

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

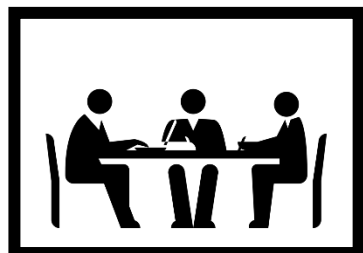
0326 **AM** ja es (.) interessiert mich auch nich
I'm not interested

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 gesagt 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)

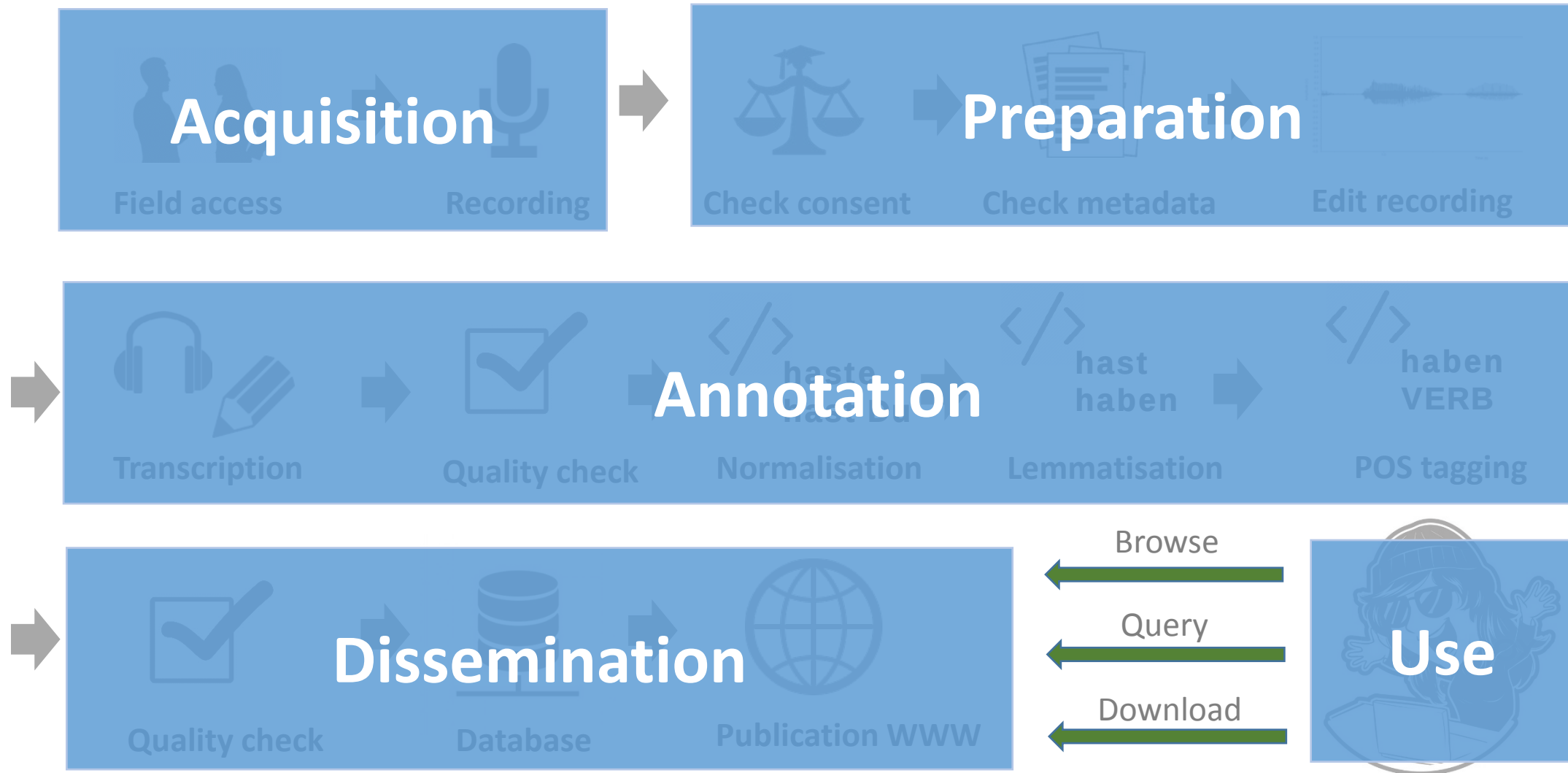
415	Jl		[des hat dann]	[jedes m]	[itkriegt d]	[o weesch]	[t]
416	HM	gott] fried hascht	[trink] amo	[l eigentli]	[ch komm mir]	[trinke mal °h wart]	[emol ich mach je] [tzet]
417	Gl	ah jo isch	[hab]	[was]		[hm_hm]	

- Many overlaps
- Difficulties in speaker identification
- Dialect
- Anonymisation of proper names
- (Background noises)
- (Non-optimal recording quality)
- (Learner language)
- (Child language)
- ...

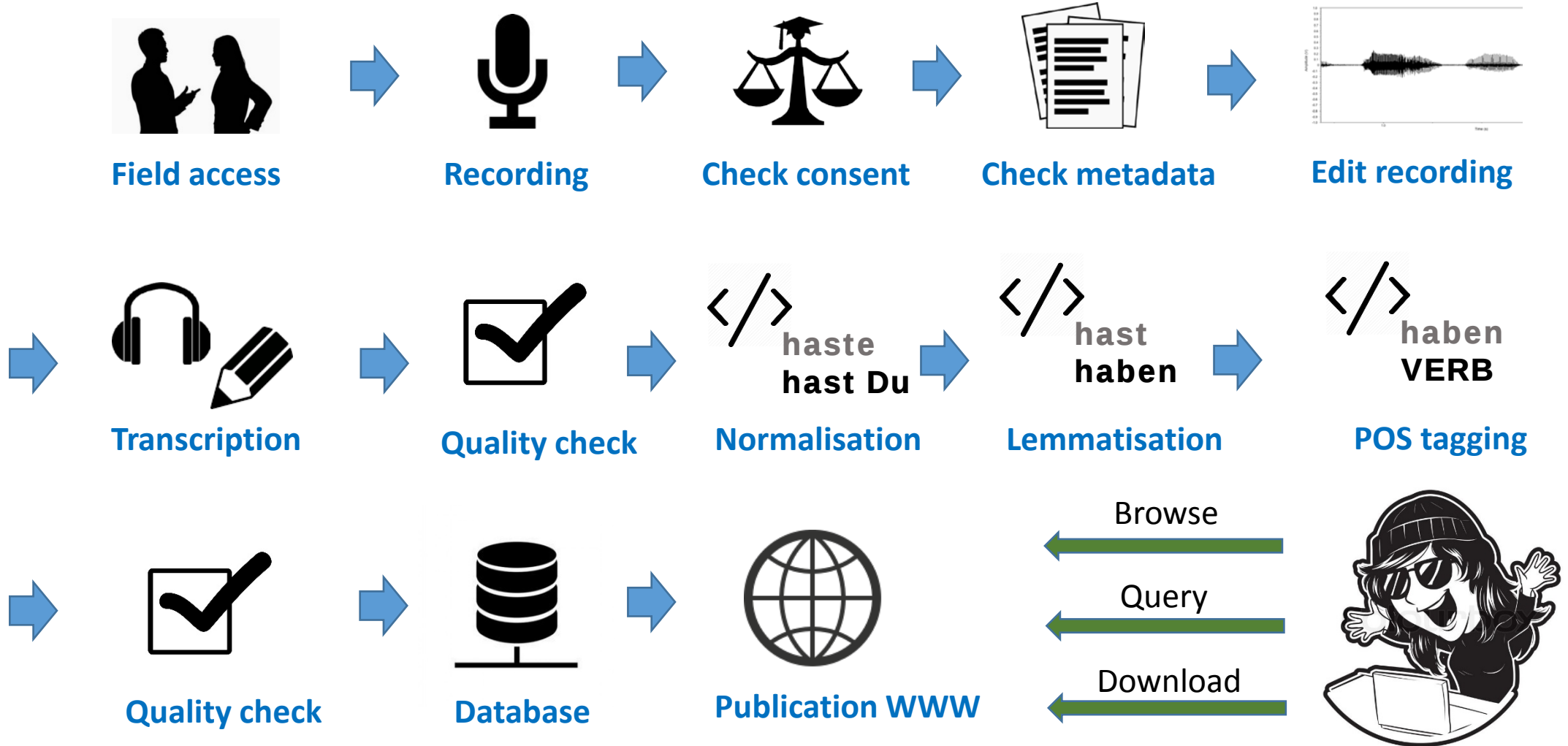
FOLK Workflow



FOLK: Workflow



FOLK: Workflow



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 transcription of pauses, markup of overlaps
 - minimal interpretation effort for transcribers – no prosody, no theory dependent annotations
- ➔ Optimize transcription efficiency and reusability

FOLK: Transcription

Folker 1.2 [C:\Users\Thomas\Dropbox\IDS\FOLK\FOLK_WGCA_01_A01\FOLK_WGCA_01_A01...]

Datei Bearbeiten Ansicht Transkription Hilfe

FOLK_WGCA_01_A01a_mask.WAV

[00:00.63 00:00.63 00:01.01] 0.387s 1px ± 2.0ms

Transcription table:

Segmente	Partitur	Beiträge				
Start	Ende	Sprecher	Transkriptionstext	Syntax	Zeit	
1	00:00.00	00:00.56	AS	((kichert))	✓	✓
2	00:00.18	00:00.56	NJ	das i	✓	✓
3	00:00.56	00:00.63	NJ	s	✓	✓
4	00:00.63	00:01.01	NJ	marvin	✓	✓
5	00:01.01	00:01.28	NJ	der wohnt	✓	✓
6	00:01.28	00:01.62	AS	ah	✓	✓
7	00:01.28	00:01.62	NJ	auch	✓	✓
8	00:01.62	00:01.74	NJ	hi	✓	✓

Segmente

Partitur

Beiträge

	3 [00:00.6]	4 [00:01.0]	5 [00:01]	6 [00:01]	7 [00:01]	8 [00:01.9]	9 [00:02]	10 [00:02.2]	11 [00:02.6]	12 [00:03]	13 [00:03]	14 [00:03]	15 [00:03]
NJ	marvin	der wohnt	auch	hi	er				genau	und			
ML					ja		hi						
DB													
SL										ich	bin die	sandy	hi

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

OrthoNormal 0.6 [U:\FOLK_DGD_2\transcripts\FOLK_MINIMAL_REFERENZ\FOLK_E_00001_SE_01_T_01_DF_01.fln]

Datei Bearbeiten Hilfe

00:16.27 00:16.27 00:19.71

	Start	Ende	Sprecher	Transkriptionstext
1	00:00.0	00:06.1		(6.1)
2	00:06.1	00:11.13	LB	so ich darf euch begrüßen heut [heute] (.) zur [zu der] unterrichtsstunde [Unterrichtsstunde] wir haben ja zuletzt äh (.) über
3	00:11.13	00:12.16		(1.03)
4	00:12.16	00:13.03	LB	die prüfung [Prüfung]
5	00:13.03	00:13.48		(0.45)
6	00:13.48	00:14.08	LB	vom [von dem]
7	00:14.08	00:14.37		(0.29)
8	00:14.37	00:15.81	LB	sekundärbereich [Sekundärbereich] gesprochen
9	00:15.81	00:16.27		(0.47)
10	00:16.27	00:19.71	LB	da müsse [müssen] mer [wir] draußen noch entsprechend [entsprechend] die (.) tests [Tests] durchführen
11	00:19.71	00:20.55		(0.85)

da müsse[n] mer [wir] draußen noch entsprechend [entsprechend] die [T]ests durchführen

Wort	Normal	Lemma	POS	p(POS)
so		so	ADV	
ich		ich	PPER	
darf		dürfen	VMFIN	
euch		ihr	PPER	
begrüßen		begrü...	VVFIN	
heut	heute	heute	ADV	
zur	zu der	zu d	APPR ...	
unterricht...	Unterrichtsstunde	Unter...	NN	
wir		wir	PPER	
haben		haben	VAFIN	
ja		ja	PTK	
zuletzt		zuletzt	ADV	
äh		äh	PTKHES	
über		über	PTKVZ	
die		d	ART	
prüfung	Prüfung	Prüfung	NN	
vom	von dem	von d	APPR ...	
sekundär...	Sekundärbereich	Seku...	NN	
gesprochen		sprec...	VVPP	
da		da	ADV	
müsse	müssen	müssen	VMFIN	
mer	wir	wir	PPER	

Modus: ☒ Normalisieren ☐ Tagging ☐ XML

☐ Automatisches Weiterücken

[13:01:23] Transkription U:\FOLK_DGD_2\transcripts\FOLK_MINIMAL_REFERENZ\FOLK_E_00001_SE_01_T_01_DF_01.fln geöffnet.

FOLK: Online Browsing

Browsing · Transkript FOLK_E_00064_SE_01_T_01

KORPUSBESCHREIBUNGEN EREIGNISDOKUMENTATIONEN SPRECHERDOKUMENTATIONEN **TRANSKRIPTE** AUDIO ZUSATZMATERIALIEN

◀ FOLK_E_00062_SE_01_T_01 | FOLK_E_00064_SE_01_T_02 ▶ Ansicht: Beitragsliste

FOLK_E_00064 ▶ 00:00:01.0

Doppelklick auf eine Stelle im Transkript zum Starten der alignierten Aufnahme (15-Sekunden Ausschnitt)
Klick auf den Stop-Button zum Anhalten der alignierten Aufnahme

0014	(3.38)
0015	MO ja was erwartet uns heute (.) bei dieser dritten schlichtungsrunde was steht auf der tagesordnung *h es soll nun heute endlich (.) und vor allem um die geplante neubaustrecke von wendlingen nach ulm gehen (.) *h (.) hier sind
0016	(0.37)
0017	MO drei themen (.) blöcke geplant
0018	(0.79)
0019	MO zum einen geht es um die allgemeine konzeption dieser neubaustrecke (.) *h es wird gehen um den personenverkehr national (.) und international
0020	(0.62)
0021	MO *h und es wird gehen (.) um (.) auswirkungen auf den güterverkehr (.) hier kommt heiner geißler (.) ((atmet ca. 1.3s ein))
0022	((Applaus vom Publikum))
0023	MO er bekommt applaus hier von den leuten im saal
0024	MO heiner geißler (.) is sehr beliebt beim publikum (.) hier im rathaus
0025	(0.21)
0026	MO *hh intressant an der diskussion heute wird natürlich sein
0027	(0.55)
0028	MO wie sieht die kostenkalkulation für diese neubaustrecke aus (.) *h da gab_s ja durchaus unterschiedliche (.) angaben in den letzten wochen *h und erst gestern wieder kursierten (.) gerüchte und zahlen (.) *h (.) dass diese strecke doch möglicherweise viel viel teurer werden könne
0029	(0.54)
0030	MO als von der bahn (.) bisher

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

➔ DEMO

FOLK: Query

SUCHE KONTEXT METADATEN ANZEIGE

Wort: z.B. 'kannst' Normalisiert: z.B. 'kannst' ?

Lemma: suchen x POS: ▾

☐ Reguläre Ausdrücke Suche starten

☒	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

- 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

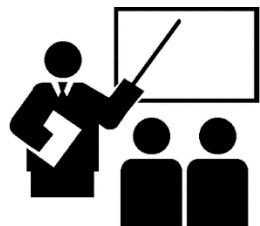
FOLK: Current status



- 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
- ...

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 <i>it is like this</i> hh° also ich sach ma wie so_n assitum <i>well let me say like an anti-social thing</i> nenn ich das jetzt mal <i>that's what I call it</i> das [hört sich jetzt bescheu]ert an <i>it sounds stupid</i>
-----	----	--

[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 jetzt mal so ba ganz banal ich s s nehm
☒	6		FOLK_00037	BÄ			bildungsbürgerlichen sag ich jetzt 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 occurrence 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

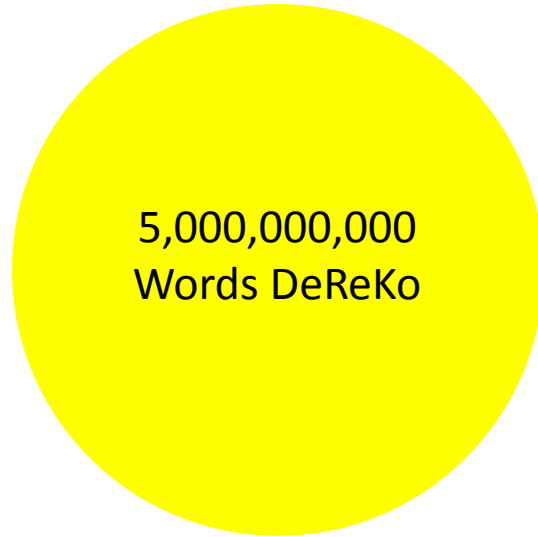
- Corpus design
 - Corpus size
 - Stratification / Sampling
 - Balance / Representativity
- Data acquisition
 - Field access
 - Recording technology
 - Meta data
 - Legal authorisation
- Corpus processing
 - Transcription / Alignement
 - Annotation
 - Segmentation
- Corpus use
 - Visualisation
 - Query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

Corpus size



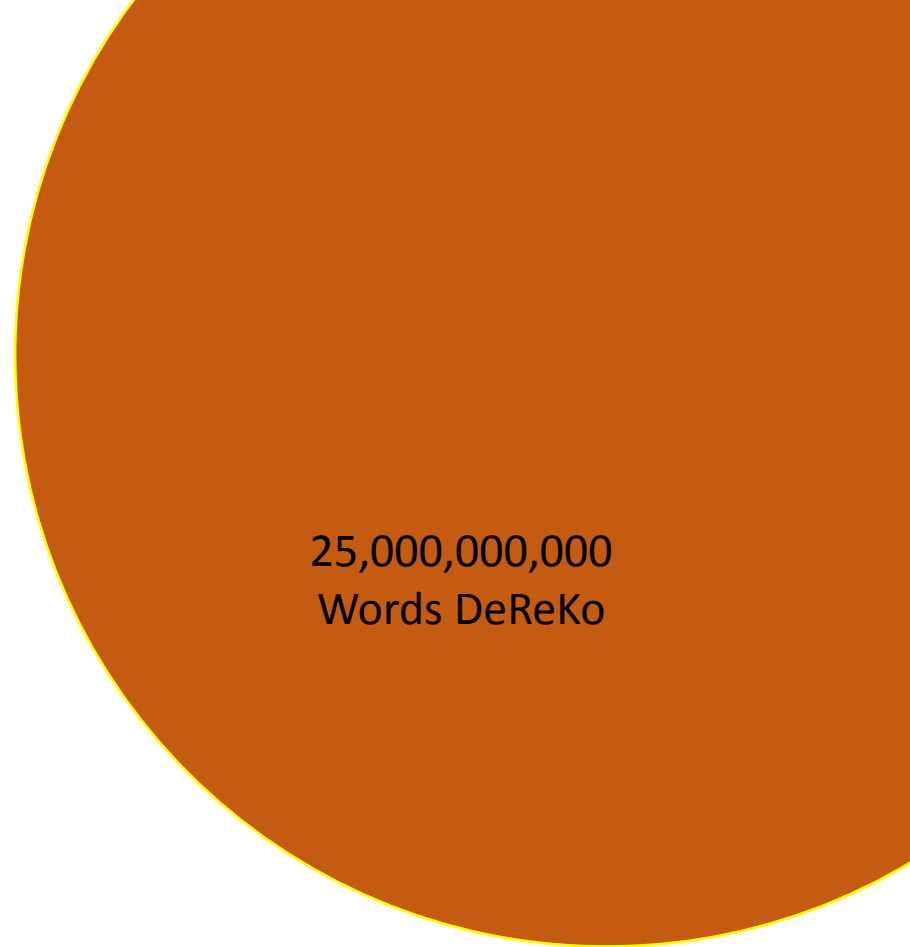
2014

1,000,000
Words FOLK



2015

1,300,000
Words FOLK



- 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)

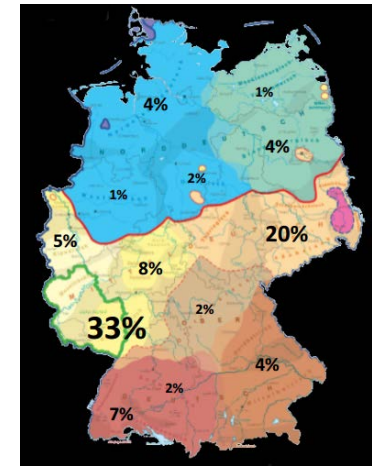
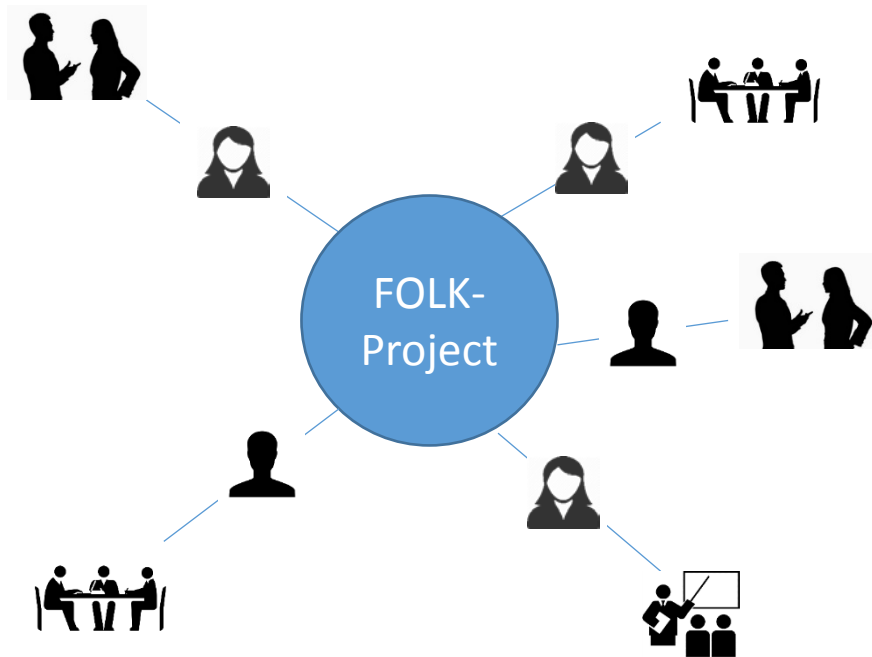
Stratification and field access

- 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



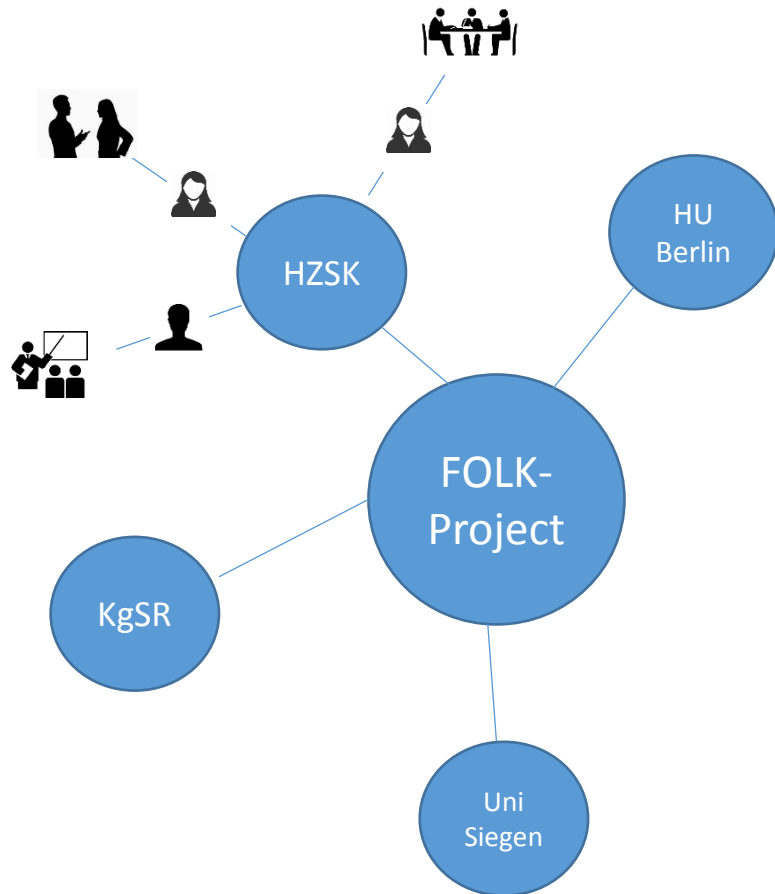
Stratification and field access

- 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 (?)

Corpus processing

FOLK



Private Kommunikation: Paargespräch (FOLK_E_00043)

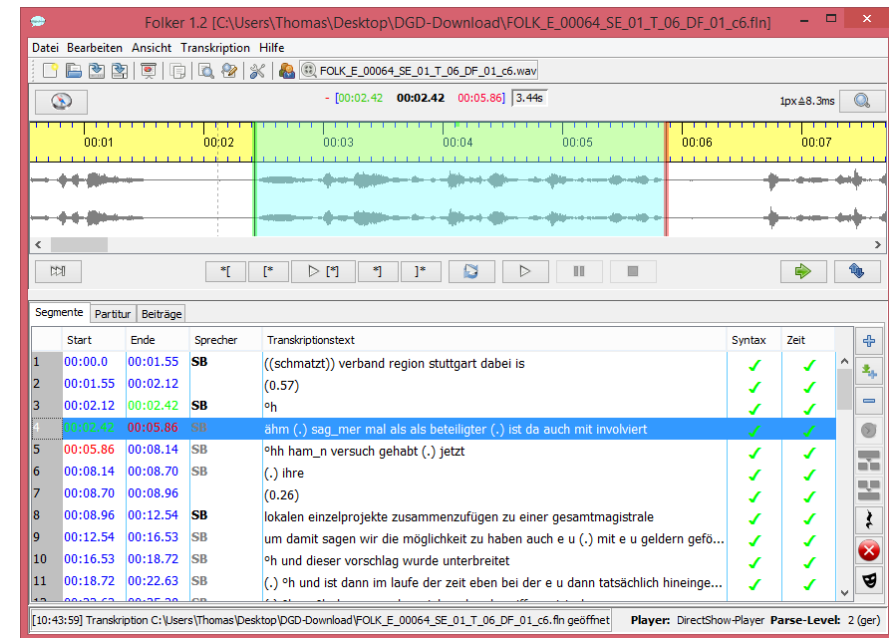
0321 PB "h zum beispiel hat so_n_kommilitone von mir
0322 (1.3)
0323 PB n jetzt den geleast n zet vier () n be em we zet vier
0324 AM () ich weiss doch überhaupt (nich was n zet vier is) () ja es ()
0325 PB [ja des dacht ich mir]
0326 (2.59)
0327 AM des wird irgend n auto sein
0328 PB ja



- „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 transcribing

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!



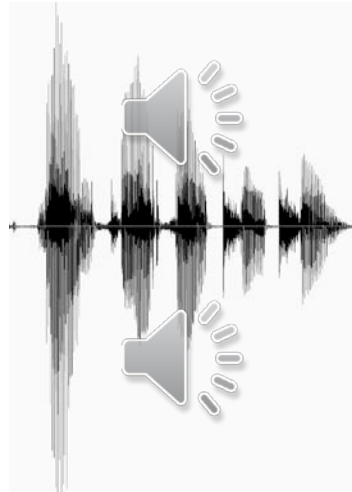
Automatic Speech Recognition

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 automatisation 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



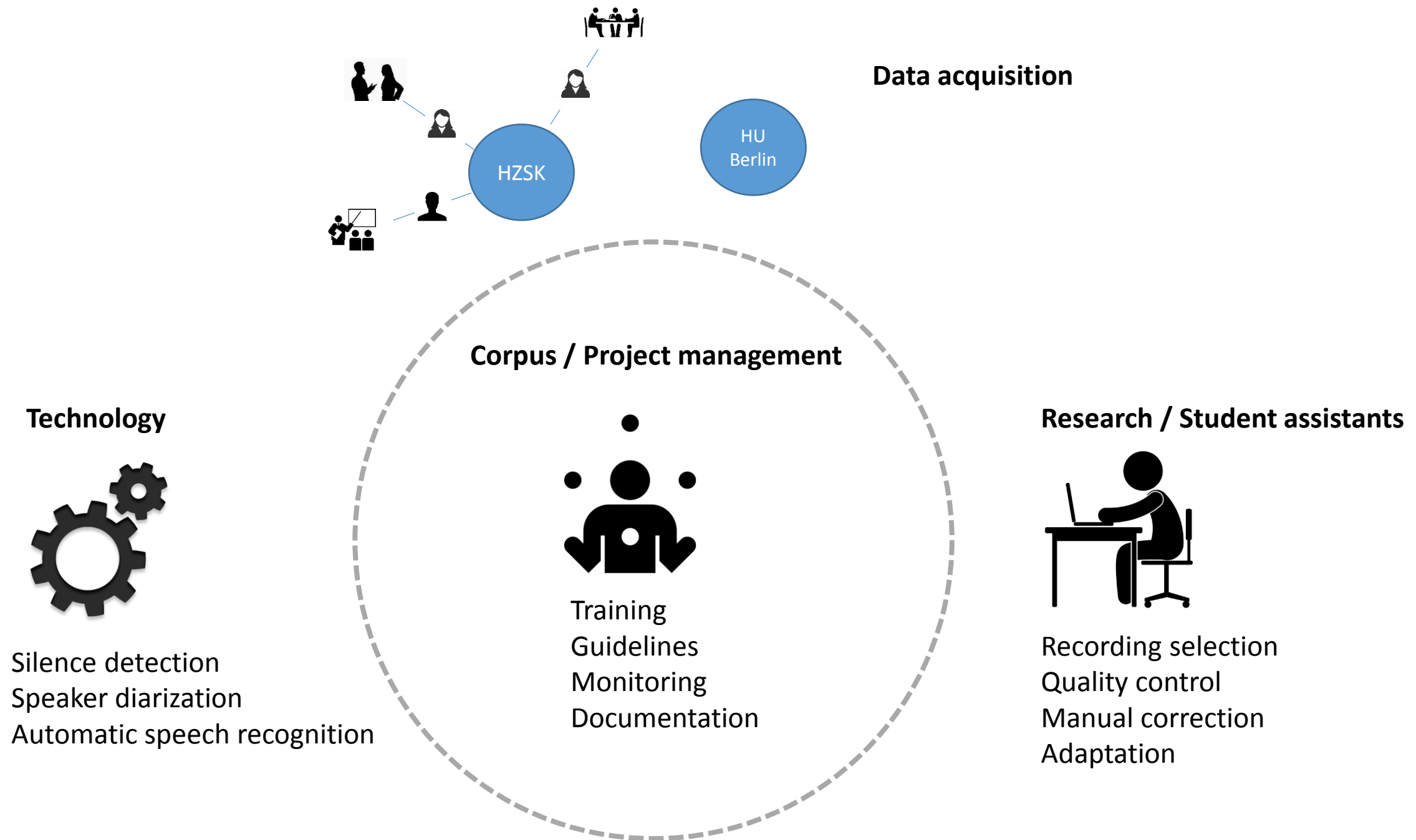
Silence detection
Speaker diarization
Automatic speech recognition



Research / Student assistants



Recording selection
Quality control
Manual correction
Adaptation



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



ESLO 2 (2008-)

CIEL-F (2008)

PCF (2004)

Rhapsodie (2008)

Discours sur la ville (2000)

CLAPI, C-ORAL-ROM, VALIBEL, ...



FOLK (2008-)

GeWiss (2009)

Multilingual Corpora at HZSK (2009)

KidKo (2008)

Texas German (2001)

...



Czech National Corpus (2006-)



GOS Corpus of Spoken Slovene (2009-)



Multimedia Corpus of Spoken Bulgarian (2013)



Reference Corpus of Contemporary Portuguese

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



Faire une requête dans le corpus ?

Si aucun corpus n'est sélectionné, tous les corpus seront interrogés.

Sélection du corpus ?

☐ ESLO1
 ☐ ESLO2

Sélection catégorie(s) ?

Entretien
Contact
Ouverture de l'entretien
Cloture de l'entretien
Repas
Magasin

Entretien
Repas
Interview de personnalités
Conférences
Itinéraire
24heures

Sélection des caractéristiques des locuteurs, des enregistrements, des transcriptions
Pour prendre en compte les critères, vous devez cocher la ligne concernée.

☐ Enregistrement ?
 ☐ Locuteur ?
 ☐ Transcription ?

Recherche d'une occurrence ?

Occurrence
 Conditions
☐ Avec balise d'évènement sonore ?

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 :

60



(Nombre de mots à afficher : 60 max)

☐ 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
Us



Télécharger 20
s



Jouer 4
Us



Télécharger 40
s

Télécharger :

l'enregistrement (mp4)

la transcription (trico)

la transcription (trs)

la transcription (rtf)

ORAL 2013 (Czech National Corpus)

kon text Query Subcorpora Save Concordance Filter Frequency Collocations View Help

Search in the corpus

Corpus:

Query Type:

Query:

Specify context ▼

Specify query according to the meta-information ▲

sp.pohlavi	sp.vek	sp.vzdelani	sp.vzdelanotyp	sp.oblast
☐ M ☐ Y ☐ Z Select all	☐ I ☐ V ☐ Y Select all	☐ A ☐ B ☐ Y Select all	☐ SŠ ☐ VŠ ☐ Y ☐ ZŠ Select all	☐ jihozápadočeská ☐ pohraničí moravské ☐ pohraničí české ☐ severovýchodočeská ☐ slezská ☐ středomoravská ☐ středočeská ☐ východomoravská ☐ Y ☐ česko-moravská Select all

Hits: 26 | i.p.m. : 7.91 (related to the whole oral2013) | ARF : 11.56 | Result is shuffled

☐ 11L002N	hrubé rozpor .] [ale já nevím ---] [hmm] [a Leonardo da Vinci žil kdy] [nebo o kerým to Leonardu bylo] [no
☐ 08A075N	..] [že byste to tam vzali . a ona ti da zpátky oranžové ..] [s tím nesouhlasím] [to on dostal jednu
☐ 09M039N	ať nam obložení ne* nezá* děla barvou] [mmm tak to da papír] [(odkašlání)] [dozadu a to musí ponatírat ..] [jo vidíš
☐ 11L002N	kerým to Leonardu bylo] [no tak to . o Leonardu da Vinci no] [no] [že on] [to tam . nák věděl
☐ 10A049N	Vinci . to bylo asi] [nejlepší] [a jo] [to toho da Vinciho vo tom vona furt mluvila .] [no to bylo]
☐ 10A062N	] [že si dovolenou vezme a bude to lepít jak se da ..] [mezi dvěma] [to říkal .] [jo .] [jo jo
☐ 11A003N	] [že tam opravdu nutně něco potřebuju .] [tak to se da domluvit jako] [zavoláš] [no] [divíte se potřebuju toto to nebo
☐ 11A005N	] [NJ] [mlaskani .] [no .] [(smích)] [---] [jen ať si da ..] [NJ] [nech ho jíst . ježíš (smích) --- ..]
☐ 10H040N	to je .] [no há dě] [jo] [to se tam da taky právě chytout ..] [no to je full há dě
☐ 08A091N	fakt vtipný dobrý žádný] [takový to ta ta dá da ta] [takový jako fakt . neznámý samby .. a je
☐ 11A009N	] [různé plesové maturitní šaty a takové ..] [takové věci a da se to] [jako vyřešit] [prostě .] [hmm .] [jo dá
☐ 11L009N	da je ruský vole ..] [no] [(cizojazyčný projev)] [no] [je německy] [da je aji rumunsky .] [aji rumunsky ? .] [možná aji
☐ 09A155N	hole] [(odmlčení)] [ty níže nesu tak dolamane . tam se da] [(odmlčení)] [počkat až . tak příští rok jako bych někam
☐ 11L009N	Kalima .] [(cizojazyčný projev)] [da ..] [ja] [(smích)] [no proto . da je ruský vole ..] [no] [(cizojazyčný projev)] [no] [je německy] [
☐ 10A005N	ho to odečetlo .] [odečetlo třeba do . teď tam da třeba ujede já nevím deset kilometrů] [tak to tam třeba
☐ 10A103N	ho budou všichni placat po ramenou že mu to něco da .] [neda mu to vůbec nic .. on právě potřebuje
☐ 11A040N	] [no a kde ty peníze kdo kdo ty peníze bance da ?] [tak ti nájemníci ať jich nebo tak] [šak z
☐ 10A169N	.] [no že jo .. no tak možná zítra ti da vědět no ..] [no a tak co ty Marcelé ?
☐ 09M039N	] [se může dat do kuchyně] [stul gauč .] [peřinač se da do kuchyně .. a dvě skříně se do chodby v*
☐ 10A063N	že] [by tam dojel .] [---] [tam se jako] [autem da dojet jo ?] [jo jo .. a že bychom dojeli
☐ 11A003N	tam vzađu .] [tak tam bude takové .] [dětské hřiště da se říct] [p* pisek skluzavky a takové pičoviny] [počkej kde
☐ 11L009N	nemám co .] [sem měl cigára u Kalima .] [(cizojazyčný projev)] [da ..] [ja] [(smích)] [no proto . da je ruský vole
☐ 11A002N	.] [nešlo otačet ne] [klíčem] [no dobře] [a když tam da celou tu vložku novou ? ..] [on říkal že má
☐ 09M039N	to jenom delka se veme a (odkašlání) ..] [to tam da] [(odmlčení)] [uhlí je tam malo už ..] [tak to jaksí
☐ 10A049N	Tutanchámon . a pak tam bylo ještě ten eee] [Leonardo da Vinci . to bylo asi] [nejlepší] [a jo] [to toho
☐ 08A091N	ale fakt vtipný dobrý žádný] [takový to ta ta dá da da ta] [takový jako fakt . neznámý samby .. a

GOS (Corpus of Spoken Slovene)

gos

☒ Iskanje po pogovornem zapisu ☐ Iskanje po standardiziranem zapisu

da

Najdi

Uporabljaš [enostavno iskanje](#)

[Napredno iskanje](#)

je je zdej [1 delno] žuriranja manj ? /// kok se le	da /// no igrata pa tud golf /// [2 smehob] niti ne ne /// ja /// ja	
ja / ja / kako vama gre golf ? ... [ime] baje	da vas clo žena premaga v golfu /// tukaj smo ga postavili	
vas clo žena premaga v golfu /// tukaj smo ga postavili	da bi mal poskusili /// je to res ? /// zaenkrat še ampak	
nehu igrat [2 smehgo] /// ne ne ne [1 neraz] /// aha /// jz bi stavu	da bo [ime] [ime] premagala v golfu stavim da grem peš	
bi stavu da bo [ime] [ime] premagala v golfu stavim	da grem peš od Ljubljane do [1 delno] ne od ne od	
povedali kaj kaj kaj nej naredim in kako nej udarim	da ker sem absolutno netalementirana /// zelo enostavno veri enostavno / tam pa	
enostavno / tam pa v lukno pa nič drugega [2 smehna] /// [ime] glej	da zmagaš /// [1 premor] [prazen] /// tak se vedno delam važnega ampak pol	
to mmm ? /// [2 smehgo] [1 glas] / [2 smehgo] [1 glas] ne [2 smehob] [2 smehob]	da znam ? /// eee ne /// upsala [2 zvok] [2 zvok] videš da ne [2 smehna] [2 smeh	
se ti zdi da znam ? /// eee ne /// upsala [2 zvok] [2 zvok] videš	da ne [2 smehna] [2 smehna] tko de boš mogu it peš ne /// [1 glas] [prazen]	
mogu it peš ne /// [1 glas] [prazen] /// dejte dejte še vi	da vidmo a ste boljši od eee [ime] ali ne [2 smehgo] / zvolte	
bom probavu brezveze bi blo pokvart ne to situacijo /// upam	da je to moj prijatelj [priimek] vido ne [2 smehob] /// v našo družbo	

FOLK (Database for Spoken German)

SUCHE KONTEXT METADATEN ANZEIGE

Wort: Normalisiert: 

Lemma:  POS:

☐ Reguläre Ausdrücke

☒	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)

EXMARaLDA EXAKT 1.3

File Edit View Concordance Columns Rows RegEx Help

Corpora

- MAPTASK
- http://...ected/MAPTASK.coma
- 24 transcriptions
- 3330 segment chains

Word lists

- Word list for MAPTASK
- 1277 types 21433 tokens

Concordances

- Käse
- MAPTASK
- 27 tokens
- 1 types

MAPTASK (27 results)

RegEx (T) Search: Käse

#	S	Communication	Speaker	Left Context	Match	Right Context	Mother tongue[S]
1	☒	MT_270110_Shirin	Sh	tartpunkt aus bitte ähm ((0,9s)) äh bis zu dieser	Käse	/ bis zu diesem Käse ((0,1s)) nach rechts ((0,3s))	Turkish
2	☒	MT_270110_Shirin	Sh	ähm ((0,9s)) äh bis zu dieser Käse/ bis zu diesem	Käse	((0,1s)) nach rechts ((0,3s)) nicht zu weit nach	Turkish
3	☒	MT_091209_Dimitri	Dim	om Start aus äh ziehst du nach rechts äh Richtung	Käse	((0,2s)) gehst ((0,8s)) unter dem Käse vorbei und	Russian, Thai
4	☒	MT_091209_Dimitri	Dim	h Richtung Käse ((0,2s)) gehst ((0,8s)) unter dem	Käse	vorbei und stoppst äh ((1,0s)) unten rechts und z	Russian, Thai
5	☒	MT_051109_Nadira	Nad	((0,4s)) dann gehst du bis zur	Käse		Pashto
6	☒	MT_051109_Nadira	Nad	((0,5s)) bis zur	Käse	dann gehst du ((0,2s)) bis zu den Äpfeln	Pashto
7	☒	MT_280110_Hoa	Hoa	geradeaus geradeaus bis zur	Käse		Vietnamese
8	☒	MT_280110_Hoa	Hoa	((0,4s)) n gehst du an der	Käse	vorbei ((0,7s)) und biegst dann da nach links ((0	Vietnamese
9	☒	MT_050110_Katinka	Kat	tart ((0,2s)) äh waagrecht äh nach rechts an dem	Käse	ein Stückchen vorbei ((0,3s)) so ungefähr/ äh ((0	Russian
10	☒	MT_050110_Katinka	Kat	((0,3s)) Kante des	Käse	s	Russian
11	☒	MT_091209_David	Dav	ken Seite einen angegessenen oder angeschnittenen	Käse	finden	French, German
12	☒	MT_091209_David	Dav	und gehen hinter diesem	Käse	((0,1s)) nach links	French, German
13	☒	MT_281009_Phuong	Phu	s also nach rechts und dann ähm weiter bis zu ein	Käse	/ Stück Käse siehst	Korean
14	☒	MT_281009_Phuong	Phu	rechts und dann ähm weiter bis zu ein Käse/ Stück	Käse	siehst	Korean
15	☒	MT_281009_Phuong	Phu	((0,5s)) und ähm ((0,4s)) ja die	Käse	gehst du vorbei	Korean
16	☒	MT_281009_Phuong	Phu	und dann äh von	Käse	aus ((0,7s)) ähm gehst du geradeaus beziehungsweise	Korean
17	☒	MT_281009_Phuong	Vic	um den König rum würde wieder zu diesem König und	Käse	fahren	French
18	☒	MT_280410_Janis	Jan	bis hinter dem	Käse	((0,5s)) und dann hinter dem Käse links vorbei ((	Greek
19	☒	MT_280410_Janis	Jan	bis hinter dem Käse ((0,5s)) und dann hinter dem	Käse	links vorbei ((0,4s)) (also) nach oben ((0,4s)) z	Greek
20	☒	MT_280410_Janis	Jan	s)) zum Apfel ((4,7s)) ja ((0,4s)) und hinter dem	Käse	((0,5s)) gehst du dann (un') nach oben zu/ bis zu	Greek
21	☒	MT_281009_Tansu	Tan	ch oben ((0,5s)) so ((0,8s)) ähm du gehst ähm die	Käse	vorbei also	Turkish
22	☒	MT_280110_Minh	Minh	echts ((0,4s)) bis zu/ also äh äh ((0,2s)) an dem	Käse	vorbei	Vietnamese
23	☒	MT_180410_Hitomi	Hit		Käse		Japanese
24	☒	MT_180410_Hitomi	Hit		Käse	((0,7s)) dann ((1s)) äh ((1,8s)) äh ((0,7s)) rech	Japanese
25	☒	MT_180410_Hitomi	Eli	ja ((1,2s)) bis zu	Käse		Polish
26	☒	MT_180410_Hitomi	Eli	bis zum	Käse	ja	Polish

Partitur

Kat [v] ((0,3s)) Kante des Käses

Kat [pho]

Kat [disfluency]

Kat [disfluency]

Fer [v]

Fer [hol]

EDIT PHASE EDIT PHASE

Concordance closed.

	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 corpus projects 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 festgelegt war untendrunter vielen dank das sind die fünf 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