

EXMARaLDA

Thomas Schmidt

SFB 538 „Mehrsprachigkeit“

University of Hamburg

Data Formats and Tools at the SFB

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

Data Formats and Tools at the SFB

Ays				
TS-Ays				
Mer	tdi©imden sonra açıld/bu zaten.	• Bulduk...	H?	Pizza/
TS-Mer	<i>n bin, wurde es erst eröffnet.</i>	<i>Wir haben es gefunden...</i>	<i>Hm?</i>	<i>Pizza/ Pizza</i>
Mut		(Pizza) Napoli mi var kar/shda?	Pizza Napoli mi var kar/shda?	
TS-Mut		<i>Befindet sich (Pizza) Napoli gegenüber.</i>	<i>Befindet sich Pizza Napoli gegenüber?</i>	
Sev				
TS-Sev				

syncWriter:

- editor for interlinear text
- MAC OS 9.x and earlier
- outputs binary data

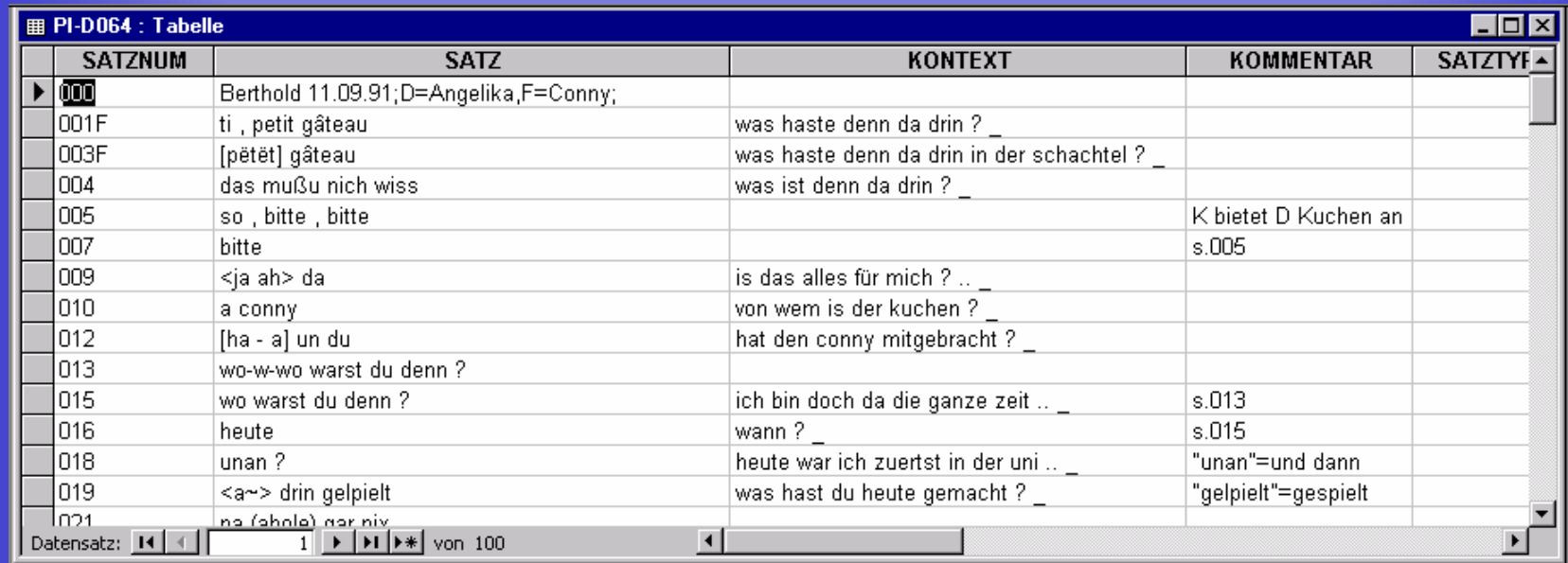
Data Formats and Tools at the SFB

```
HIAT22.EXE
F1: 33  Pos: 1  Transkriptname: A6T003  HIAT-DOS 2.2
Bezugszeile < [
Kontrollzeile = [
PL
N1
N2
TC
N1
N2
NE
N1
N2
N1
N2
N1
N2
[
etinget . for första g*ngen.
Enhetslistan <
Og Enhedslisten..
]
```

HIAT-DOS:

- editor for HIAT-transcription
- MS-DOS/Windows
- outputs text files

Data Formats and Tools at the SFB



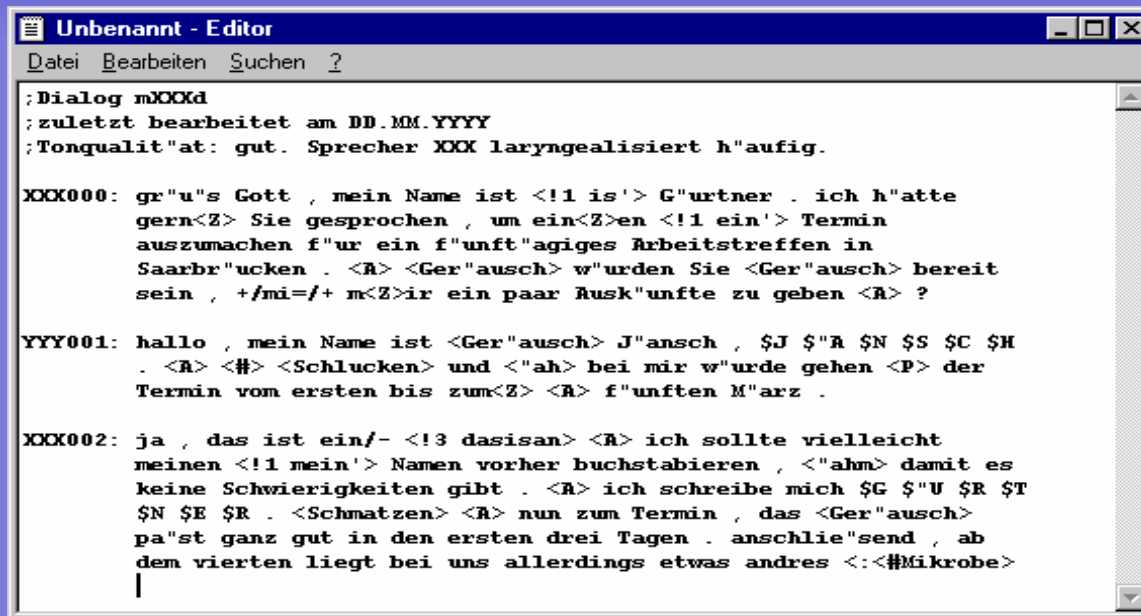
SATZNUM	SATZ	KONTEXT	KOMMENTAR	SATZTYF
000	Berthold 11.09.91;D=Angelika,F=Conny;			
001F	ti , petit gâteau	was haste denn da drin ? _		
003F	[pëtët] gâteau	was haste denn da drin in der schachtel ? _		
004	das mußu nich wiss	was ist denn da drin ? _		
005	so , bitte , bitte		K bietet D Kuchen an	
007	bitte		s.005	
009	<ja ah> da	is das alles für mich ? .. _		
010	a conny	von wem is der kuchen ? _		
012	[ha - a] un du	hat den conny mitgebracht ? _		
013	wo-w-wo warst du denn ?			
015	wo warst du denn ?	ich bin doch da die ganze zeit .. _	s.013	
016	heute	wann ? _	s.015	
018	unan ?	heute war ich zuerst in der uni .. _	"unan"=und dann	
019	<a~> drin gelpielt	was hast du heute gemacht ? _	"gelpielt"=gespielt	
021	na (ahole) gar nix			

Datensatz: 1 von 100

dBase/Access/4th Dimension

- utterance databases

Data Formats and Tools at the SFB



```
Unbenannt - Editor
Datei Bearbeiten Suchen ?

;Dialog mXXXd
; zuletzt bearbeitet am DD.MM.YYYY
;Tonqualitaat: gut. Sprecher XXX laryngealisiert h"aufig.

XXX000: gr"u"s Gott , mein Name ist <!1 is'> G"urtner . ich h"atte
gern<Z> Sie gesprochen , um ein<Z>en <!1 ein'> Termin
auszumachen f"ur ein f"unft"agiges Arbeitstreffen in
Saarbr"ucken . <R> <Ger"ausch> w"urden Sie <Ger"ausch> bereit
sein , +/mi=/+ m<Z>ir ein paar Rusk"unfte zu geben <R> ?

YYY001: hallo , mein Name ist <Ger"ausch> J"ansch , $J $"A $N $$ $C $H
. <R> <#> <Schlucken> und <"ah> bei mir w"urde gehen <P> der
Termin vom ersten bis zum<Z> <R> f"unften M"arz .

XXX002: ja , das ist ein/- <!3 dasisan> <R> ich sollte vielleicht
meinen <!1 mein'> Namen vorher buchstabieren , <"ahm> damit es
keine Schwierigkeiten gibt . <R> ich schreibe mich $G $"U $R $T
$N $E $R . <Schmatzen> <R> nun zum Termin , das <Ger"ausch>
pa"st ganz gut in den ersten drei Tagen . anschlie"send , ab
dem vierten liegt bei uns allerdings etwas andres <:<#Mikrobe>
|
```

Verbmobil:

- 7-bit ASCII files

Database „Multilingualism“

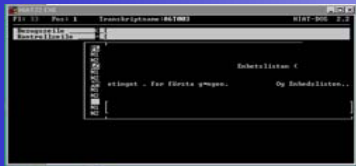
Goals:

1. To have 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)
2. To have different tools reflecting the habits and needs of the different projects
 - different input methods (Score, column, vertical notation)
 - different output methods (dito)

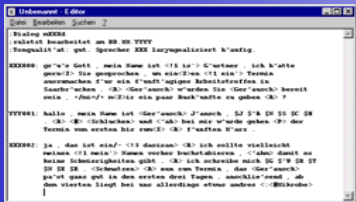
Database „Multilingualism“



SyncWriter



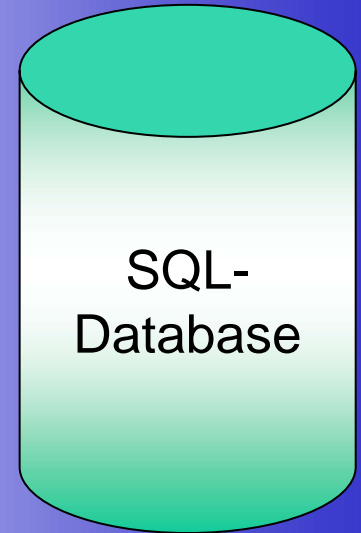
HIAT-DOS



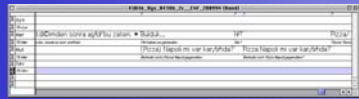
Verbmobil



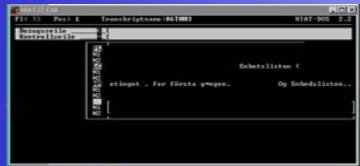
ACCESS /
dBase



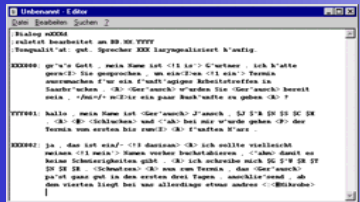
Database „Multilingualism“



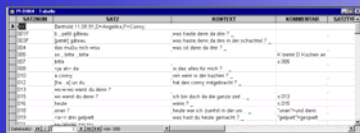
SyncWriter



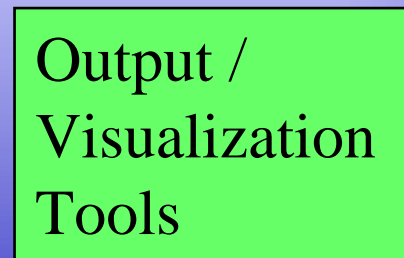
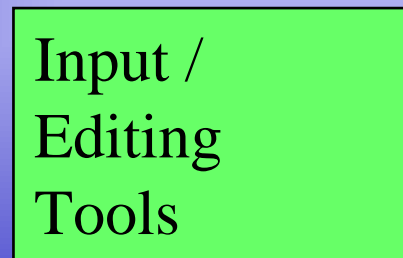
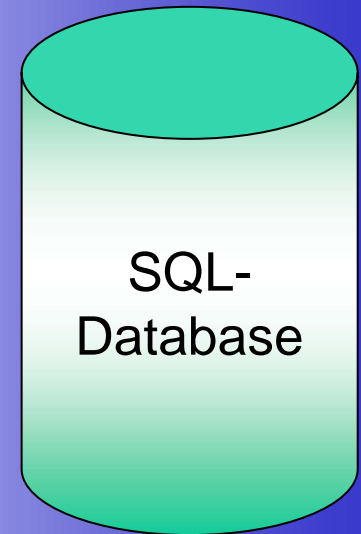
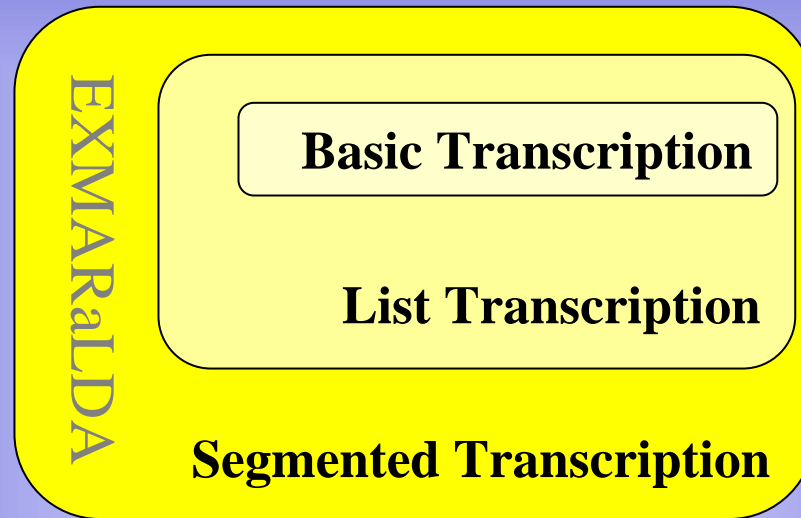
HIAT-DOS



Verbmobil



ACCESS /
dBase



„Traditional“ layout principles

1. Score notation („Partitur“)

MAX [v] You keep interrupting me, Tom.

MAX [nv] ----- *pointing at Tom* -----

TOM [v] Oh, I'm sorry for that.

TOM [nv] ----- *smiling* -----

„Traditional“ layout principles

1. Score notation („Partitur“)

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> -----

Tiers

„Traditional“ layout principles

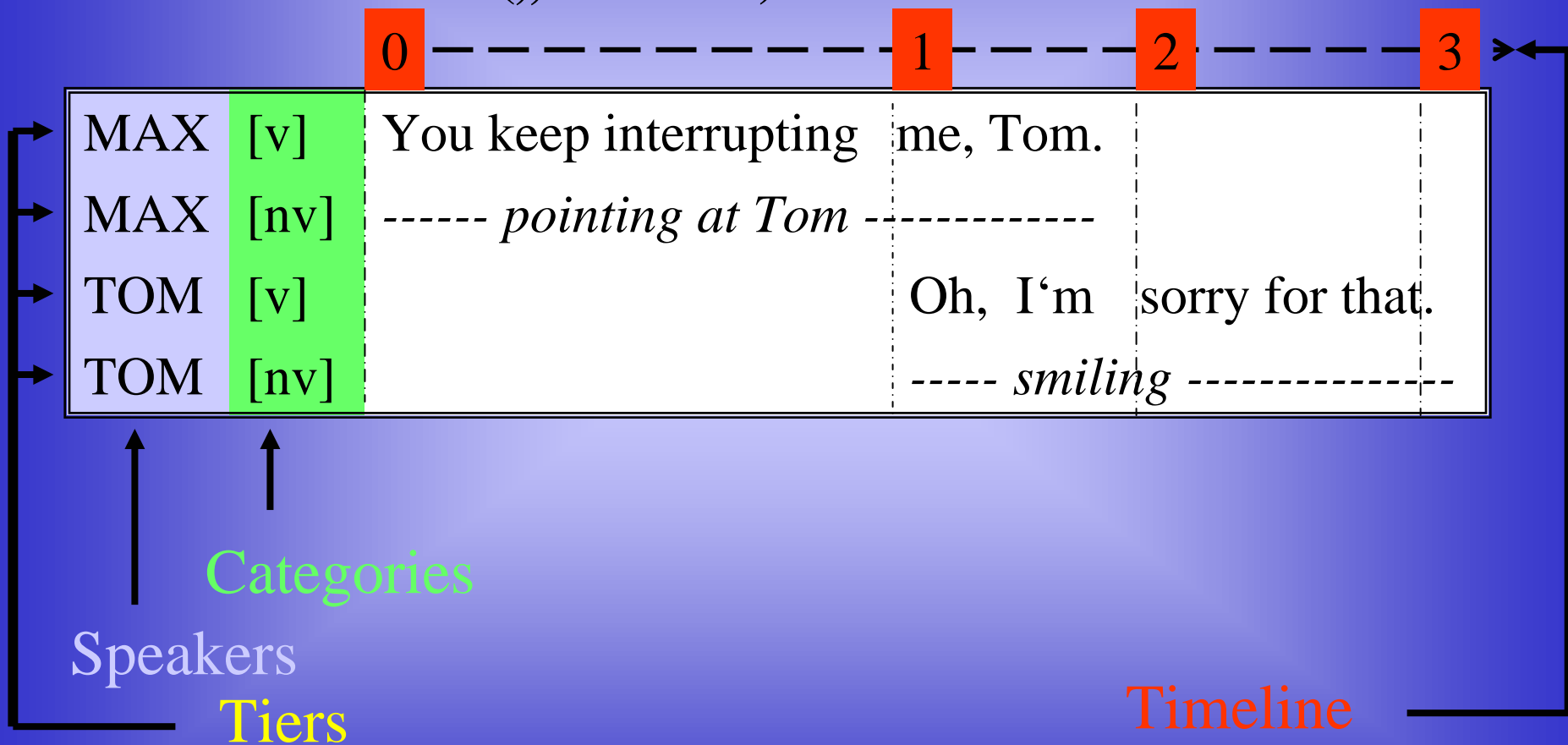
1. Score notation („Partitur“)

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> -----

↑
↑
Categories
Speakers
Tiers

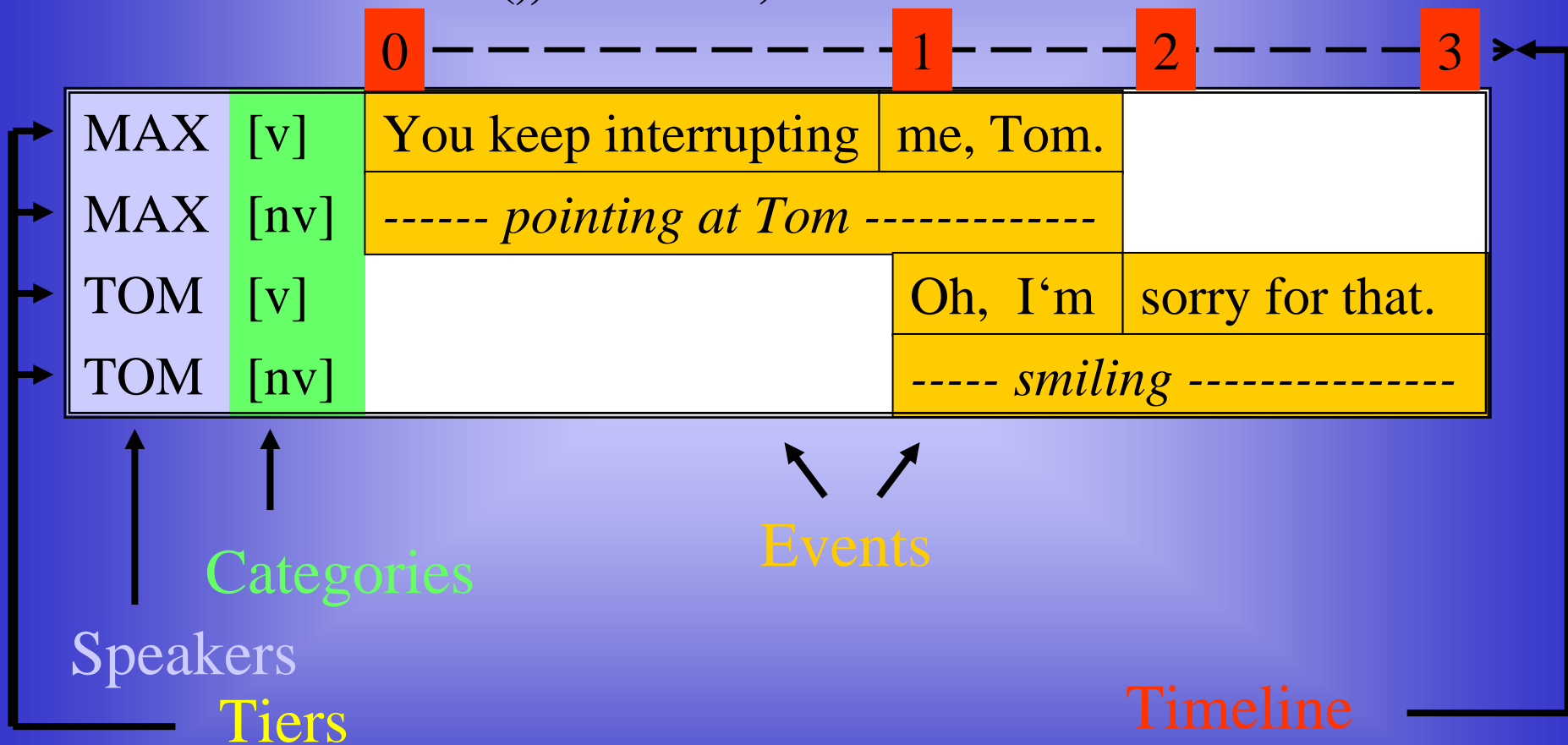
„Traditional“ layout principles

1. Score notation („Partitur“)



„Traditional“ layout principles

1. Score notation („Partitur“)



„Traditional“ layout principles

1. Score notation („Partitur“) → Basic Transcription

```
<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>
```

Categories

Speakers

Events

Timeline

Tiers

„Traditional“ layout principles

2. Column notation

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

„Traditional“ layout principles

2. Column notation → Basic Transcription

	MAX [v]	MAX [nv]	TOM [v]	TOM [nv]
0	You keep interrupting	<i>pointing at Tom</i>		
1	me, Tom.		Oh, I'm	<i>smiling</i>
2		sorry for that.		
3				

Categories
 Speakers
 Events
 Timeline
 Tiers

„Traditional“ layout principles

3. Vertical notation

MAX (*pointing at Tom*)

You keep interrupting [me, Tom.]

TOM (*smiling*)

[Oh, I'm] sorry for that.

„Traditional“ layout principles

3. Vertical notation

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

Categories

Speakers

Events

Timeline

Tiers

„Traditional“ layout principles

3. Vertical notation

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

Speaker-Turns

Categories Speakers Events Timeline Tiers

Structure Of Annotated Data

You keep interrupting me, Tom.

Oh, I`m sorry for that

Events (temporal structure)

Structure Of Annotated Data

You keep interrupting me, Tom.

Immer unterbrichst Du mich, Tom

Oh, I `m sorry for that

Oh, das tut mir Leid.

Events (temporal structure)

Utterances (linguistic structure)

Structure Of Annotated Data

You	keep	interrupting	me,	Tom.
Immer unterbrichst Du mich, Tom				
Pro	V	Vpart	Pro	PN.

Oh,	I`m	sorry	for	that
Oh, das tut mir Leid.				
Int	P	V	Adj	Prep Pro

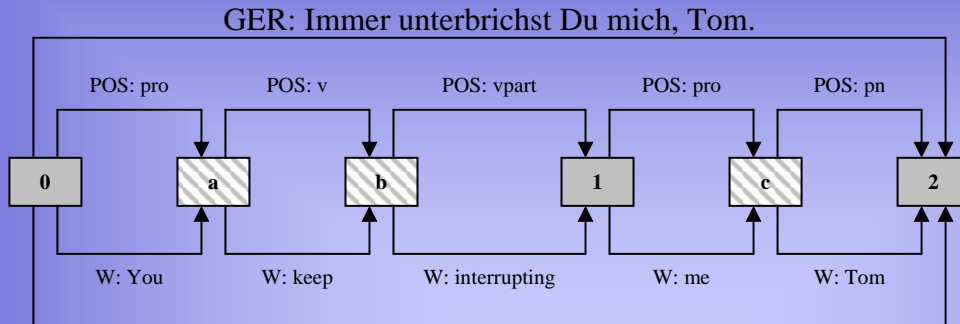
Events (temporal structure)

Utterances (linguistic structure)

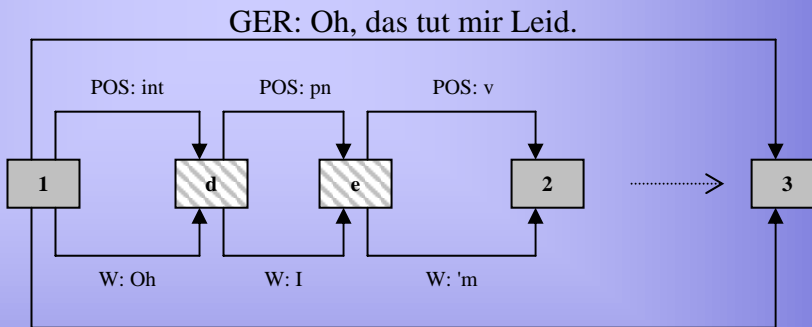
Words (linguistic structure)

.....

Structure Of Annotated Data



U: You keep interrupting me, Tom.



U: Oh, I'm sorry for that.