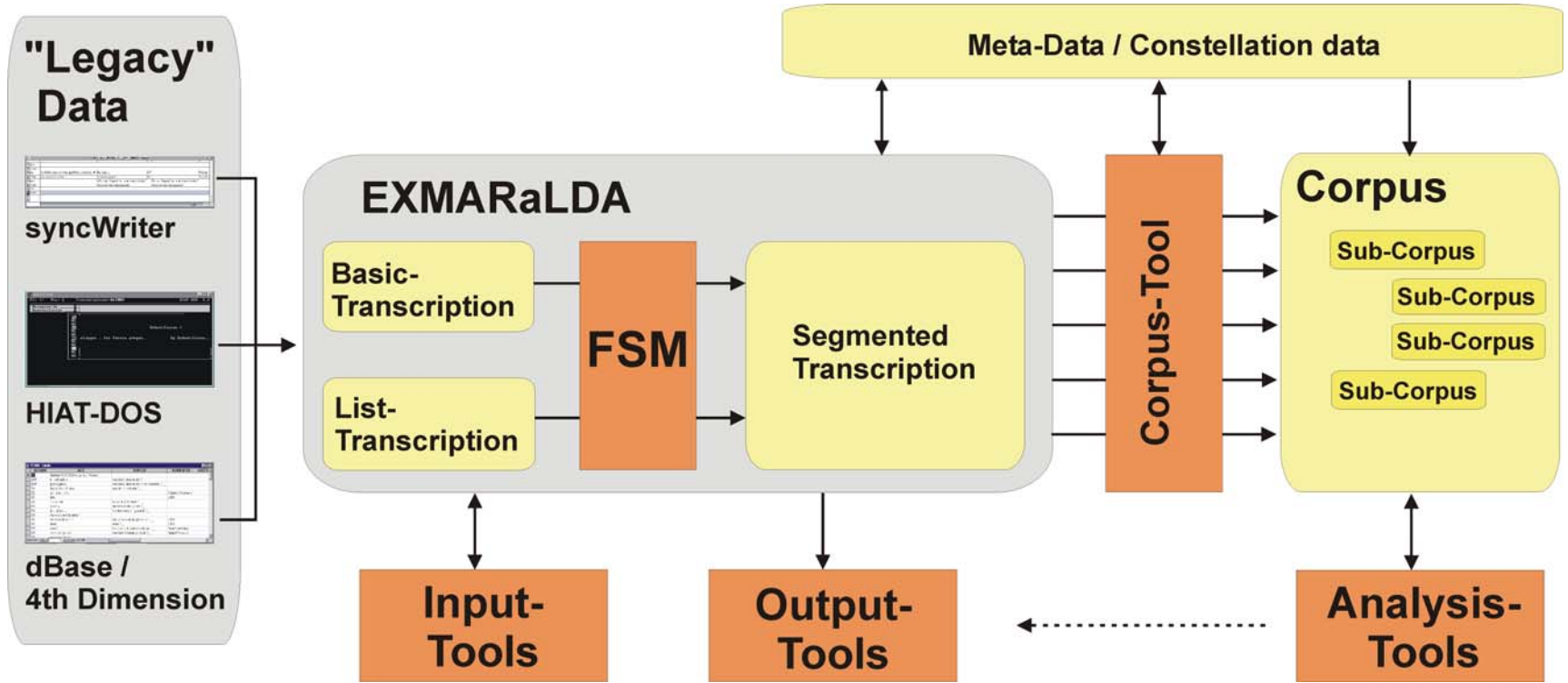
Transcription Framework



Vision: A framework for computer transcription

- Let software and formats operate on a common conception of transcription data
- Make transcription systems a parameter rather than a principle for data models
- Use standard technologies (JAVA, XML, Unicode) to achieve “platform independence”
 - Use one tool with different transcription systems
 - Use different tools with one data format
 - Use one tool on different operating systems
 - **Facilitate collaboration in corpus construction and analysis**

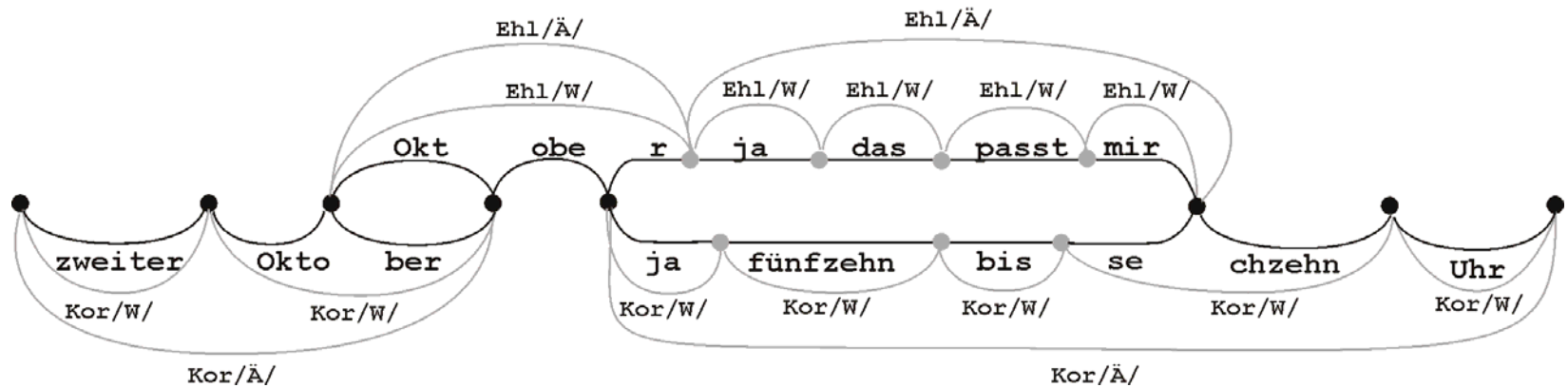
Database „Multilingualism“



System architecture

EXMARaLDA

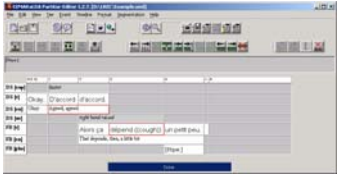
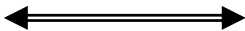
Ehl Oktober. Ja, das paßt mir.
Kor Zweiter Oktober. Ja, fünfzehn bis sechzehn Uhr.



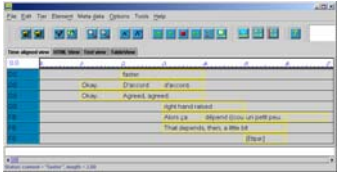
- Separate model and visualization / three level architecture
- Describe models as **Directed Acyclic Graphs**
 - Time reference of all transcription entities (Annotation Graphs)
- **Calculate** visualization(s) from model
- Store as **XML** files

EXMARaLDA

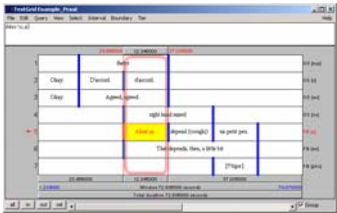
- HIAT
- GAT
- DIDA
- IPA
- CHAT



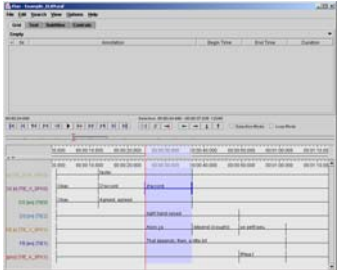
EXMARaLDA
Partitur-Editor



TASX
Annotator

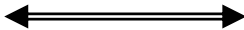


Praat



ELAN

**Collaborative
Commentary
Tools**



Collaborative Commentary I

[1]

| | | | |
|-----------------|------------------------|--------------------------|-------------------|
| Sel | | He he. | |
| Yıl | Okul kitabıdır. | O okül/ okul kitabın | içinde |
| TL-Yıl | school book-PSS3SG-COP | DEI school- book-GEN | inside-PSS3SG-LOC |
| Yıl [en] | It's a textbook. | What's in this textbook? | |

Sel [k]

affirmative

Yıl [k]

for: kitabının

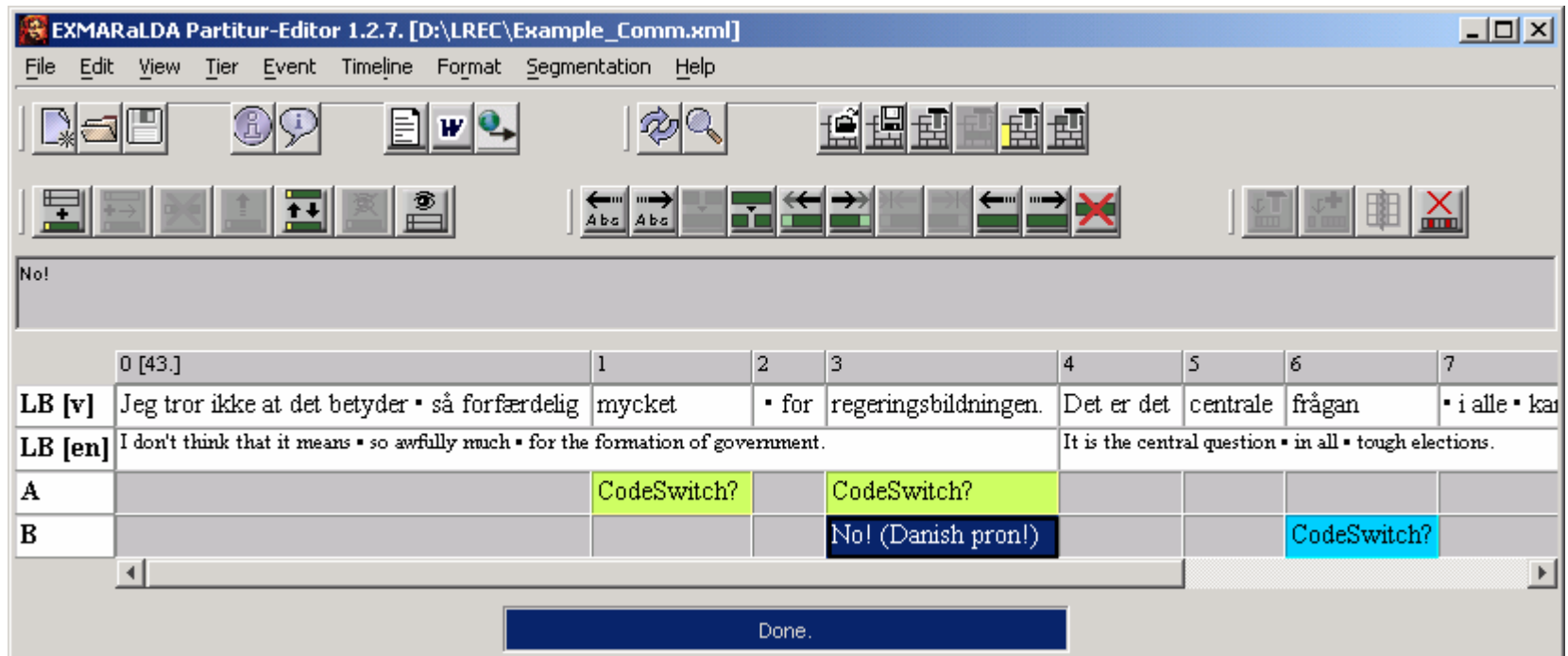
[2]

| | | | |
|-----------------|----------------------------|----------------------------------|------------------------|
| Sel | ((1,5s)) | Hep yazıyo. • Nasıl/ • böyle hep | gonuşuyo. |
| TL-Sel | | always write-PRS | How/ so always SPK-PRS |
| Sel [en] | It's always written there. | How/ • they always sort of like | speak. |
| Yıl | neler var? | | |
| TL-Yıl | what-PL there are | | |
| Yıl [en] | | | |

Collaborative transcription and annotation

1. Transcription
2. Transcription control ←
3. Utterance translation
4. Translation control ←
5. Morphological transliteration
6. Transliteration control ←

Collaborative Commentary II



The screenshot shows the EXMARaLDA Partitur-Editor 1.2.7 interface. The title bar indicates the file path is [D:\LREC\Example_Comm.xml]. The menu bar includes File, Edit, View, Tier, Event, Timeline, Format, Segmentation, and Help. The toolbar contains various icons for file operations, editing, and analysis. The main workspace displays a table with columns numbered 0 to 7. The table content is as follows:

| | 0 [43.] | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | | | | | | | | |
|---------|------------------------------|------------------|-------------|-------|----------------------|-------------------|----------|--------|----------|------------------------------------|--|----------------------------|--|----------|--|--------------------|--|
| LB [v] | Jeg tror ikke at det betyder | ▪ så forfærdelig | mycket | ▪ for | regeringsbildningen. | Det er det | centrale | frågan | ▪ i alle | ▪ kan | | | | | | | |
| LB [en] | I don't think that it means | | | | | ▪ so awfully much | | | | ▪ for the formation of government. | | It is the central question | | ▪ in all | | ▪ tough elections. | |
| A | | | CodeSwitch? | | CodeSwitch? | | | | | | | | | | | | |
| B | | | | | No! (Danish pron!) | | | | | CodeSwitch? | | | | | | | |

A blue button labeled "Done." is located at the bottom center of the interface.

Collaborative analysis

- Negotiate categorizations / interpretations

HTML example

Collaborative Commentary III

Jochem Rehbein, Thomas Schmidt, Bernd Meyer, Franziska Watzke, Annette Herkenrath

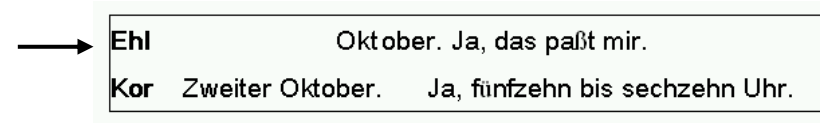
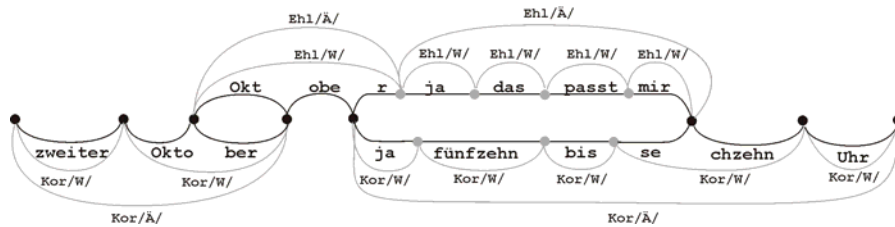
Handbuch für das computergestützte Transkribieren nach HIAT

Version 0.4, 07. April 2004

Collaborative publication

- Negotiate transcription conventions
- Get user feedback

Collaborative Commentary: Technology



Model
(Time-Based /
XML-files)

Visualisation(s)
(HTML
documents)

EXMARaLDA data model

ProjectPad data model

ProjectPad

Annozilla



Database „Multilingualism“ – Perspectives for collaborative corpus construction and collaborative commentary

Summary

- ❖ Research on multilingualism is a “market” for collaborative commentary:
 - collaborative transcription and annotation
 - collaborative analysis
 - collaborative publication
- ❖ A common framework for computerized transcription data
 - use different tools on (different flavors of) the same data structure
- ❖ Collaborative commentary can simply be the task of one of those tools
- ❖ Time based data models and tools like ProjectPad seem to go with one another