Oral corpora

Current opportunities and challenges between methodology, technology, and organisation

Dr. Thomas Schmidt

Programmbereich ‚Mündliche Korpora‘

IDS Mannheim

Outline

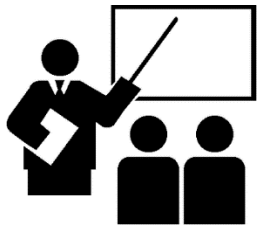
1. Presentation of FOLK – Research and Teaching Corpus of Spoken German
   - Data
   - Workflow
2. Some pilote studies carried out on FOLK
3. Challenges
   - Corpus size
   - Stratification and field access
   - Transcription Bottleneck
4. Beyond FOLK
   - Oral corpora
   - Standardization issues in oral corpora

FOLK

- Forschungs- und Lehrkorpus Gesprochenes Deutsch = Research and Teaching Corpus of Spoken German
- Large corpus of spoken interaction in German
- Aim: to **capture the variety of interaction in society**
- Freely available to the scientific public
- Using the technical and methodological state of the art

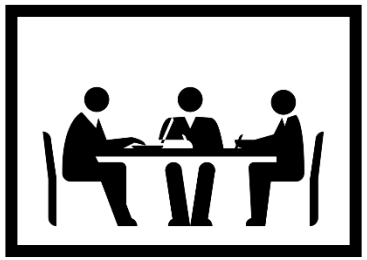
FOLK Data

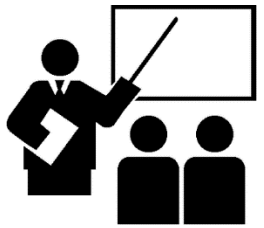
FOLK: Private communication - Couple talk (FOLK_E_00043)



0321	**PB**	°h zum beispiel hat so_n kommilitone von mir *for example, a fellow student of mine*
0322		(1.3)
0323	PB	n jetzt den geleast n zet vier (.) n be em we zet vier *has leased a BMW Z4*
0324	**AM**	(.) ich weiss doch überhaupt [nich was n zet vier is] (.) *I don‘t even know what a Z4 is*
0325	**PB**	[ja des dacht ich mir] *yes I suspected that*
0326	**AM**	ja es (.) interessiert mich auch nich *I‘m not interested*
0327		(2.59)
0328	**AM**	des wird irgend n auto sein *that will be some kind of car*
0329	**PB**	ja

FOLK: Professional communication - team meeting (FOLK_E_00150)





0249	**IW**	wir wissen ja von welchem team wir da (.) und von welcher *we know which team and which*
0250		(0.65)
0251	IW	agentur wir da sprechen *agency we talk about*
0252	**JS**	ma
0253		(0.31)
0254	**IW**	oh
0255		(0.29)
0256	IW	°h sind wir so harmlos [((lacht))] *are we that harmless*
0257	**JS**	[((lacht))]
0258		(0.25)
0259	**IW**	°hh (.) nein ich hab den auch gesacht unser team is verkleinert worden *no I told them our team was downsized*

FOLK: Public communication – Arbitration (FOLK_E_00069)

0096	**BP**	°h die aussage war nicht dass der tunnel °h genau so liegt *the statement was not that the tunnel is exactly like this*
0097	**BP**	von daher is der vorwurf nicht richtig an [he s] an herrn sierig *therefore the accusation to Mr. Sierig is not correct*
0098	**WW**	[doch] *yes*
0099	**BP**	°h nicht zu halten jetzt hab ich ein paar fra[gen++++++] *untenable / I have a couple of questions*
0100	**WW**	[der bleibt rich]tig *it remains correct*
0101	**WW**	darf ich dazu etwas sagen der vorwurf bleibt dennoch richtig weil °h *can I say something / the accusation remains correct because*

FOLK: Authentic interaction

Private communication: Coffee table talk / older speakers from the Mannheim region (FOLK_E_00143)

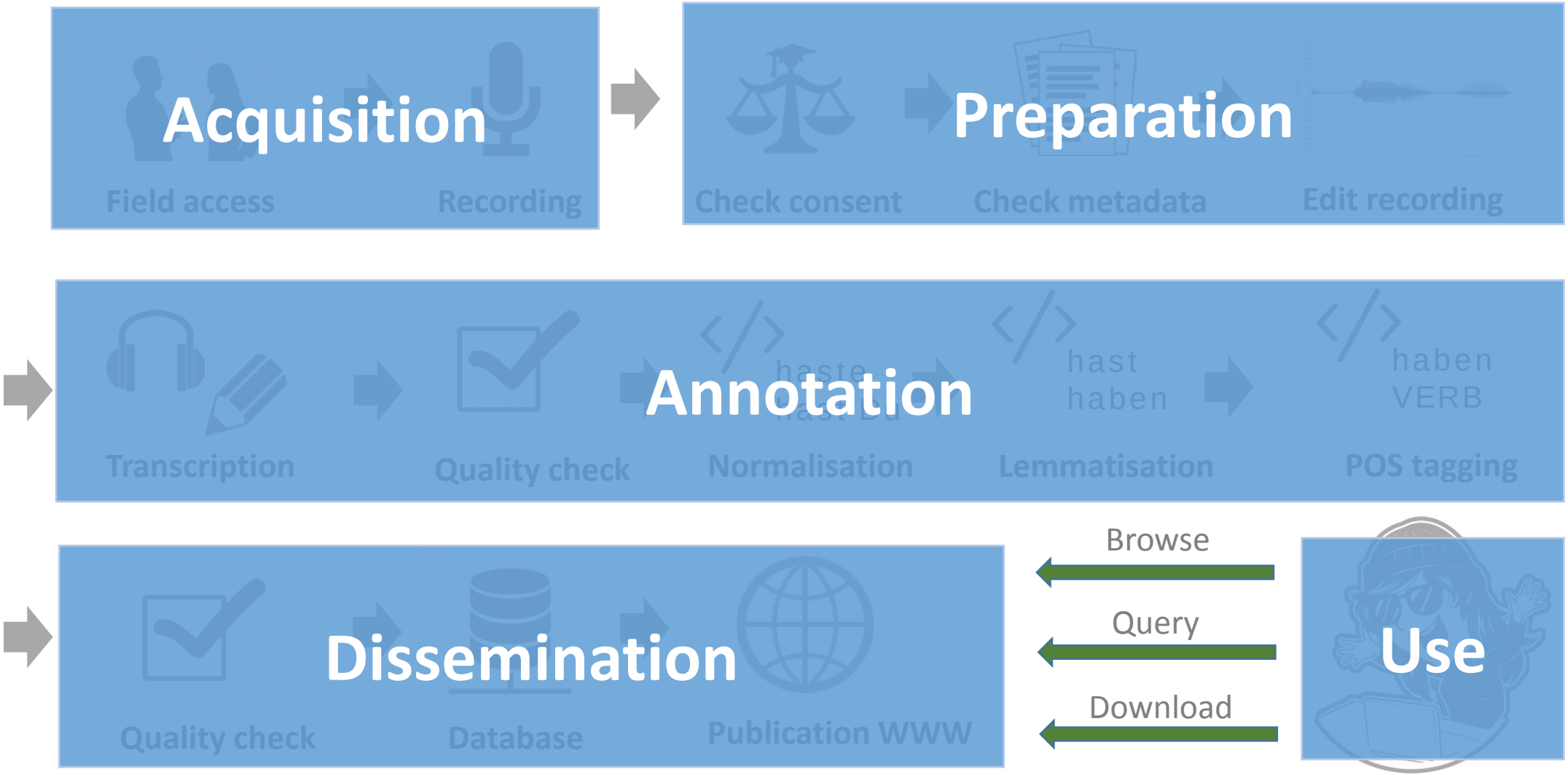
415	JI					[des hat dann]	[jedes m]		[itkriegt d]		[o weesch]	[t]
416	HM	gott]	fried hascht		[trink]	amo [I eigentli]	[ch komm mir]		[trinke mal °h wart]		[emol ich mach je]	[tzet]
417	GI			ah jo isch	[hab]		[was]				[hm_hm]	

- Many overlaps
- Difficulties in speaker identification
- Dialect
- Anonymisation of proper names

- (Background noises)
- (Non-optimal recording quality)
- (Learner language)
- (Child language)
- …

FOLK Workflow

FOLK: Workflow



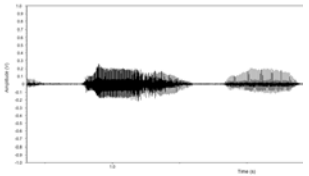
FOLK: Workflow

Field access

Recording

Check consent

Check metadata



Edit recording

Transcription

Quality check

Normalisation

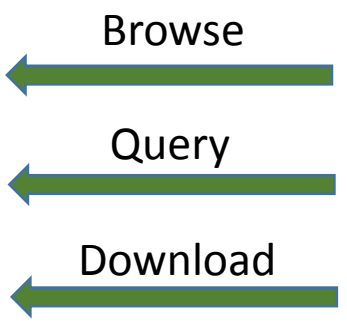
Lemmatisation

POS tagging

Quality check

Database

Publication WWW

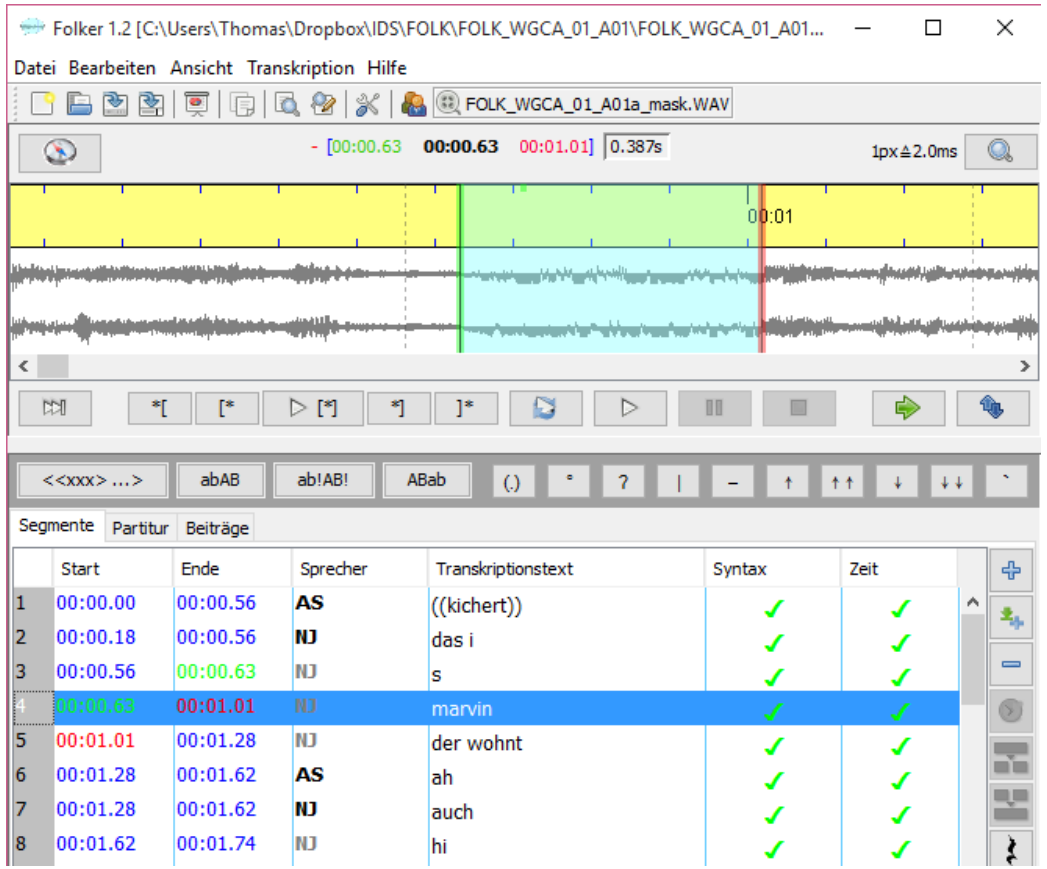


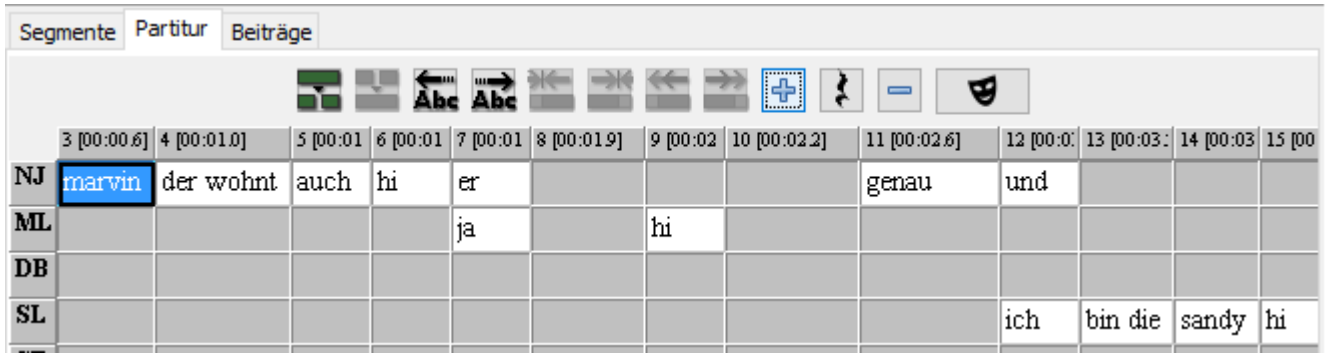
FOLK: Transcription

- Around 1500 hours transcriber time per year
- Using FOLKER – FOLK EditoR
  - time aligned transcription
  - machine readable data (XML markup)
  - automatic checks for consistency (temporal structure, conventions)
- Using cGAT-Minimal conventions
  - based on GAT, related to CA
  - detailed transcription of words in modified orthography
  - detailed transription of pauses, markup of overlaps
  - minimal interpretation effort for transcribers – no prosody, no theory dependent annotations

➔ Optimize transcription efficiency and reusability

FOLK: Transcription





0042	GF	((hustet))
0043	AK	vielleicht [vorher schön]er
0044	LB	[okay]
0045	LB	(.) schreibe ma [rein zündfunke un]
0046	GF	[((zieht die Nase hoch))((grunzt))]
0047		(0.25)
0048	LB	ruhig (.) sie können glei rechts nebendran schreiben wenn sie wollen

FOLK: Token annotation

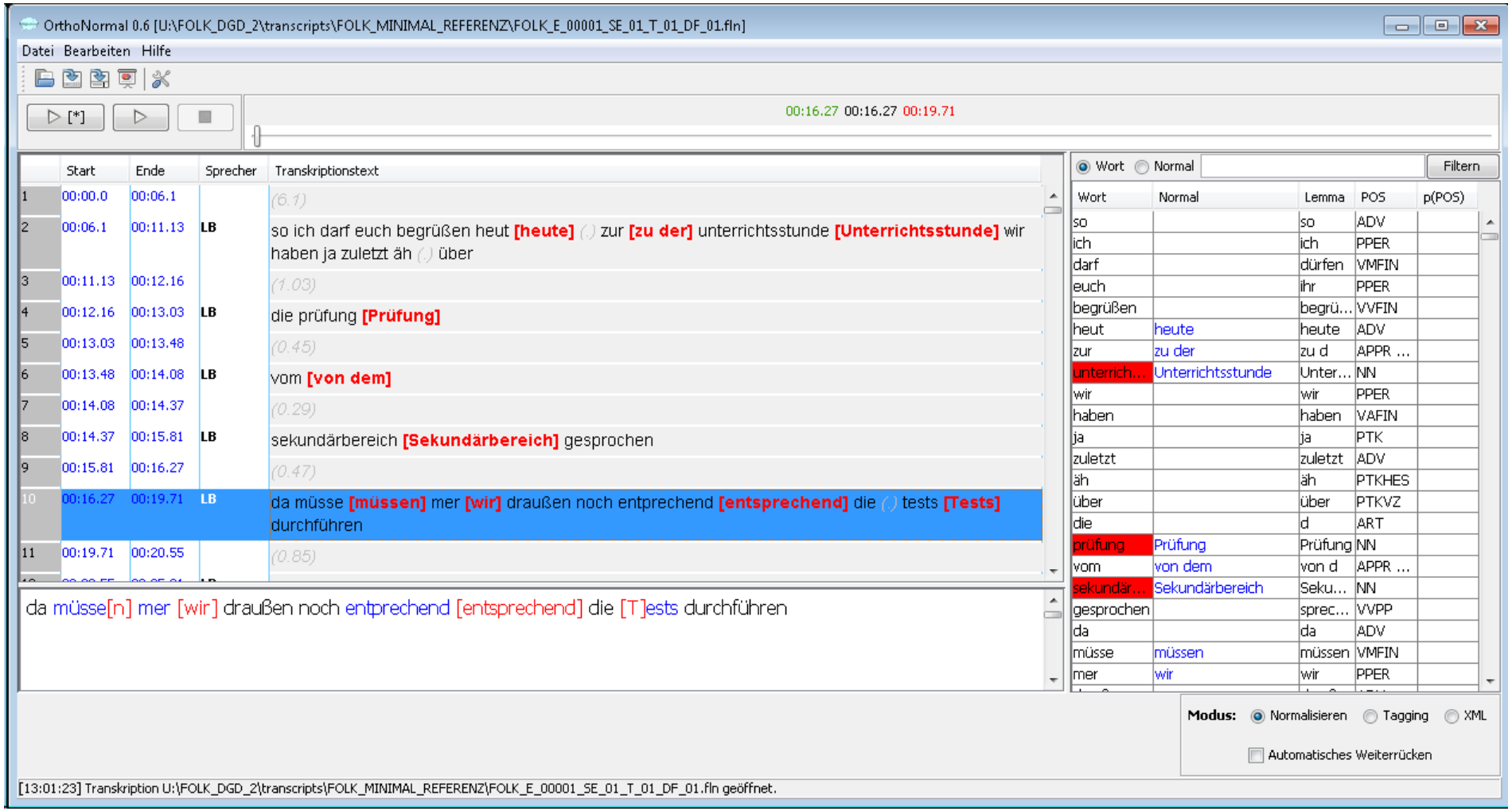
Transcription	**ich**	**nehm**	**auch**	**n**	**stück**	**nee**	**lieber**	**zitrone**
Normalisation	ich	**nehme**	auch	**ein**	**Stück**	**nein**	lieber	**Zitrone**
Lammatisation	ich	**nehmen**	auch	ein	Stück	nein	**lieb**	Zitrone
POS-Tagging	PPER	VVFIN	ADV	ART	NN	SIERSP	ADJA	NN

- Modified orthography („literary transcription“) adequate for
  - representation of pronunciation deviations, disfluencies
  - Visualisation / readability of transcripts
- Inadequate / insufficient for
  - systematic queries (false negatives/underselection, false positives/overselection)
  - automatic processing (e.g. POS tagging)

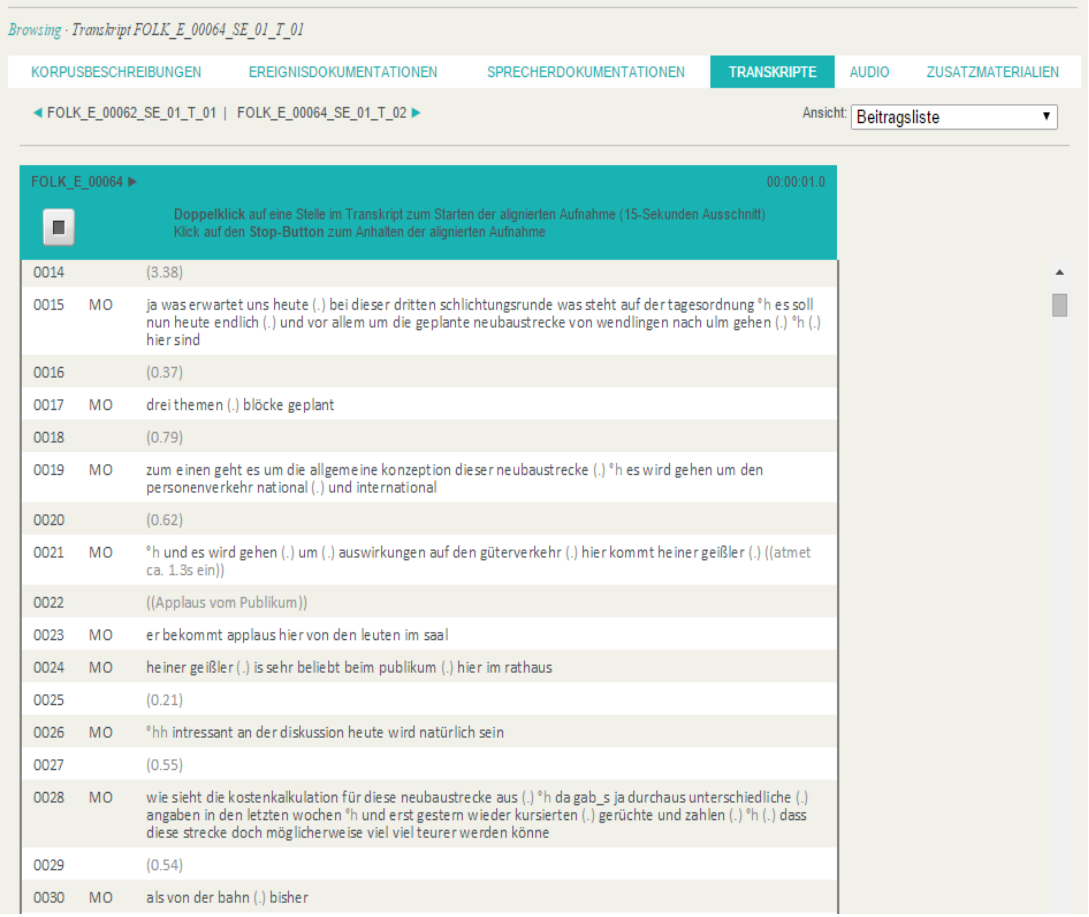
➔Orthographic normalisation as a second annotation layer

➔Lemmatisation and POS tagging based on normalisation

FOLK: Token annotation



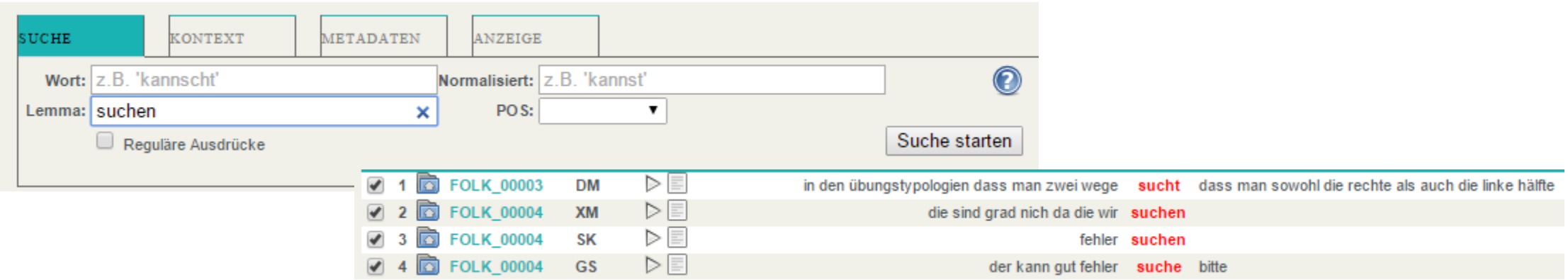
FOLK: Online Browsing



- Read and listen to:
  - Metadata
  - Transcripts
  - Recordings
- Get to know the data
- Qualitative analysis
- Build example collections

➔ DEMO

FOLK: Query

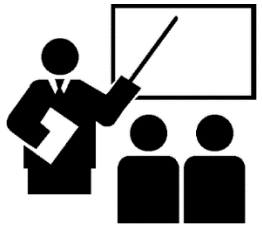


- KWIC concordancing
- Correlate with speaker and interaction metadata
- Manual inspection with access to transcript context and aligned recording

- Quantification
- Corpus driven analysis
- Download excerpts for further analysis / annotation

➔ DEMO

FOLK: Current status



- 169 Interactions, 130h recordings, 1.3 million tokens
- At least + 30h (=300,000 tokens) / year
- 4000 registered DGD users, FOLK most popular corpus
  - Corpus linguistics (computer linguistics / speech technology)
  - Conversation analysis (and related)
  - Language teaching (DaF/DaZ)
  - Constrastive: written / CMC
  - grammar
  - lexicography
  - …

FOLK Analyses

Lemma frequencies

- Compare frequency classes of lemmas between
  - FOLK: spontaneous interaction
  - DeReKo: large „canonic“ written corpus
- For lemmas with pronounced frequency differences:
  - Explain
  - Categorize
  - Determine structure for adequate lexical description

➔ Dictionary of German in Spoken Interaction

Lemma frequencies (Klosa/Schmidt 2014)

Lemma	Häufigkeit FOLK	HK FOLK	HK DeReWo
d	73287	0	0
sein	32475	1	3
ja	31428	1	8
ich	29246	1	5
und	21532	2	2
…	…	…	…
okay	3360	4	14
halt	3182	5	13
eigentlich	1773	5	9
glauben	1632	5	9
irgendwie	1585	6	12
bißchen	1419	6	11
kriegen	1058	6	12
gucken	928	6	13
ungefähr	602	7	12
sozusagen	569	7	13
deswegen	526	7	11
irgendetwas	389	8	14
gell	376	8	17

- Register („umg.“ / colloquial)
- Interaction
  - 1./2. Person
  - Backchannels
  - …
- Modality
  - Modal particles and adverbials
  - Mental verbs
  - Vagueness
- But: small absolute frequencies, „rare“ words not analysable

“Early electronic corpora […] had little impact on lexicography […]. The reason was simple: these were each **corpora of only one million words – corpora so small** that it was impossible to distinguish statistically significant co-occurrences of words from chance co-occurrences.”

Hanks, P. (2012): Corpus evidence and electronic lexicography.
In: Granger, S. & Paquot, M. (2012): Electronic lexicography.

Discourse marker „ich sag mal“

- Grammaticalization?
- Variation in form and function?
- Contexts?
- Frequencies?
- Co-Occurrences?

040	AJ	dat is so dieses *it is like this* hh° also **ich sach ma** wie so_n assitum *well **let me say** like an anti-social thing* nenn ich das jetz mal *that‘s what I call it* das [hört sich jetz bescheu]ert an *it sounds stupid*

[FOLK_E_00191: Biographical Interview, Westphalia]

1	FOLK_00121	SM	des is auch typisch also	sage	mas mal vorsichtig eher zurückgenommen ja
2	FOLK_00130	NL	also also ich	sag	jetzt mal ganz plump politisch rechten
3	FOLK_00099	HUS1	dann ähm	sagen	wir einfach mal das ist ein schornsteinfeger also du weißt
4	FOLK_00121	SM		sagen	ma s mal vorsichtig s gibt nämlich auch längere parabeln
5	FOLK_00015	CH	ich	sag	jetz mal so ba ganz banal ich s s nehm
6	FOLK_00037	BÄ	bildungsbürgerlichen	sag	ich jetz mal ungeschützt familie sich zu einem nicht oder
7	FOLK_00160	PB_aa	dass sie jetzt zumindest auf menschlicher basis	sag	ich einfach mal
8	FOLK_00069	HG	passiert ich	sag	s mal so ganz allgemein

Discourse marker „ich sag mal“ (Schmidt 2014)

In FOLK:

- Frequency in FOLK comparable to lexems of basic vocabulary
- Limited variation in form and function
- Cooccurrence with adverbs like „vorsichtig“ (‚careful‘), „grob“ (‚roughly‘)
- Across all interaction types, regions, speaker types

In other oral corpora:

- NO occurence in data from Western Germany before 1990
- In Berliner Wendekorpus (1990s)
  - Very few occurrences in West Berlin data
  - Many more in East Berlin data

➔„Imported“ from East Germany after reunification

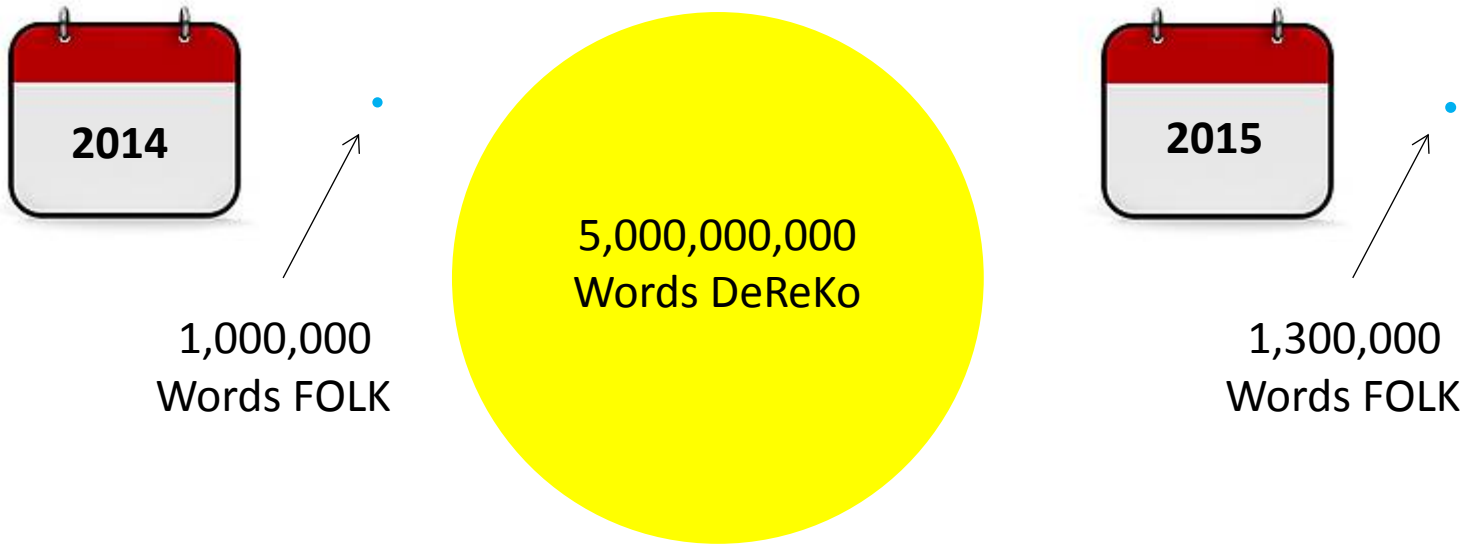
➔But: no (transcribed) data from East Germany before 1990

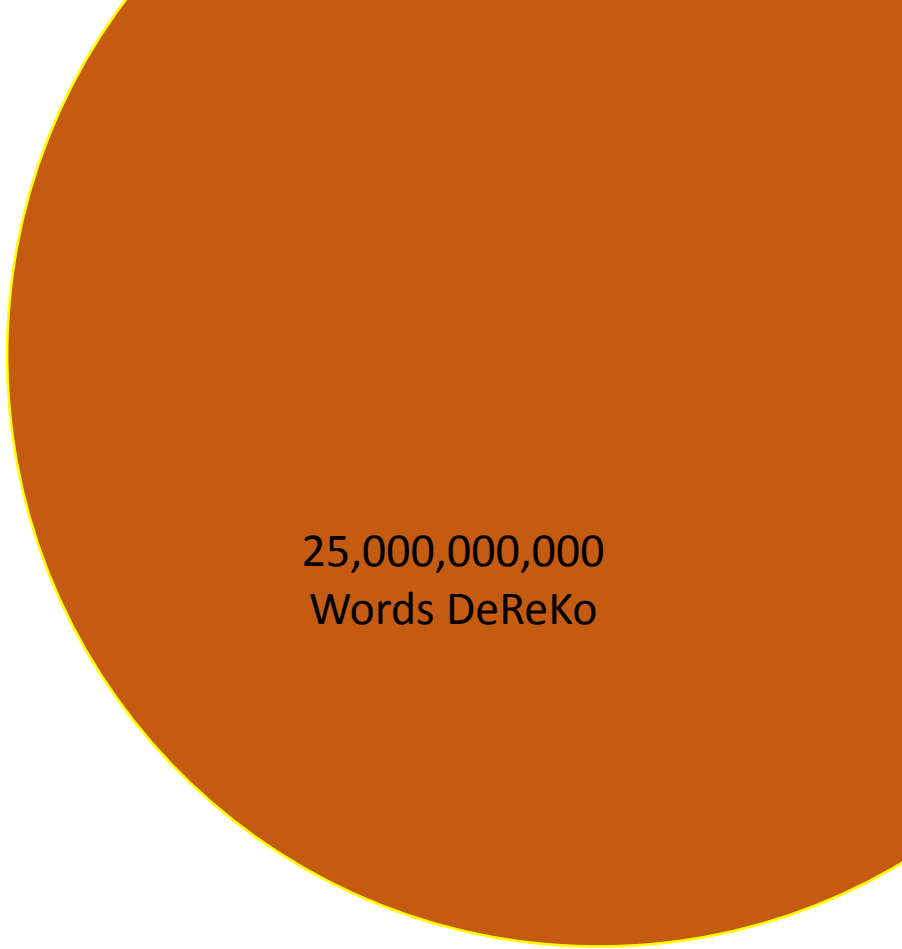
FOLK Challenges

Challenges in building / expanding FOLK

- Corpus design
  - Corpus size
  - Stratification / Sampling
  - Balance / Representativity
- Data acquisition
  - Field access
  - Recording technology
  - Meta data
  - Legal authorisation
- Corpus processing
  - Transcription / Alignement
  - Annotation
  - Segmentation
- Corpus use
  - Visualisation
  - Query
  - Correlation of linguistic and extra-linguistic data
  - Quantification

Corpus size





- Millions vs. billions of Words
- „Exponential“ vs. linear growth / automatic vs. manual data acquisition and processing
- Realistic mid-term goal for FOLK: 10 million words / 1000 hours
  ➔ BNC Spoken, Corpus Gesproken Nederlands, Deutsch Heute

Stratification / Sampling

Primary level: interaction types					
Private		**Institutional**			**Public**
Everyday talk	**Activities**	**Education**	**Work**	**Services**	
Couple talk Going for a walk Table conversation …	Cooking Playing poker …	German lesson Seminar Examination …	Team meeting Job interview …	Sales talk At a police station …	Panel discussion …
Secondary level: speaker properties					
Gender	**Region**	**Age**	**Education**	**…**	
female male	North Germany Bavaria … (15)	-20 20-60 60-	Lower Higher		

Stratification / Sampling

```
Dyadic
Telephone conversation
Administration
Unfamiliar speakers
Northern Germany
Between 30 and 60
Higher education
```

„to **capture the variety of interaction in society**“

- No one-dimensional classification of interaction types
- Feature bundles / prototypes
- Combinatorial explosion when intersecting feature bundles

➔Multidimensional corpus design

➔Not a globally balanced corpus

➔Strategy:

„Maximal Dispersion“ (Kupietz/Schmidt 2015)

„Breadth before depth“ (Deppermann/Hartung 2011)

Stratification and field access

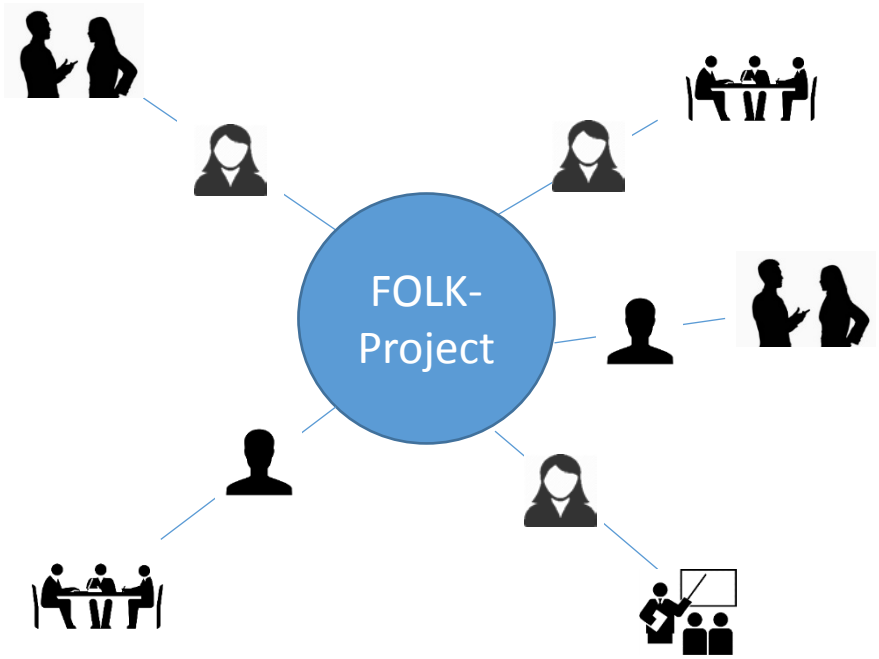
- Corpus design + stratification strategy require:
  - Authentic, spontaneous recordings
  - Good documentation, authorization
  - Maximum variation of interaction types
  - Full coverage of secondary variables (region, age, …)

➔No centralised planning possible

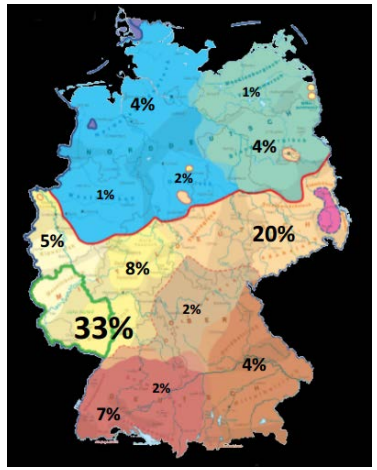
➔Need many personal contacts for field access

➔Network of recording opportunities

Stratification and field access

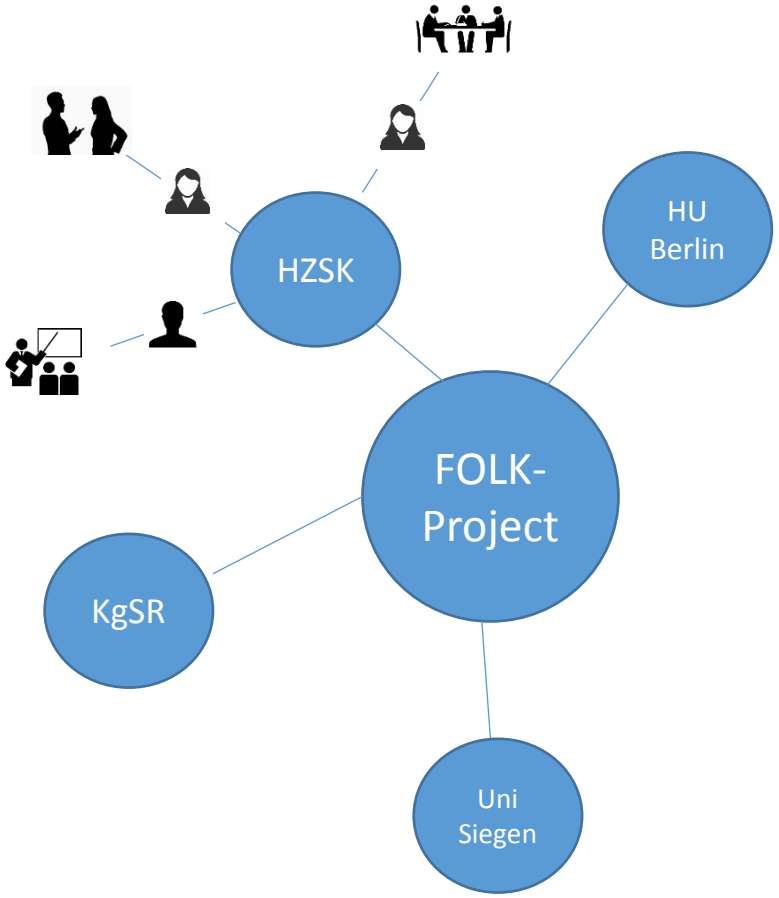


- First project phase (1 million tokens): field access via students and project members



➔ 33% of interactions from the Mannheim region
➔ 42% of speakers between 20 and 30 years old
➔ 57% of speakers with higher education

Stratification and field access



- Currently: More cooperations and „data donations“
- „wish list“ + FOLK-“starter kit“

➔ More options / better control of stratification
➔ Increased costs for management and documentation of good practices

- Ideally:
  - Expand the network
  - Use a platform for management and documentation
  - „Zero“ Marginal Costs (?)

BNC BRITISH NATIONAL CORPUS 2014

We […] encourage people […] to record their interactions and send them to us […]. For each hour of good quality recordings […] we will pay £ 18.

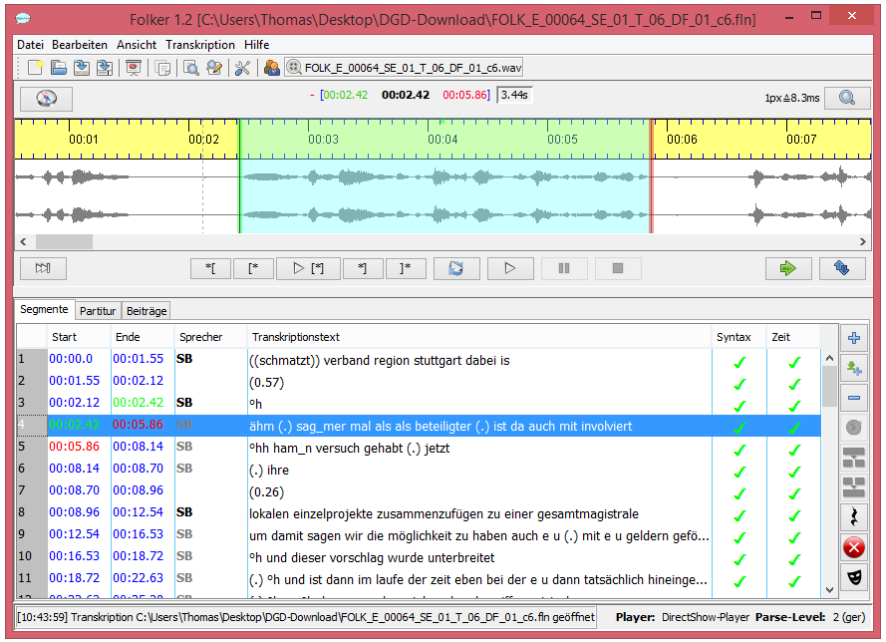
Corpus processing

- „Transcription bottleneck“ (Brinckmann 2009)
  - Up to 100h manual transcription for 1h of recording
  - 1000 EUR for 10.000 tokens → 10 cent per Token / around 30 years for 10 million tokens!
  - ➔Transcription bottleneck decisive factor for attainable corpus size / variation

- Dependencies:
  - Characteristics of the interaction / Quality of the recording
  - Experience of the transcriber / External conditions for transribing

Corpus processing: transcription bottleneck

- Methodological measures
  - Preselection: „Simple“ recordings
  - Simplify transcription conventions
  - Improve guidelines
- Organisational measures
  - Professional transcribers
  - Constantly check quality and efficiency
  - Modularize workflow
- Technological measures
  - Optimize efficiency of transcription tools
  - Automate!



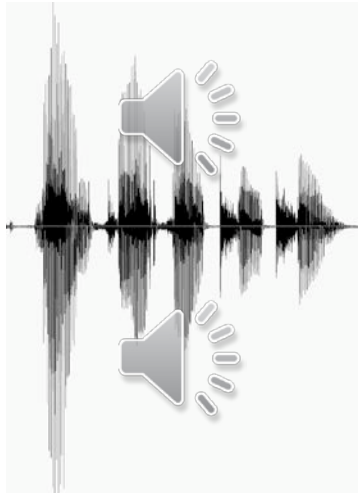
Automatic Speech Recognition

Pilot study with a commercial platform – 100 EURO for 10h (5000 EURO for unlimited amount)

zu den Optimierungspotentialen die der Herr Doktor Käfer hier schon ausführlich vorgestellt hat

die Tür zumachen (0.3) äh die die die äh (.) Terrassentür zumachen

„close the terrace door“



zu den Optimierungspotentialen die der Herr Doktor **Kiefer wir** schon ausführlich vorgestellt hat

und das Fahrrad dagegen ist erst nach dem Tod

„the bicycle by contrast is after death“

- Full automisation an illusion
- Save up to 50% transcription time with suitable data (see AVATECH-Projekt)
- Only small proportion (<10%) of data „suitable“
- Combine automatic and manual methods to realise the potential

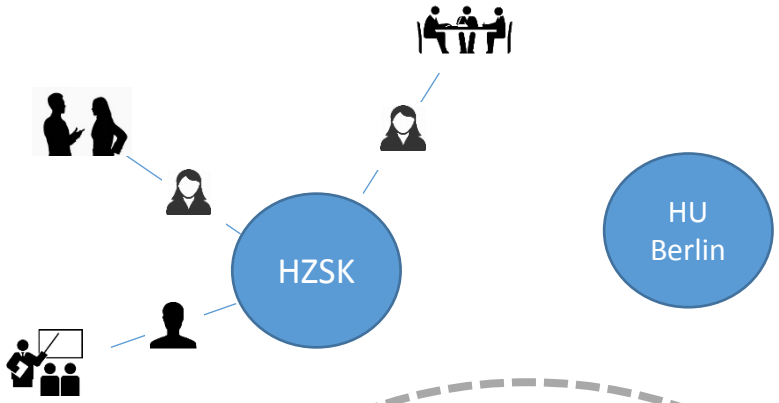
- Not: Transcriber + Tool + Recording = Transcription
- More complex workflows:
  - automatic methods, evaluation, manual work, modification of automatic methods, …
- Different for different recordings
- Additional expertise required (speech technologists)
- Additional organisational overhead

Technology

Silence detection
Speaker diarization
Automatic speech recognition

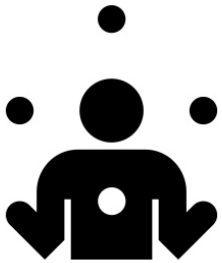
Research / Student assistants

Recording selection
Quality control
Manual correction
Adaptation



Data acquisition

Corpus / Project management



Training
Guidelines
Monitoring
Documentation

Technology

Silence detection
Speaker diarization
Automatic speech recognition

Research / Student assistants

Recording selection
Quality control
Manual correction
Adaptation

Management summary

- FOLK is interesting
- FOLK is small
- Enlarging FOLK is expensive
- Cost reduction requires a mix of methodological, technical and organisational measures
- The more we mix the more organisational overhead we get
- Crowdsourcing is a concept to explore
- There is much unused potential in technology

Beyond FOLK

Oral corpora in an international perspective

BNC Spoken 2014

AusNC Spoken (2012-)

BNC Spoken (1990s)

ICE, SBCSAE, MICASE, …

Spoken Dutch Corpus (1998)

Spokes/PELCRA

Czech National Corpus (2006-)

FOLK (2008-)

GeWiss (2009)

Multilingual Corpora at HZSK (2009)

KidKo (2008)

Texas German (2001)

…

ESLO 2 (2008-)

CIEL-F (2008)

PCF (2004)

Rhapsodie (2008)

Discours sur la ville (2000)

CLAPI, C-ORAL-ROM, VALIBEL, …

GOS Corpus of Spoken Slovene (2009-)

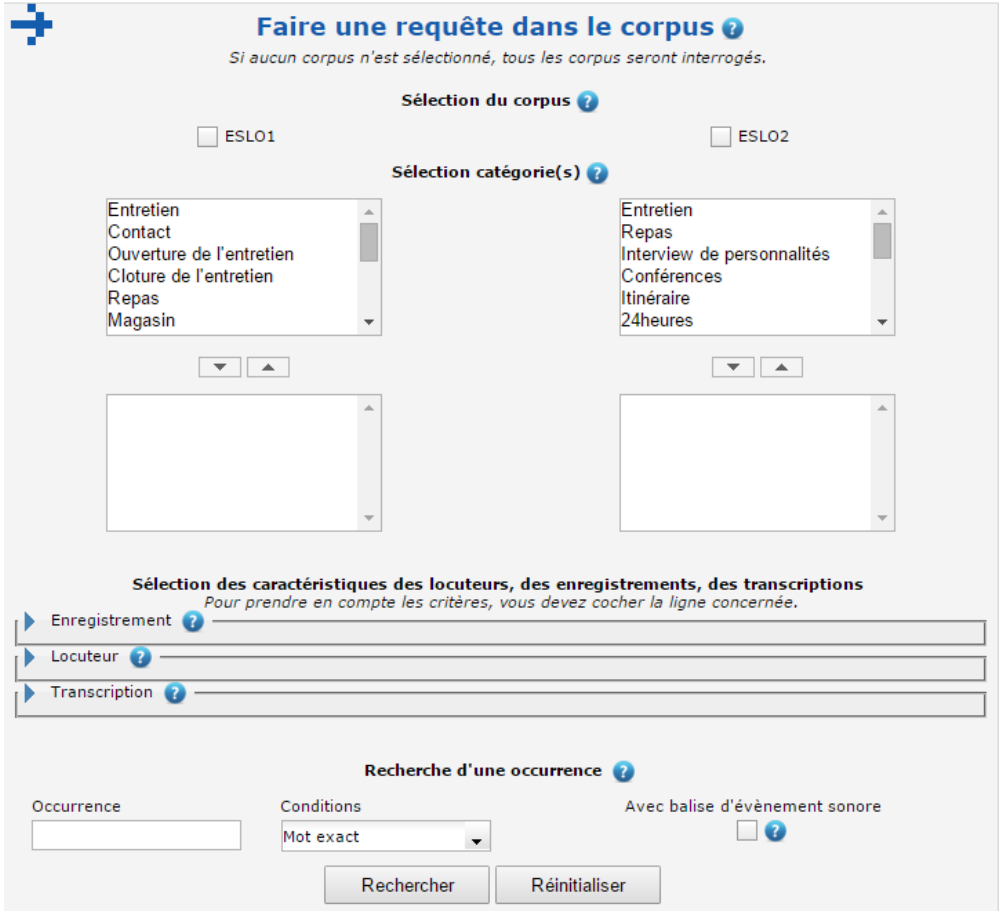
Reference Corpus of Contemporary Portuguese

Multimedia Corpus of Spoken Bulgarian (2013)

Oral corpora in an international perspective

- Many recent/ongoing ambitious projects
- Different corpus designs with common basis: spontaneous talk
- Similar workflows, similar needs and challenges
- But so far: little commonalities in terms of
  - metadata schemes
  - transcription tools: Transcriber, FOLKER, EXMARaLDA, ELAN, …
  - transcription conventions: cGAT, ESLO, …
  - further annotation tools / methods / formats
  - dissemination methods
  - browsing and query interfaces

ESLO Corpus



Enregistrement	Locuteur(s)	Transcription(s)
ESLO1_ENT_001	BA725 001LOC1 001LOC2 OU	ESLO1_ENT_001_C
OU	encore oui **maintenant** oui	
BA725	encore **maintenant** oui	
BA725	euh la vente sous cellophane malheureusement va s'étendre puisque c'est dans les traditions **maintenant**	
BA725	à l'époque euh on m'a pas demandé mon avis vous savez c'était pas comme **maintenant** hein	
BA725	j'aimerais visiter les musées par exemple monter à Paris d'ici **maintenant** c'est la banlieue	
BA725	mais qui sont des métiers rémunérateurs croyez-moi **maintenant** hein	
BA725	**maintenant** hein euh un bon chef peut gagner cent cinquante cent quatre vingt mille francs par mois	
BA725	mais autrement on on peut dire que les trois quarts des des des ouvriers **maintenant** hein	
BA725	parce que je crois que les études ont été plus poussées et puis qu'on aperçoit **maintenant** que les enfants ont une facilité beaucoup plus grande	
OU	**maintenant** est-ce que	
BA725	puis c'est devenu une habitude et puis **maintenant** c'est elle qui écrit	

CLAPI Database

Accueil > Requêtes > Rechercher un mot

Rechercher un mot - Concordancier multimedia

Etape 1 : Saisie des critères > Etape 2 : Nombre d'attestations > Etape 3 : Concordancier multimedia

Type de requête: **Recherche par mot** | Recherche par référence

Mot recherché : (ou token recherché) maintenant

Contexte : (Nombre de mots à afficher : 60 max) 60

☐ Inclure les formes élidées

☐ Respecter les accents

Rechercher : ◉ Par ordre de fréquence dans les corpus ○ Par nom de corpus

> Afficher le nombre d'attestations

Apéritif entre ami(e)s - chat ~ Apéritif entre ami(e)s - chat ~ Aperitif chat -xml — Informations du corpus

1 sur 318

Apéritif entre ami(e)s - chat ~ Apéritif entre ami(e)s - chat ~ Aperitif chat -xml
Transcription : Aperitif chat -xml
Retrouver directement cette attestation dans le concordancier avec son code **15c/c4n**

JUL parce qu'après en fait c'est pas directement lié à la linguistique forcément\ mais en fait y a plein d` métiers aussi qui sont pas directement liés à:&
JUL &j` sais pas moi géographie ou histoire et quand même c'est utile d'avoir un: background [quoi]
JEA [hm hm]
JUL donc après ça dépend vraiment c` qu'on en fait\ (.) .h mais moi en fait j` suis vraiment contente avec mes études `fin c'est clair que maintenant quand je regarde des langues/ j'ai vraiment un regard qu'est complètement différent (.) °tu vois ou alors vraiment `fin: develop`ment euh: d'où ça vient:
JEA ouais [c'est ouais ouais ça peut être]
JUL [y a vraiment euh les différentes] stratégies de: c'est ouais

Passage/Timing

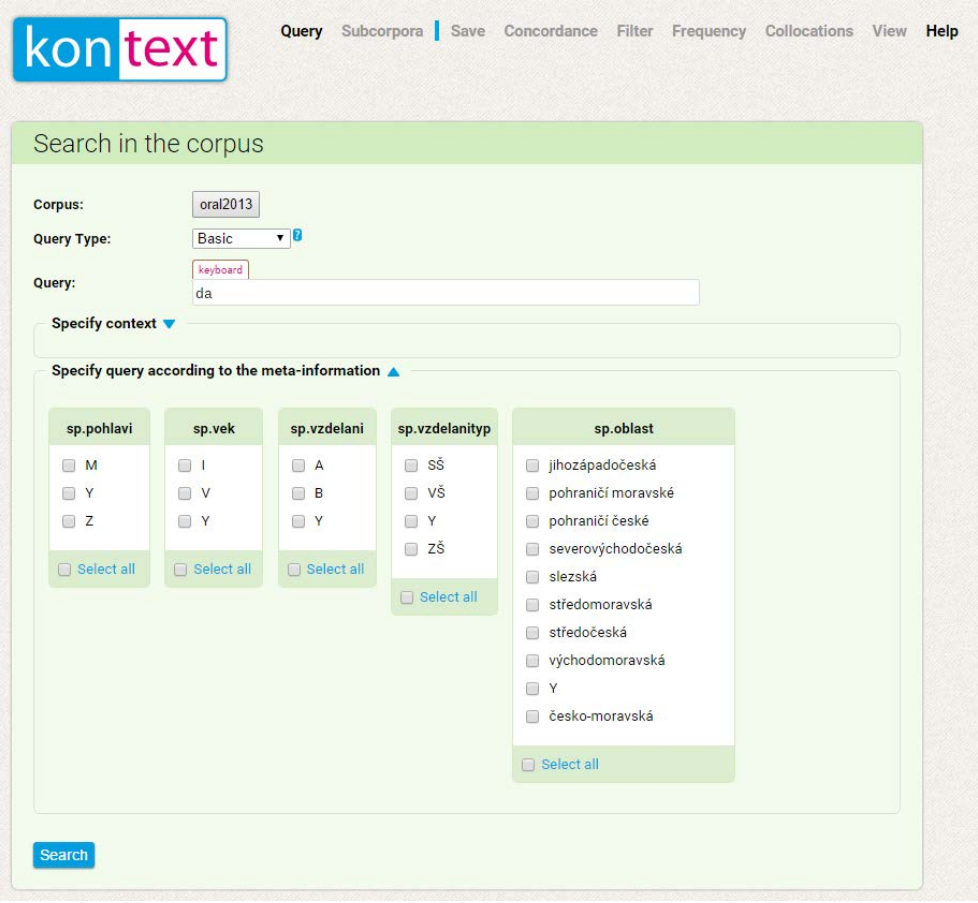
Jouer 2 s | Télécharger 20 s

Jouer 4 s | Télécharger 40 s

Télécharger :

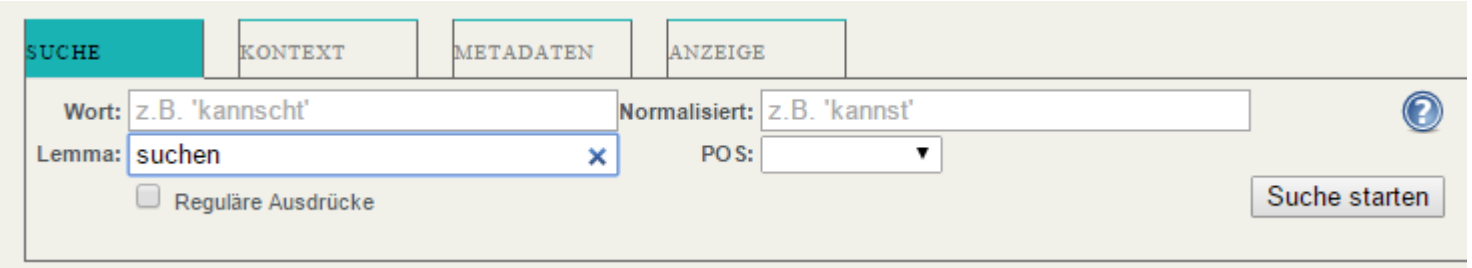
- l'enregistrement (mp4)
- la transcription (trico)
- la transcription (trs)
- la transcription (rtf)

ORAL 2013 (Czech National Corpus)



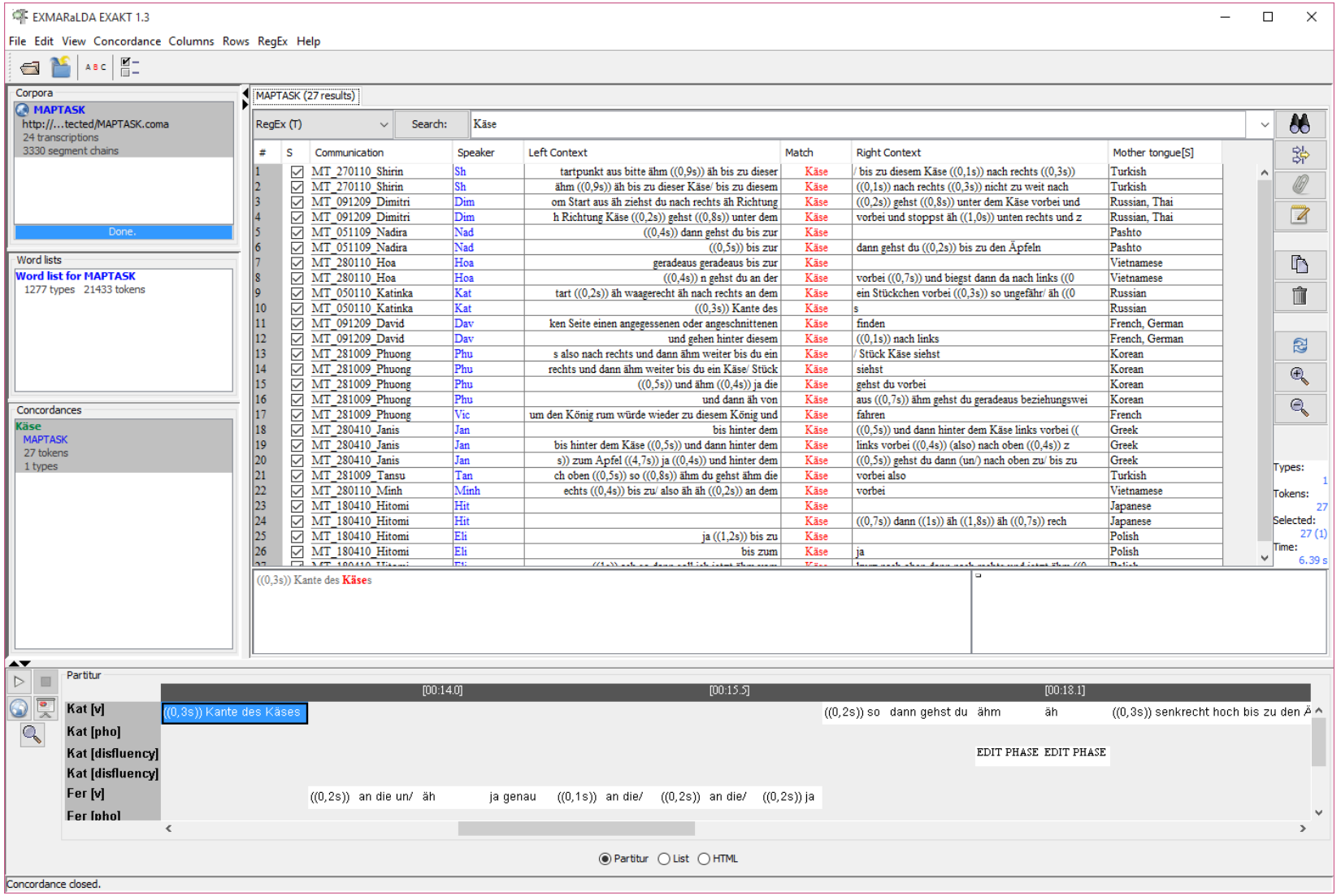
GOS (Corpus of Spoken Slovene)

FOLK (Database for Spoken German)



☑	1	FOLK_00003	DM	in den übungstypologien dass man zwei wege	sucht	dass man sowohl die rechte als auch die linke hälfte
☑	2	FOLK_00004	XM	die sind grad nich da die wir	suchen	
☑	3	FOLK_00004	SK	fehler	suchen	
☑	4	FOLK_00004	GS	der kann gut fehler	suche	bitte

Hamburg Maptask Corpus (EXAKT tool)



	HZSK	AGD	MPI	CHILDES/Talkbank
Corpus data model / format	Coma XML	DGD XML	IMDI	CHAT/XCHAT
Corpus / metadata editor(s)	EXMARaLDA Corpus Manager	Web interface	IMDI Editor / Arbil	CLAN
Transcription data model / format	EXMARaLDA XML	FOLKER XML	ELAN XML	CHAT/XCHAT
Transcription tool(s)	EXMARaLDA Partitur-Editor	FOLKER	ELAN	CLAN
Preferred transcription convention	HIAT	GAT	no preference	CHAT
Corpus query / dissemination tool(s)	EXAKT / Fedora repository	DGD	LAMUS / IMDI Browser / TROVA	CLAN / Browsable database

From Schmidt (2014): (More) Common Ground for Spoken Language Corpora?

- On a sufficiently abstract level, all oral corpus projects work on the same set of problems
- Software stacks / workflows of different oral corpusprojects have little common ground

What if…

- … we could share experiences with workflows…
- … we could jointly develop or improve …
  - workflows
  - transcription and annotation tools and methods
  - dissemination methods
- … we could reuse …
  - guidelines and conventions
  - platforms for project organisation
  - portals for corpus exploitation

Oral corpora in an international perspective

- Exchange
  - Corpus linguistic conferences and journals?
  - Speech and language technology?
  - Conversation analysis / phonetics / language learning / dialogue systems / …

➔JLC 2015

- Cooperation
  - Larger infrastructure projects (DARIAH / CLARIN)?
  - Digital Humanities?

➔ IRCOM

- Best practices
  - Documentation of corpus creation processes?
  - Legal issues?

➔Thompson (2004)
➔Baude/Benveniste (2006)
➔Ruhi et al. (2014)
➔Kirk/Andersen (2016)
…

- Standards
  - For metadata?
  - For transcription and annotation?

➔TEI/ISO transcription standard

21	FOLK_00070	HS		zeit auch schon festgelecht war untendrunter vielen dank das sin die fümf komma
22	FOLK_00069	HG		äh vielen dank äh das äh bedeutet
23	FOLK_00064	HG		gut äh v äh vielen dank
24	FOLK_00126	SF		vielen dank
25	FOLK_00069	HG		ja herr professor wittke vielen dank
26	FOLK_00021	DK		ey viel vielen dank subber
27	FOLK_00070	PC		vielen dank
28	FOLK_00173	TB		super vielen herzlichen dank
29	FOLK_00064	HG		gut vielen dank also jetzt käme glaub der herr conradi dran nich oder
30	FOLK_00021	CH		pascal vielen dank
31	FOLK_00007	GS		dann mal vielen dank bis nächste woche