

## Database of multilingual spoken discourse

### Is:

- ≈2200 transcriptions of spoken language (30 min recording each)
- Language acquisition data, interviews, expert discourse, classroom discourse, presentation discourse, interpreted discourse,...
- 14 languages (German, English, Swedish, Norwegian, Danish, French, Spanish, Portuguese, Turkish, Italian, Basque, Japanese, Chinese, Russian)
- 9 different data formats / tools (dBase, syncWriter, HIAT-DOS, Verbmobil, ...)
- 3 different operating systems (MAC OS 9.x, Windows, Linux) + MAC OS X
- research interests: phonetics, syntax, discourse, ...

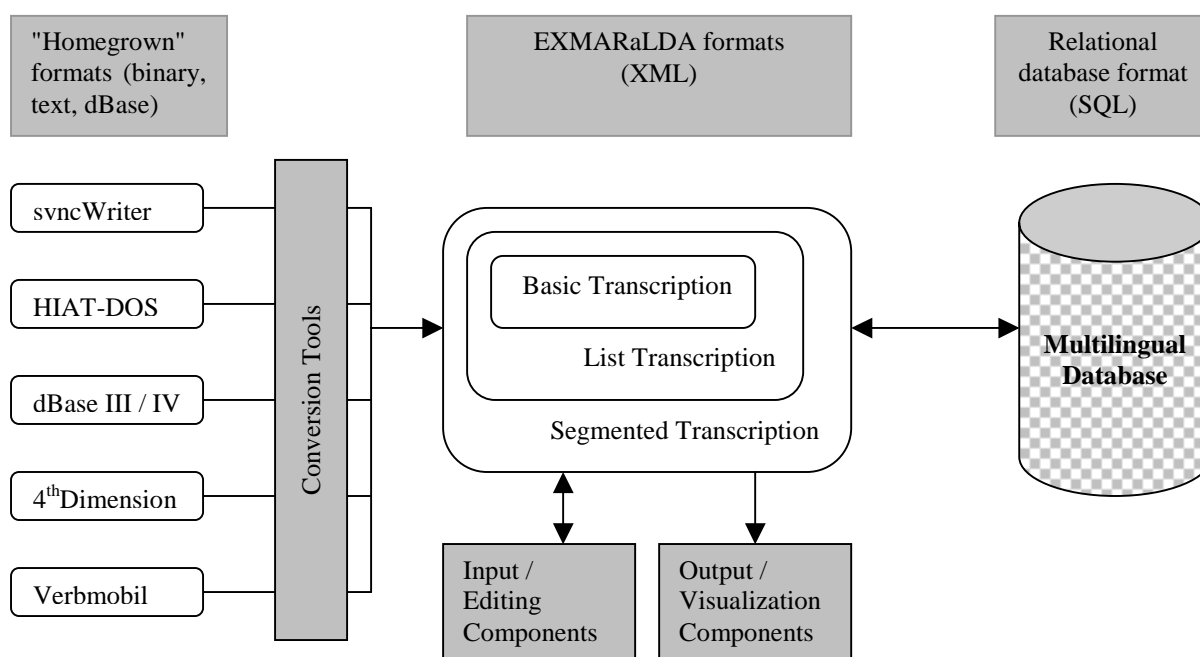


Figure 1: System architecture

### Should be:

- one common tool for accessing (querying) the data
  - Data must come in one format (AG)
  - Multilingual issues must be taken care of (UNICODE)
  - Data format should be software independent (XML)
  - Software should work across different OS (JAVA)
- different tools reflecting the habits and needs of the different projects
  - different input methods (Score, column, vertical notation)
  - different output methods (dito)

## "Traditional" layout principles for transcriptions of spoken language

MAX [v]:	You keep interrupting me, Tom.
MAX [nv]:	--- <i>pointing at Tom</i> -----
TOM [v]:	Oh, I'm sorry for that.
TOM [nv]:	--- <i>smiling</i> -----

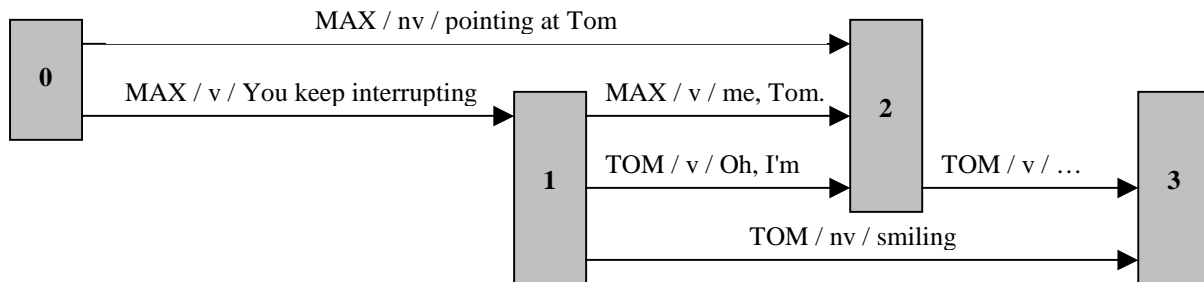


Figure 2: Score notation (*Partitur*) and underlying data structure

→ **Basic Transcription:** tiers, speakers, categories, timeline, events

```
<transcription>
  <speakertable>
    <speaker id="SPK1" name="MAX"/>
    <speaker id="SPK2" name="TOM"/>
  </speakertable>

  <timeline>
    <timepoint id="T0"/>
    <timepoint id="T1"/>
    <timepoint id="T2"/>
    <timepoint id="T3"/>
  </timeline>

  <tier speaker="SPK1" category="v">
    <event start="T0" end="T1">You keep interrupting </event>
    <event start="T1" end="T2">me, Tom. </event>
  </tier>
  <tier speaker="SPK1" category="nv">
    <event start="T0" end="T2">pointing at Tom</event>
  </tier>
  [...]
</transcription>
```

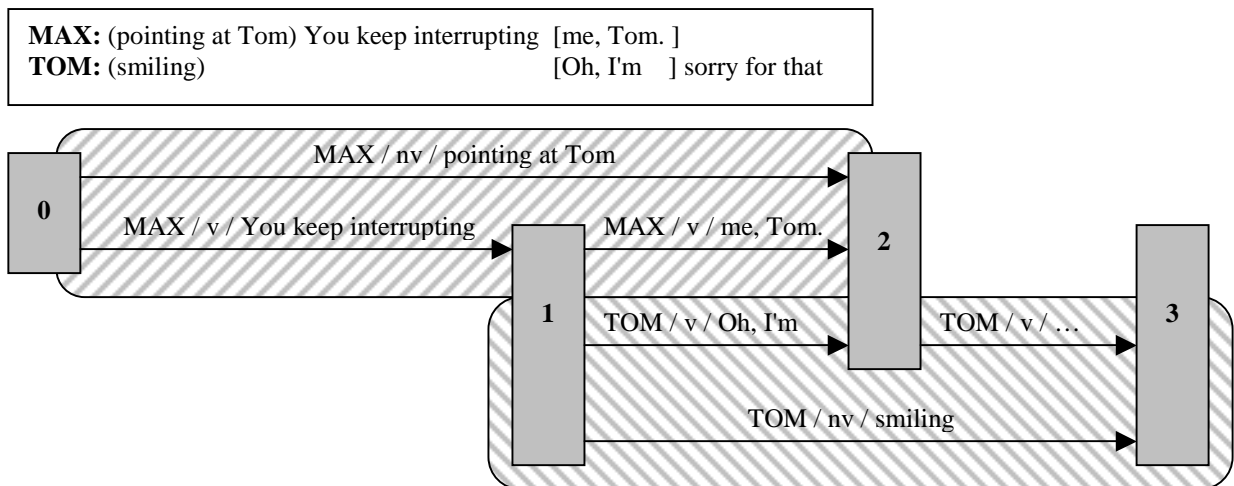


Figure 3: Vertical notation and underlying data structure

→ **List Transcription** = Basic Transcription (tiers, speakers, categories, timeline, events) + **speaker turns**

Traditional layout principles reflect book formats, but:

"[...] book format is an attribute not of speech, but of Western writing systems. There is no reason beyond established custom and practice to present speech in this way. On the contrary, since there are often several annotations relating to the same piece of data, book format is in many cases inappropriate. The use of book format without consideration of other possibilities is based on a confusion between the organization of the data itself, and the presentation of the data on the printed page." (Knowles 1995)

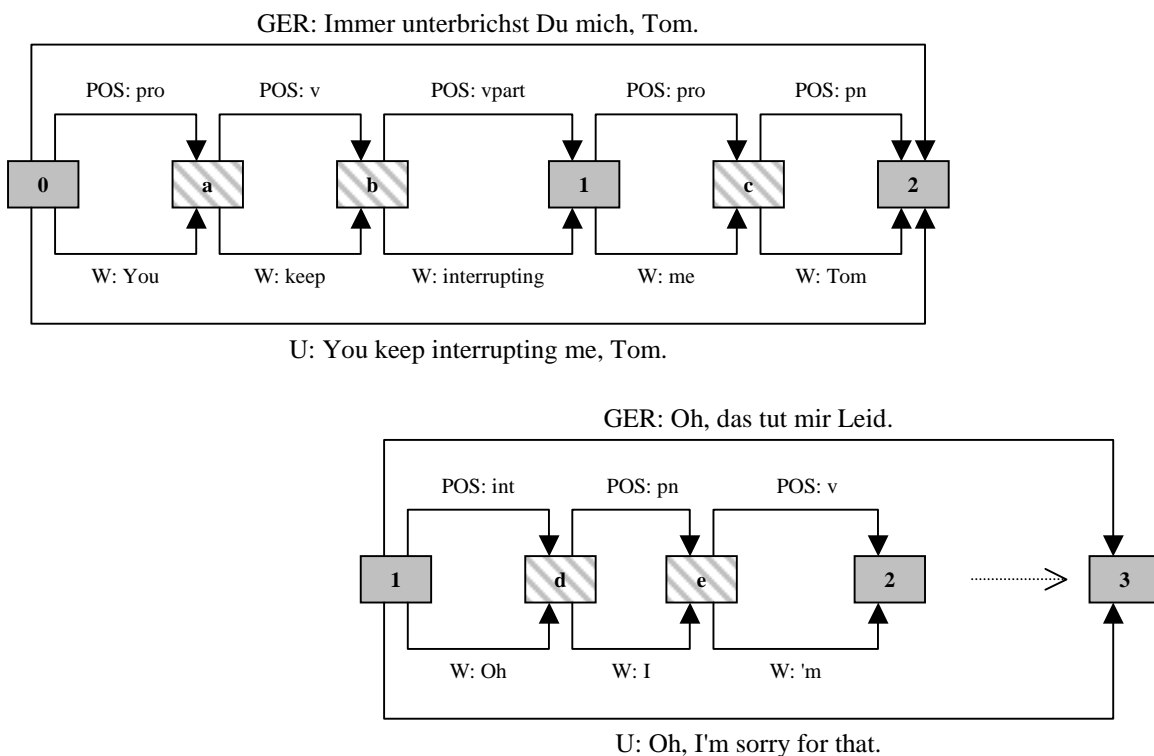


Figure 4: Structure of annotated data

→ **Segmented Transcription**: several, partially intersecting, partially independent timelines

## EXMARaLDA Tools

### Output / visualization tools:

- Output in vertical notation (HTML or RTF)
- Output in score notation (HTML or RTF)

MAX [v]	You keep interrupting me, Tom.
MAX [nv]	<i>pointing at Tom</i>
MAX [ger]	Immer unterbrichst Du mich, Tom.
TOM [v]	Oh, I'm sorry for that.
TOM [nv]	<i>smiling</i>
TOM [ger]	Oh, das tut mir Leid.

### Input / Editing tools:

- Input in vertical notation in a text file (e.g. MS Word): Simple Exmaralda

MAX:	[pointing at Tom]
	You keep interrupting <me, Tom.>1>
	{Immer unterbrichst Du mich, Tom.}
TOM:	[smiling]
	<Oh, I'm >1> sorry for that.
	{Oh, das tut mir Leid.}

- Input in score notation: Partitur-Editor

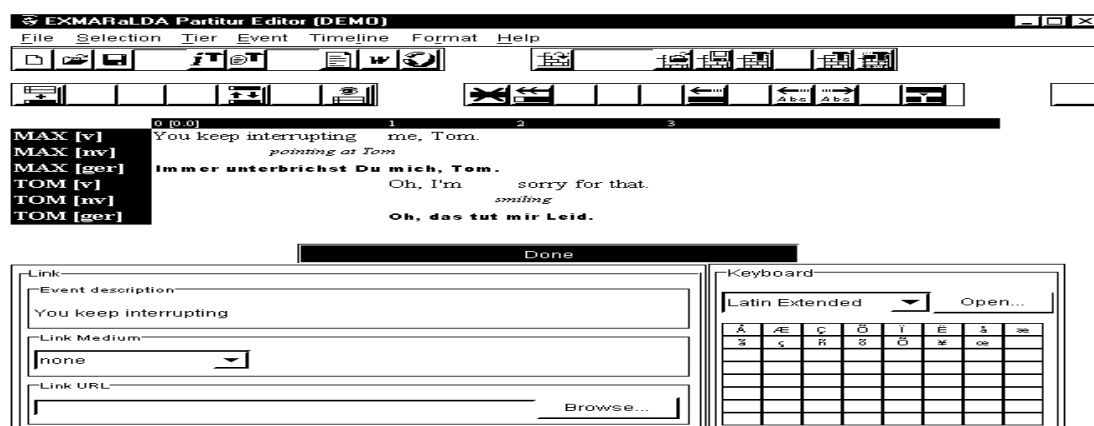


Figure 5: Screenshot of Score editor

### References:

Gerry Knowles 1995. *Converting a corpus into a relational database: SEC becomes MARSEC*. In: Geoffrey Leech et al. (ed.). *Spoken English on Computer: Transcription, Markup and Application*: 208—219. Harlow: Longman

Thomas Schmidt 2001. *Gesprächstranskription auf dem Computer - das System EXMARaLDA*. To appear in: *Gesprächsforschung* (2). [<http://www.gespraechsforschung-ozs.de>]