

EXMARaLDA

Creating, managing, analysing
and sharing spoken language
corpora



- Background
- Aims and principles
- System overview
- Demonstrations
 - Transcription (Partitur-Editor)
 - Analysis (EXAKT)
 - Publication

Background



- Research Centre on Multilingualism, University of Hamburg
- Since 1999, until June 2011
- More than 20 projects in 4 groups
 - Multilingual **individuals**
 - Multilingual **communication**
 - **Historical** multilingualism
 - Transfer group (Practical **applications**)

Background



- All work empirical, based on corpus data
- Mostly oral data
 - Language **acquisition** corpora (bilingual children: French, Spanish, Turkish, Portuguese/German Spanish/Basque)
 - Language **attrition** corpora (bilingual adults: Polish, Italian/German)
 - **SLI** data (German / Turkish)
 - **Interpreting** data (Portuguese, Turkish / German)
 - Interscandinavian **communication**
- also: written data (e.g. historical texts, translation corpora)

Example: French/German bilinguals



UH DUFDE - AN-F069 [Prev] [Next] 

ANNIKA	Interviewer	
00002 de kaka		
	ha ha	
00003 ca pas		An choisit l'un des jeux
	tu veux pas jouer à celui-là?	
00004 non ca c'est un		An mélange les cartes
	tu mélanges bien	
00006 comme ca		An montre à P comment ranger les cartes
	on fait comme ca toutes	
00007 maintenant comme ca		An mélange les cartes
	et moi aussi	
00008 oh tu (x)		
00009 /tu/=toutes les cartes		
	toutes les cartes	
00012 les deux de ca		
00013 de ca		An montre les cartes sur la boîte
	il faut retrouver les deux cartes	
00015 maintenant ici		An se met sur le ventre et continue à ranger les cartes
	attends je finis ici d'abord	
00016 ah maintenant (xx) là		
	et ensuite quel côté	

Example: interpreted doctor/patient communication



[2]

..	6	7	8	9	10	11	12	13	14
Intonation									
A [v]	Krankenhaus kamen.								
D [v]	((1,5s))Eh ela tá a perguntar ((holt hörbar Luft)) eehm... • Como é que eu digo? ••• ((schnalzt))								
D [de]	<i>Äh sie fragt gerade</i>			<i>äähm...</i>		<i>Wie sage ich es?</i>			
P [v]	Ja.								
P [de]	((1s))Quee/ por • que, por que <i>Dass/ warum, warum ich ins</i>								

[3]

..	15	16	17	18	19	20	21	22	23		
D [v]	Certo. E que problemas é que você tinha.					Hm̃'			Sim'		
D [de]	<i>Genau. Und was für Probleme Sie hatten.</i>								<i>Ja'</i>		
P [v]	vine al hospital.					Sí. ((holt hörbar Luft)) •Hace o cho dias, ((1,5s))yo me fui al control • normal					
P [de]	<i>Krankenhaus gekommen bin.</i>					<i>Ja.</i>		<i>Vor acht Tagen</i>		<i>bin ich zur normalen Kontrolle</i>	<i>((unverständlich,</i>

[4]

..	24	25	26	27	28	29	30	31	32	33	34	35
D [v]	Jä'		Oito dias, sim.					Ao ECG, hm̃.				
D [de]			<i>Acht Tage, ja.</i>					<i>Am EKG, hm.</i>				
P [v]	((unverständlich, 0,5s))la doctora mia, ¿no?		Hace hoy, hoy ocho días.			Sí ((1s))Y ella me tomó loo que... ¿Como llama?			Ja,			
P [de]	<i>0,5s)) meiner Ärztin gegangen, ne?</i>		<i>Heute, heute sind es acht Tage.</i>			<i>Ja'</i>		<i>Und sie hat mir genommen das... Wie heißt das?</i>		<i>Ja,</i>		
D [k]			zögerlich									

Background



- Our project: „Computer-assisted methods for creating and analysing multilingual data“
- Software development → **EXMARaLDA**
- Corpus curation and dissemination → **Data sharing**
- Methodological questions → **Workflows, quality control, transcription conventions**
- Sustainability of data (long-term archiving etc.) → **standardisation, infrastructures**

Aims and principles



- Develop a system which...
 - makes intelligent use of **multimedia** and **hypertext** capabilities of the computer
 - is usable for **different** theoretical **approaches**, different **research interests**
 - is usable in **different technological environments**
 - produces data which is easily **shared**
 - can ingest existing „**legacy data**“
 - produces data which has a chance of surviving the next 30 years

Aims and principles



- Data formats based on open standards (**XML, Unicode**)
- Theory-neutral data model (based on **annotation graphs**)
- Cross-Platform software tools (**JAVA**)
- Interoperability with other widely used tools (**Praat, ELAN, CLAN, Transcriber, ...**)
- Interoperability with common desktop formats (**HTML, RTF, PDF**)
- Compatibility with emerging/possible standards (e.g. **TEI**)

System overview: EXMARaLDA



- Data model, data formats
 - for **transcriptions** (+ annotations)
 - for **corpora** (+ metadata)
- Tools
 - **Partitur-Editor**: Transcription
 - **Corpus Manager**: Corpus Administration, Metadata
 - **EXAKT**: Query, Analysis
 - Utilities for data conversion, corpus publication etc.

System overview: Partitur-Editor



	17 [42.3]	18 [44.3]	19 [45]	20 [47.3]	21 [49.0]	22 [51.0]	23 [51.3]	24 [53.5]
PRE [v]				I mean your/ your new theory.		Yeah.		
ELK [v]		Ehem.		... Oh what is my theory?		Ooh what is my theory that it is!	Well, C	
[m]	((laughter))		((laughter))					
PRE [k]								
ELK [k]								

Audio/Video panel [JMF]

0.0 00:00.0 00:22.2 03:36.4 03:36.4 sec

Loop Playback halted

IPA Panel

Vowels / Suprasegmentals Consonants (pulmonic) Diacritics

Front Near front Central Near back Back

Close i y i u u

Near close I • Y • U

Close mid c ɔ o o ɤ o

Mid e e a a A ɔ

Near open æ e e

Open a e a A D

SUPRASEGMENTALS

- Primary stress
- Secondary stress
- Long
- Short
- Syllable break
- Extra stress
- Half-long
- Extra-short
- Linking

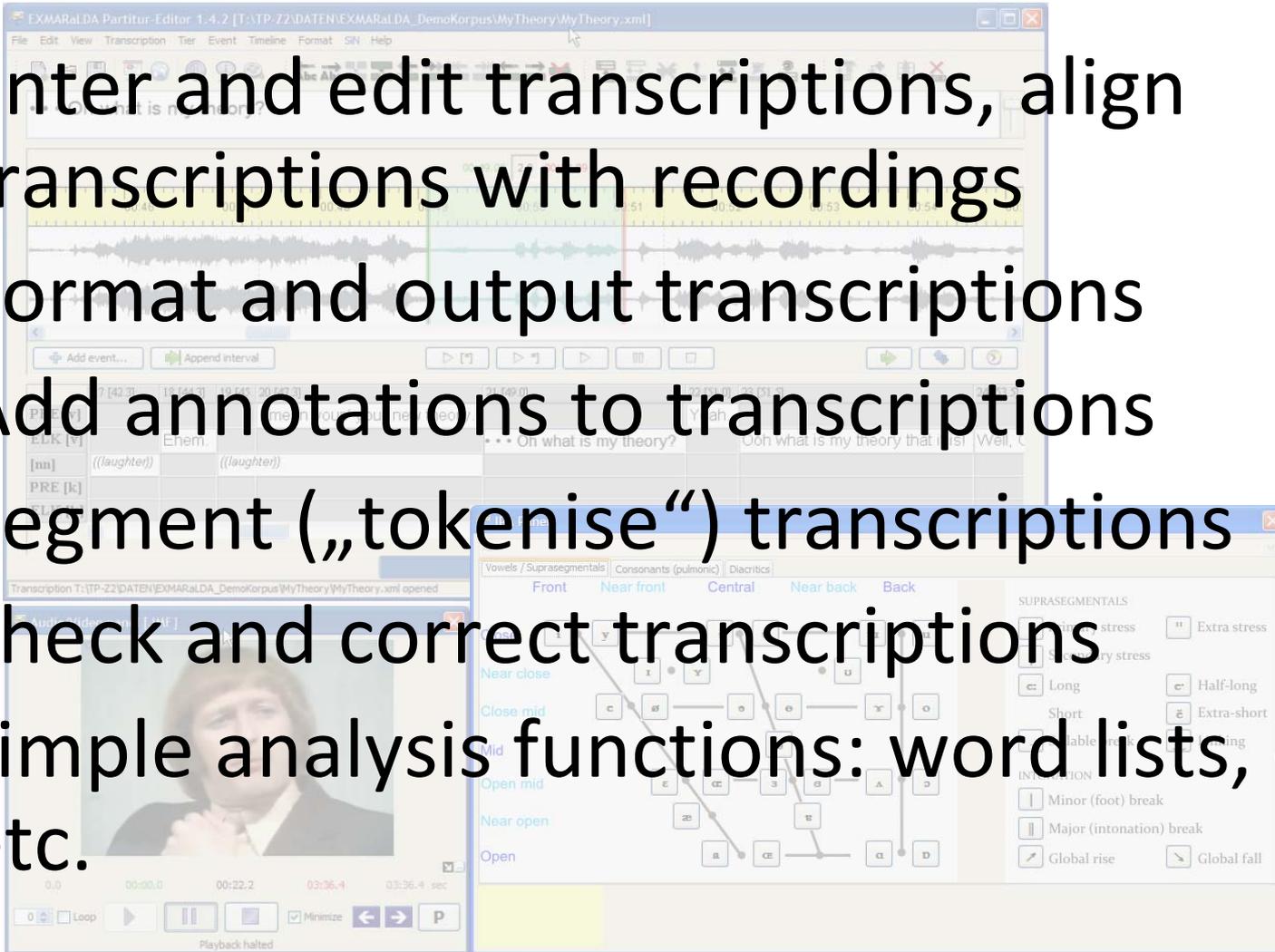
INTONATION

- Minor (foot) break
- Major (intonation) break
- Global rise
- Global fall

System overview: Partitur-Editor



- Enter and edit transcriptions, align transcriptions with recordings
- Format and output transcriptions
- Add annotations to transcriptions
- Segment („tokenise“) transcriptions
- check and correct transcriptions
- simple analysis functions: word lists, query, etc.



System overview: Coma



EXMARaLDA Coma | MapTask_Test.coma | Corpus: MapTask_Test

File View Analysis Maintenance Help

Corpus Data Basket (0) Settings

Filter remove all filters 20 Communications

Communication

S	Name	Aufnehmende
	MT_270110_Shirin	Kim Chi Hamze
	MT_091209_Dimitri	Kim Chi Hamze
	MT_280410_Hamit	Kim Chi Hamze
	MT_051109_Nadira	Secil Yusun
	MT_280110_Lucy	Kim Chi Hamze
	MT_031109_Liang	Secil Yusun
	MT_051109_Matteo	Secil Yusun
	MT_280110_Hoa	Kim Chi Hamze
	MT_281009_Phuong	Secil Yusun
	MT_091209_David	Martina Schwalm
	MT_280410_Janis	Kim Chi Hamze
	MT_280110_Minh	Kim Chi Hamze
	MT_281009_Tansu	Secil Yusun
	MT_270110_Zhi_Zhi	Kim Chi Hamze
	MT_281009_Victor	Secil Yusun
	MT_280110_Hussein	Kim Chi Hamze
	MT_091209_Rufus	Martina Schwalm
	MT_031109_Hayat	Secil Yusun
	MT_281009_Stella	Secil Yusun
	MT_091209_Ali	Kim Chi Hamze

Search

Communication **MT_280410_Hamit**

Description (Communication)

Aufnahmedatum 280410

Aufnahmegerät M-Audio - Microtrack II

Aufnehmende Kim Chi Hamze

comment Die beiden Map-Task-Teilnehmer kannten sich vorher.

project-name Maptask

transcription-convention HIAT

transcription-name MT_280410_Hamit

No Location +

Language(s) +

Language LanguageCode deu (German)

Description (Language)

1 Recording +

Recording: MT_280410_Hamit

Description (Recording)

Filter remove all filters 24 Speaker

S	Sigle	Muttersprache
	Hus	Arabisch
	Tan	?
	Raf	Portugiesisch, Deutsch
	Vic	Französisch
	Sh	Türkisch
	Dim	Russisch, Thai
	Phu	Koreanisch
	AufS	?
	Dav	Deutsch, Französisch
	Minh	Vietnamesisch
	Nad	Afghanisch
	Hay	Türkisch
	Auf	?
	AufK	?
	Af	?
	Hoa	Vietnamesisch
	Ham	Arabisch
	Li	Chinesisch
	Jan	Griechisch
	Ali	Dari
	Lucy	Bulgarisch
	St	?
	Mat	Spanisch
	Zhi	Chinesisch

System overview: Coma



- **Corpus-Manager**
- Bundle transcriptions and recordings into a corpus
- Enter and organise metadata for communications, speakers, recordings
- Query metadata, build subcorpora
- Corpus maintenance (consistency checks etc.)
- Corpus analysis (quantification etc.)

S	Name	Aufnehmende
MT_270110	Shirin	Kim Chi Hamze
MT_051109	Nadira	Secil Yusun
MT_280110	Lucy	Kim Chi Hamze
MT_031109	Matteo	Secil Yusun
MT_280110	Hoa	Kim Chi Hamze
MT_281009	Phuong	Secil Yusun
MT_091209	David	Martina Schwaim
MT_280110	Minh	Kim Chi Hamze
MT_281009	Tansu	Secil Yusun
MT_270110	Zhi_Zhi	Kim Chi Hamze
MT_091209	Lucy	Martina Schwaim
MT_031109	Hayat	Secil Yusun
MT_281009	Stella	Secil Yusun

S	Sigle	Muttersprache
Hus		Arabisch
Tan		?
Sh		Türkisch
Dim		Russisch, Thai
Phu		Koreanisch
AufS		?
Dav		Deutsch, Französisch
Minh		Vietnamesisch
Nad		Afghanisch
Hay		Türkisch
Auf		?
AufK		?
Af		?
Hoa		Vietnamesisch
Li		Chinesisch
Jan		Griechisch
Ali		Dari
Lucy		Bulgarisch
St		?
Mat		Spanisch
Zhi		Chinesisch

System overview: EXAKT



EXMARaLDA EXAKT 0.8

File Edit View Concordance Help

Corpora

- MAPTASK**
 - S:\TP-Z...nahmen\MAPTASK.coma
 - 21 transcriptions
 - 2362 segment chains

Done.

Concordances

wie, wer, warum, was

MAPTASK

95 tokens

4 types

Word lists

Word list for MAPTASK

1210 types 16862 tokens

MAPTASK (95 results)

RegEx (T) Search: \b[Ww](er|ie|as|arum)\b

#	S	Communication	Speaker	Left Context	Match	Right Context	Age[S]
1	<input checked="" type="checkbox"/>	MT_270110_Shirin	Zhi	2s)) darf ich an • der Zahnbürste vorbei	wie		19
2	<input checked="" type="checkbox"/>	MT_091209_Dimitri	Dim	hnbürste also vo/ ((2s)) also im Prinzip	wie	/ bist du dann links bei der Zahnbürste i	24
3	<input checked="" type="checkbox"/>	MT_280410_Hamit	Ham	Mann • der mit den Salami oder Schlange	wer	was weiß ich	31
4	<input checked="" type="checkbox"/>	MT_280410_Hamit	Ham	n • der mit den Salami oder Schlange wer	was	weiß ich	31
5	<input checked="" type="checkbox"/>	MT_280410_Hamit	Ham	von diese Box da oben oder dieser Karton	was	weiß ich •• dann gehst du gleich rechts	31
6	<input checked="" type="checkbox"/>	MT_280410_Hamit	Jan	((lacht, 0,5s)) und jetzt	was	ist los was mach ich jetzt	21
7	<input checked="" type="checkbox"/>	MT_280410_Hamit	Jan	((lacht, 0,5s)) und jetzt was ist los	was	mach ich jetzt	21
8	<input checked="" type="checkbox"/>	MT_051109_Nadira	Nad	zu diesem Weizen diesen einen/ ((1,5s))	wie	ein Strauß aussieht also	21
9	<input checked="" type="checkbox"/>	MT_051109_Nadira	Nad	o zwei sch/ ((2,9s)) ähm Sträube oder so	wie	((unv.))	21
10	<input checked="" type="checkbox"/>	MT_051109_Nadira	Mat	((0,3s)) ja und von dort	was	mache ich	25
11	<input checked="" type="checkbox"/>	MT_051109_Nadira	Mat	mhm ((1s))	wie	(es)	25
12	<input checked="" type="checkbox"/>	MT_051109_Nadira	Mat	((1,1s))	wie	((0,7s)) welcher Strauß	25
13	<input checked="" type="checkbox"/>	MT_280110_Lucy	Lucy	((1,1s)) äh	was	siehst du dann links so •• schräg Motor	37
14	<input checked="" type="checkbox"/>	MT_280110_Lucy	Lucy	• dein Strich ist rechts ((lacht)) also	was	du •• meinst •• also du l/•• läufst	37
15	<input checked="" type="checkbox"/>	MT_280110_Lucy	Lucy	••• bis •••	was	weiß ich was ist das - die/ der •• bis	37
16	<input checked="" type="checkbox"/>	MT_280110_Lucy	Lucy	••• bis ••• - was weiß ich	was	ist das - die/ der •• bis ••• zur Lab	37
17	<input checked="" type="checkbox"/>	MT_280110_Lucy	Lucy	liei nach oben ((lacht)) (so zu) hast du	was	verstanden	37

weiter nach oben bis fast ((1s)) zum Box •• auf die Linie von diese Box da oben oder dieser Karton
was weiß ich •• dann gehst du gleich rechts

Age[S]	31
Mother tongue[S]	Arabic
Are the participants acquainted?[C]	Yes

Types: 4
 Tokens: 95
 Selected: 95 (1)
 Time: 2.5 s

Partitur

Ham [v] n bis fast ((1s)) zum Box •• auf die Linie von diese Box da oben oder dieser Karton **was weiß** ich •• dann gehst du gleich rechts ((1,8s)) dann wieder nach unten ((2,6s)) dann wi
 Jan [v] ja ja •• ja
 AfP2 [v]
 [nn]

Partitur List HTML

System overview: EXAKT

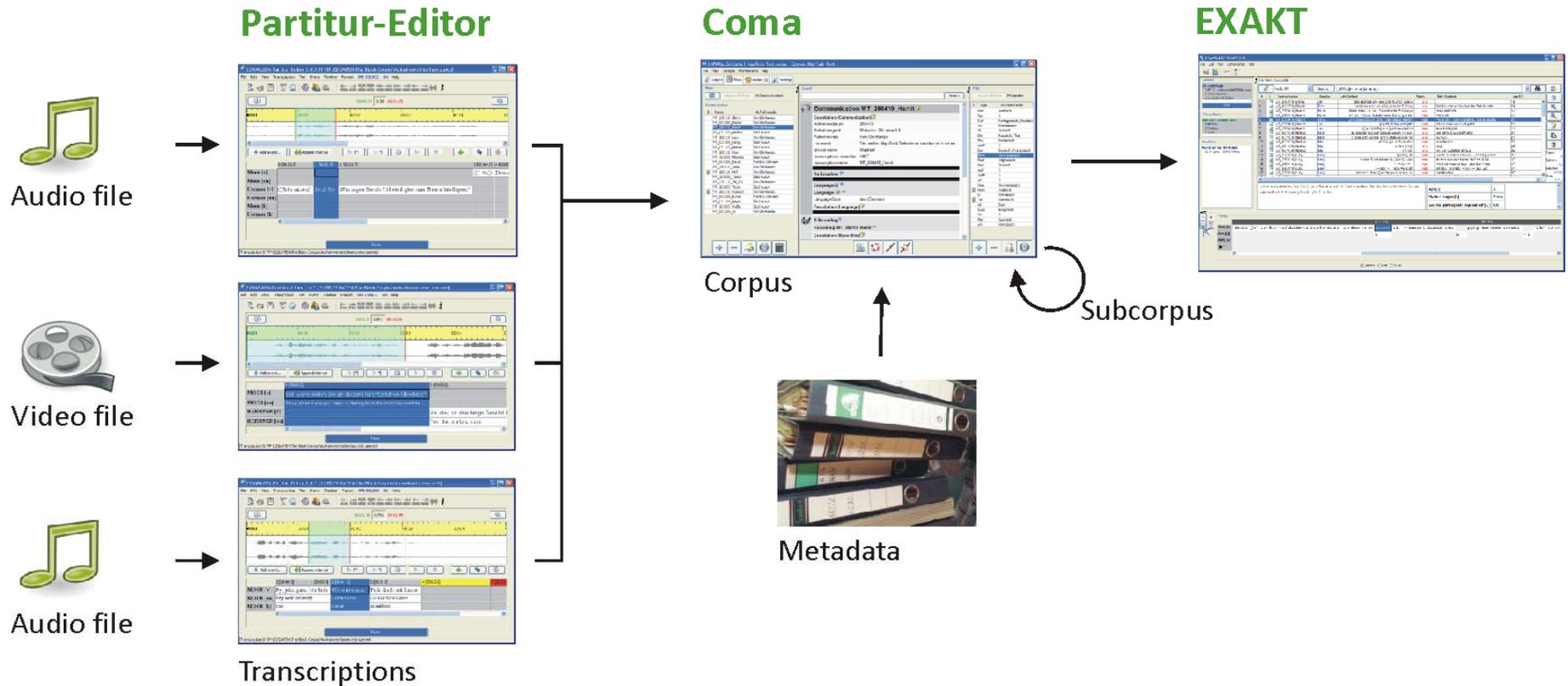


- **EXMARaLDA Analysis and Concordance tool**
- Query corpora
- KWIC-Concordancing
- Word list functionality
- Correlate query results and metadata
- Filter, categorise, quantify query results

The screenshot displays the EXMARaLDA EXAKT 0.8 software interface. The main window shows a concordance table with columns for #, S, Communication, Speaker, Left Context, Match, Right Context, and Age[S]. The search criteria are set to RegEx (T) with the pattern \b[Ww](er|ie|as|arum)\b. The table lists several concordances, with the fifth one highlighted. A search results window is open at the bottom, showing the selected concordance and its context.

#	S	Communication	Speaker	Left Context	Match	Right Context	Age[S]
2				2s)) darf ich an • der Zahnbürste vorbei	wie	/ bist du dann links bei der Zahnbürste	19
3				hnbürste also vo/ ((2s)) also im Prinzip	wie	was weiß ich	24
4				Mann • der mit den Salami oder Schlange	wer	weiß ich	31
4				n • der mit den Salami oder Schlange	was	weiß ich	31
5				von diese Box da oben oder dieser Karton	was	weiß ich •• dann gehst du gleich rechts	31
6				((lacht, 0,5s)) und jetzt	was	ist los was mach ich jetzt	21
6				((lacht, 0,5s)) und jetzt was ist los	was	mach ich jetzt	21
8				zu liegen W in diesen einen/ ((1,5s))	wie	ein Strauß aussieht also	21
8				ahm Strauße oder so	wie	((unv.))	21
10				((0,3s)) ja und von dort	was	mache ich	25
11				mhm ((1s))	wie	(es)	25
12				((1,1s))	wie	((0,7s)) welcher Strauß	25
17				((1,1s)) ah	was	siehst du dann links so •• schräg Motor	37
17				nts ((lacht)) also	was	du •• meinst •• also du l/ •• läufst	37
17				••• bis •••	was	weiß ich was ist das - die/ der •• bis	37
17				••• - was weiß ich	was	ist das - die/ der •• bis ••• zur Lab	37
17				lieh nach oben ((lacht)) (so zu) hast du	was	verstanden	37

System overview: Corpus workflow



Demonstration: Transcription



- Example: TV debate Royal / Sarkozy
- Video excerpt
- Transcription according to HIAT conventions
- „Verbal“ tier with orthographic transcription, pauses, non-verbal behaviour
- English translation tier
- http://www1.uni-hamburg.de/exmaralda/files/demokorpus/Royal/presentation/royal_partiture.html

Demonstration: Analysis



- Example: Hamburg Map Task Corpus
(http://www.exmaralda.org/corpora/en_sfb_z2.html)
- Task: Explaining a route on a map
- Advanced L2 learners of German with (very) different L1s
- Orthographic transcription only
- Metadata about speakers' language biographies
 - Learned German when and where?
 - Language use in everyday life?

Demonstration: Analysis



- Methodological issues:
 - Different type of relevant context:
 - Interactional context: other speakers, other modalities
 - Situational context: communication metadata
 - Biographic context: speaker metadata
 - „Corpus-driven“ analysis: development of hypotheses in a „dialogue“ with the data
 - Stepwise or iterative refinement and specification of queries, „explorative“ approach
 - Manual processing of search results
 - Ad-hoc categorisation (categories in the corpus vs. categories in the analysis)

Demonstration: Corpus publication



- Example: Hamburg Map Task Corpus
- Similar for other completed corpora at the Research Centre on Multilingualism:
<http://corpora.exmaralda.org>
- Password protected access will be given upon request
- Methods for generating publication format are part of EXMARaLDA
- Similar Corpora elsewhere (e.g. METU Spoken Turkish Corpus in Ankara)
- Online corpus browsing in a web browser
- Online query through EXAKT



- Free download from
<http://www.exmaralda.org>
- Documentation, mailing list, demo data, tutorials from the same website
- Published corpora from the same website

Thank you, Merci!