

Accéder aux corpus oraux: méthodes et technologies

Qu'est que (se) représenter la parole?

Orléans, 13/01/2023

Thomas Schmidt -- linguisticbits.de -- thomas@linguisticbits.de

Introduction

[We know] far less about how best to support **access to extended sessions of spontaneous speech**. There is also a need for focused assessment of the **needs of specific user groups** that to date have been understudied. Some examples include teachers, scholars in the humanities and social sciences [...]

[Goldman, J. / Renals, S. / Bird, S. / de Jong, F. / Federico, M. / Fleischhauer, C. / Kornbluh, M. / Lamel, L. / Oard, Douglas W. / Stewart, C. / Wright, R. (2005):
Accessing the Spoken Word.
In: International Journal on Digital Libraries, 287-298]

Strictly speaking, a corpus by itself can do nothing at all, being nothing other than a store of used language. **Corpus access software**, however, can re-arrange that store so that observations of various kinds can be made. [...] A corpus does not contain new information about language, but the software **offers us a new perspective on the familiar**.

[Hunston, Susan (2002): *Corpora in Applied Linguistics*. Cambridge University Press]

Projet ZuMult (Accès aux corpus multimodaux ...)

- 2018-2021: Université de Leipzig (Christian Fandrych), Université de Hambourg (Kai Wörner), Institut de la Langue Allemande, Mannheim
 - Collaborateurs: Franziska Wallner, Annette Portmann, Cordula Meißner, Matthias Schwendemann (Leipzig), Elena Frick, Julia Kaiser, Josip Batinić (Mannheim)
 - Basés sur les corpus FOLK (Corpus de l'allemand parlé pour la recherche et l'enseignement) et GeWiss (Corpus multilingue de la langue académique parlée)
 - Basés sur les technologies développées autour de EXMARaLDA* et de la Base de Données de l'Allemand parlé**
 - Avec une attention particulière pour l'enseignement de l'Allemand comme langue étrangère (DaF) / langue seconde (DaZ) et pour la recherche en analyse conversationnelle
- Cette communication:
 - Motivation et arrière-plan des développements
 - Démonstration: Rencontre ZuMult <> ESLO

* <https://www.exmaralda.org>

** <https://dgd.ids-mannheim.de>

Corpus oral

Audio 

Vidéo   

Métadonnées   

Transcriptions   

Corpus oral

Audio 

Vidéo 

Métadonnées 

Transcriptions 

Grand effort



- Accès au terrain
- Enregistrement
- Transcription
- Organisation

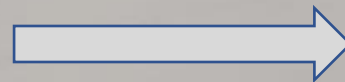
Corpus oral

Audio 

Vidéo 

Métadonnées 

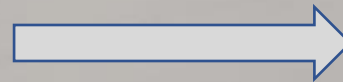
Transcriptions 



(Tiroir de l'oubli)

Corpus oral

Audio	
Vidéo	
Métadonnées	
Transcriptions	



(Tiroir de l'oubli)



Plateforme de corpus

- Téléchargement
- Parcourir les données
- Recherche de corpus
- Optimisé pour un objectif de l'application

Sociolinguistique:
Variation en fonction de
caractéristiques socio-
biographiques



Corpus oral

Audio 

Vidéo 

Métadonnées 

Transcriptions 



?  Traitement automatique
de langue

?  Oral History

?  Enseignement de
langues:
Trésor d'exemples
authentiques

?  Analyse
conversationnelle:
Séquences d'échanges
verbaux

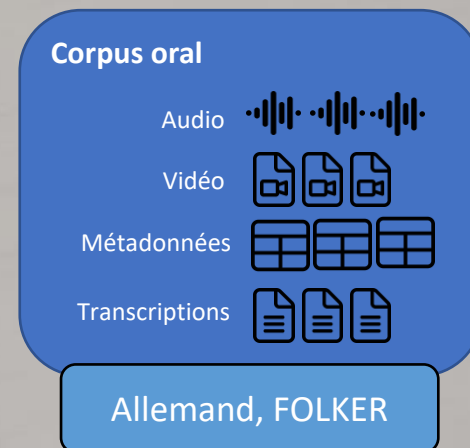
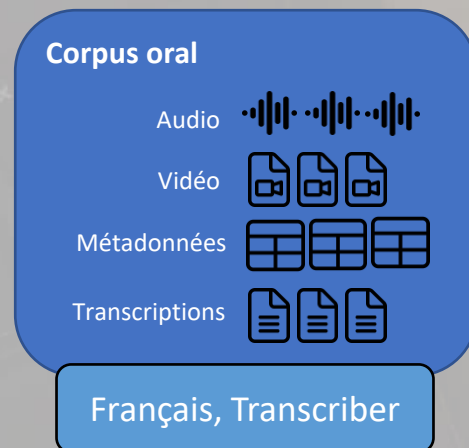
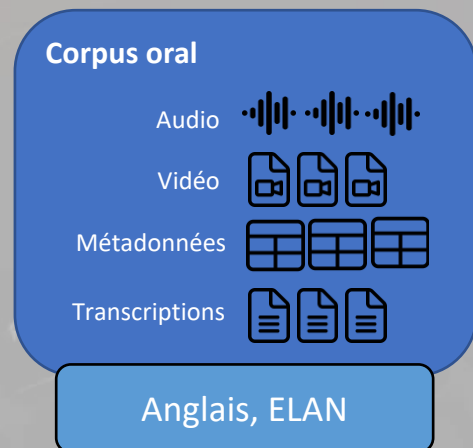
Variation en
perspective /
méthodologie de
recherche /
d'application



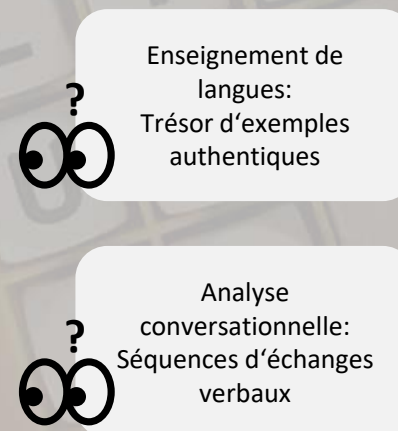
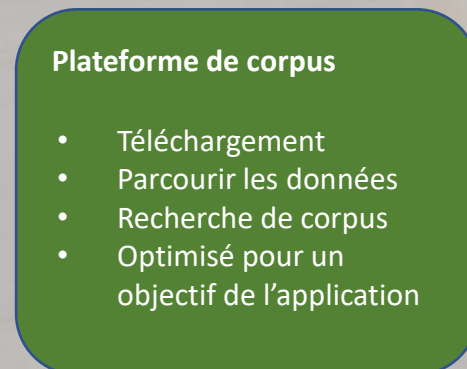
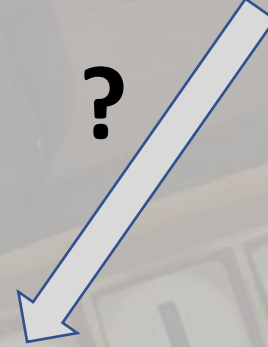
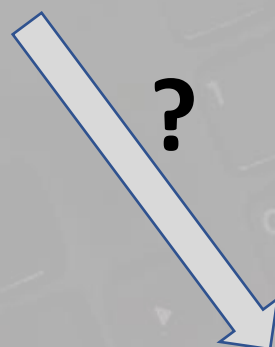
Plateforme de corpus

- Téléchargement
- Parcourir les données
- Recherche de corpus
- Optimisé pour un objectif de l'application

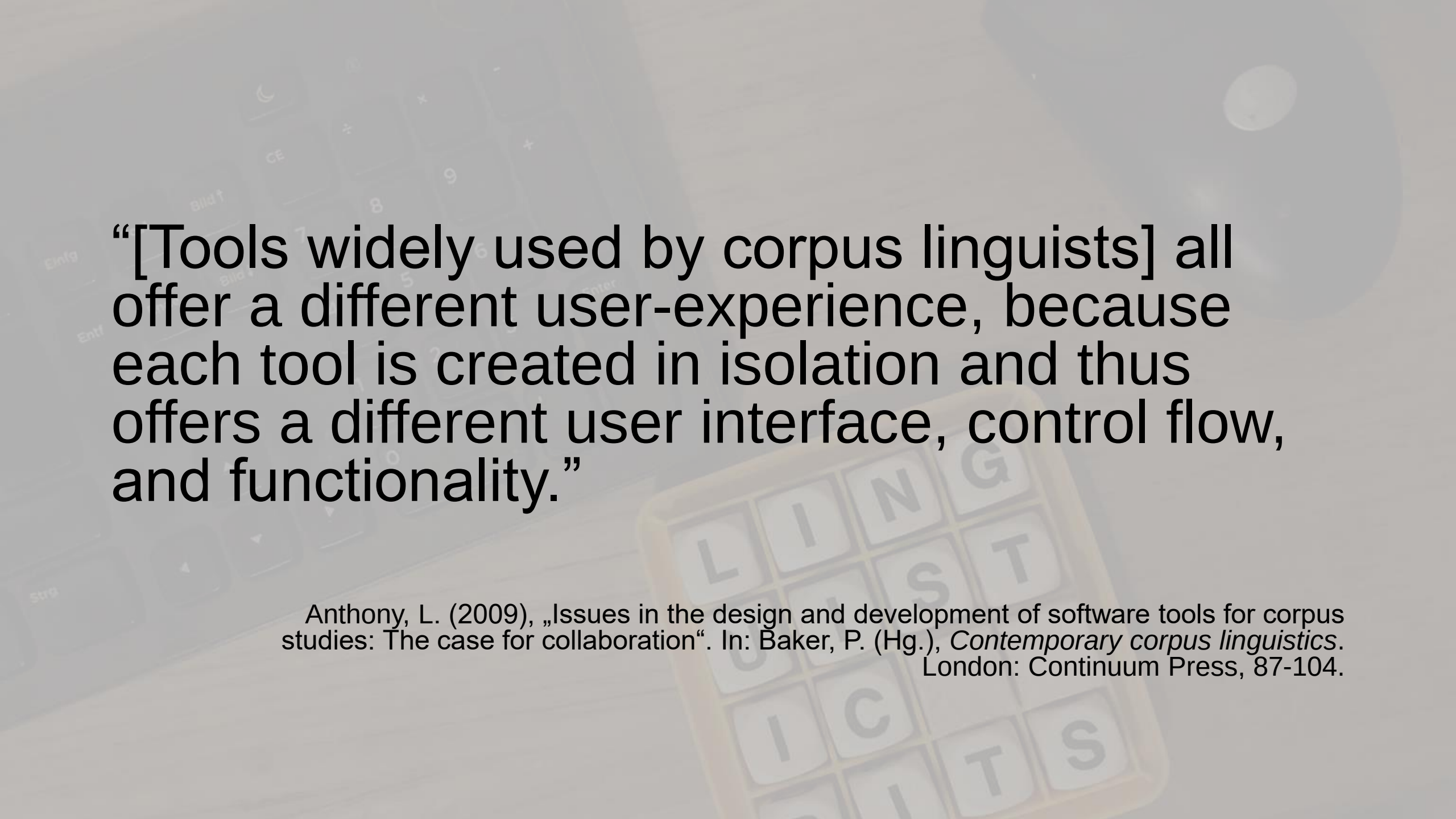
!  Sociolinguistique:
Variation en fonction de
caractéristiques socio-
biographiques



Variation en contenu
et en réalisation
technique

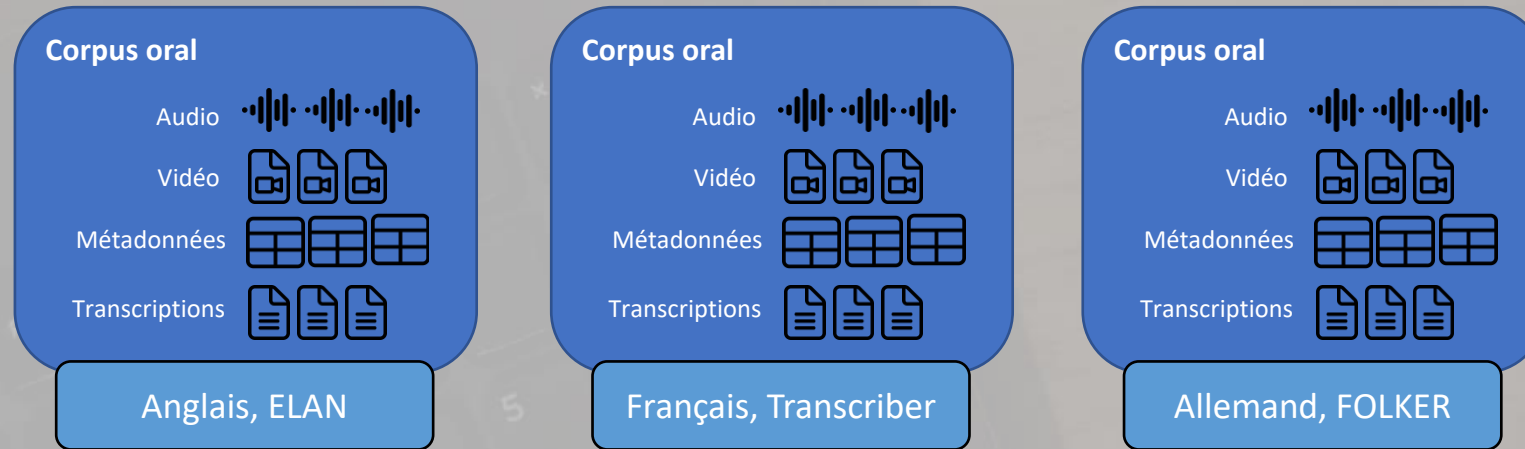


Variation en
perspective /
méthodologie de
recherche /
d'application

The background of the slide is a grayscale image of a computer keyboard. A prominent feature is a custom keycap in the lower right area that reads 'LINGUISTICS' in a grid. The text of the quote is overlaid on the upper left portion of the keyboard image.

“[Tools widely used by corpus linguists] all offer a different user-experience, because each tool is created in isolation and thus offers a different user interface, control flow, and functionality.”

Anthony, L. (2009), „Issues in the design and development of software tools for corpus studies: The case for collaboration“. In: Baker, P. (Hg.), *Contemporary corpus linguistics*. London: Continuum Press, 87-104.



Variation en contenu
et en réalisation
technique

Modèle de données
complet,
manipulation d'objets

!

Sociolinguistique:
Variation en fonction de
caractéristiques socio-
biographiques

Plateforme de corpus

- Téléchargement
- Parcourir les données
- Recherche de corpus
- Optimisé pour un objectif de l'application

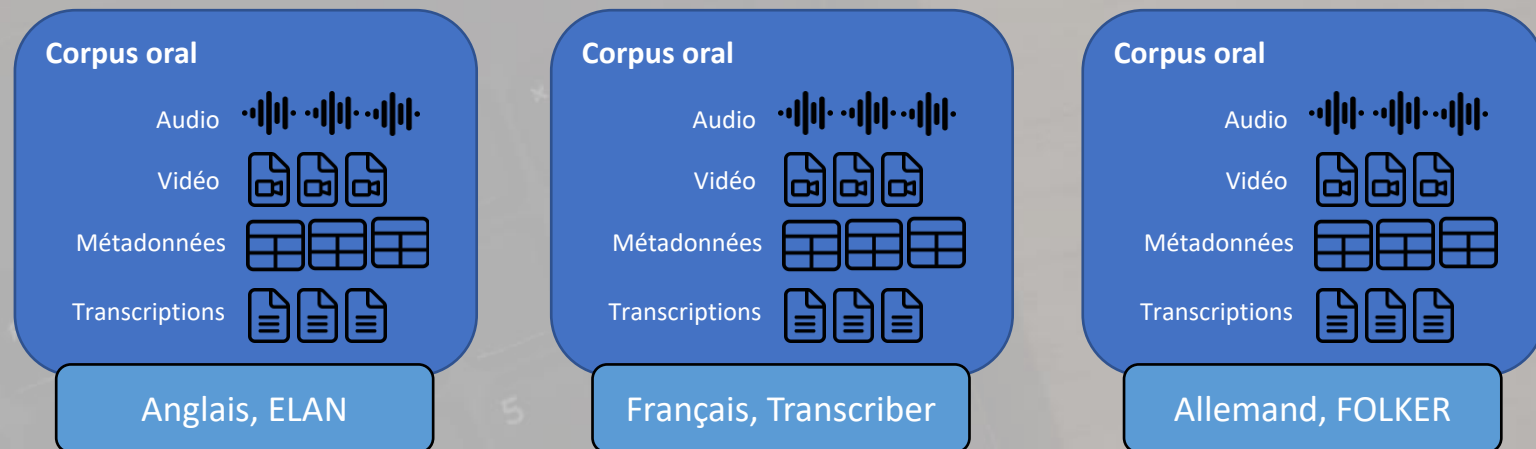
?

Enseignement de
langues:
Trésor d'exemples
authentiques

?

Analyse
conversationnelle:
Séquences d'échanges
verbaux

Variation en
perspective /
méthodologie de
recherche /
d'application



Variation en contenu
et en réalisation
technique

Modèle de données
complet,
manipulation d'objets

Sociolinguistique:
Variation en fonction de
caractéristiques socio-
biographiques

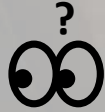
Enseignement de
langues:
Trésor d'exemples
authentiques

Analyse
conversationnelle:
Séquences d'échanges
verbaux

Variation en
perspective /
méthodologie de
recherche /
d'application

Cas d'application

FOLK



- Corpus de référence por l'allemand parlé dans l'interaction
- transcription littéraire, aligné au son, normalisation orthographique, lemmatisation, étiquetage POS
- 300+ heure d'enregistrement audio et vidéo
- Variation en domaine d'interaction, type d'interaction, dialectes des locuteurs, ...
- Accessible par la Base de Données de l'Allemand parlé (DGD)
- Optimisé vers un application en analyse conversationnelle et/ou linguistique de corpus

```
≡ 0010 EL ((lacht)) °hh der wollte uns rich[tig schön uffloof]en la[ssen] ((lacht))
≡ 0011 NO ((lacht))
≡ 0012 NO [°hh ]
≡ 0013 NO [j]ə
≡ 0014 (0.4)
```

Enseignement de Langues

- „Il me faut un bon exemple“*
- ...pour l'usage du marqueur de discours „ich sag mal“
- Pas trop difficile
 - avec un vocabulaire connu par les apprenants de niveau B1
 - où on ne parle pas trop vite
 - avec peu de chevauchements
 - assez proche à la prononciation standard
- Pour montrer et discuter/expliquer un extrait dans mon cours
- TROP DE DONNÉES!

*Fandrych, C. / Schwendemann, M. / Wallner, F. (2021): „Ich brauch da dringend ein passendes Beispiel ...“: Sprachdidaktisch orientierte Zugriffsmöglichkeiten auf Korpora der ge-sprochenen Sprache aus dem Projekt ZuMult. Informationen Deutsch als Fremdsprache, vol. 48, no. 6, 2021, pp. 711-729.

Annotation: FOLK_E_0039 – Conversation de couple

The screenshot displays the FOLKER 1.3 software interface. At the top, the file path is shown: C:\Users\bernd\Dropbox\work\2022_ORLEANS\FOLK_E_00039\FOLK_E_00039_SE_01_T_03_DF_01_c17.flv. Below this is a menu bar with 'Datei', 'Bearbeiten', 'Ansicht', 'Transkription', and 'Hilfe'. A toolbar contains various icons for file operations and playback. The main window features a timeline with a yellow background and a green bar indicating the current segment. Below the timeline is a waveform display. At the bottom, a table lists the segments with columns for 'Start', 'Ende', 'Sprecher', 'Transkriptionstext', 'Syntax', and 'Zeit'.

	Start	Ende	Sprecher	Transkriptionstext	Syntax	Zeit
1	00:00.0	00:00.40		(0.4)	✓	✓
2	00:00.40	00:00.82	EL	oh	✓	✓
3	00:00.82	00:01.17		(0.34)	✓	✓
4	00:01.17	00:02.82	EL	oh tanzt mai euren walzer hier	✓	✓
5	00:02.82	00:03.20	EL	hm hm	✓	✓
6	00:02.82	00:03.20	NO	ja genau	✓	✓
7	00:03.20	00:03.52	EL	((lacht))	✓	✓
8	00:03.52	00:03.85		(0.33)	✓	✓
9	00:03.85	00:04.73	EL	mach det ma ((Lachansatz))	✓	✓

[12:10:58] Transkription C:\Users\bernd\Dropbox\work\2022_ORLEANS\FOLK_E_00039\FOLK_E_00039_SE_01_T_03_DF_01_c17.flv geöffnet IS-Player Parse-Level: 2

1. Transcription selon cGAT minimal en FOLKER

- Transcription littéraire („ouais“ pour „oui“)
- Pausen, comportement non-verbal
- Chevauchements
- Alignement son en segments de 2s à 5s

Annotation: FOLK_E_0039 – Conversation de couple

Transcription	mach	det	ma
(fr.)	<i>faites</i>	<i>le</i>	<i>une-fois</i>
Normalisation	macht	das	mal

1. Transcription selon cGAT minimal en FOLKER
2. Normalisation orthographique en OrthoNormal
 - Automatique avec taux d'erreurs < 20%
 - Correction manuelle

OrthoNormal 1.0 [C:\Users\bernd\Dropbox\work\2022_ORLEANS\FOLK_E_00039\FOLK_E_00039_SE_01_T_03_DF_01_c17.fin]

Datei Bearbeiten Ansicht Annotation Hilfe

[JDS] [P] [N] [S] [M]

00:03.85 00:03.85 00:04.73

	Start	Ende	Sprecher	Transkriptionstext
1	00:00.0	00:00.40		(0.4)
2	00:00.40	00:00.82	EL	oh
3	00:00.82	00:01.17		(0.34)
4	00:01.17	00:03.52	EL	oh tanzt mai [mal] euren walzer [Walzer] hier hm [hmhm]_hm [&] ((lacht))
5	00:02.82	00:03.20	NO	ja genau
6	00:03.52	00:03.85		(0.33)
7	00:03.85	00:04.73	EL	mach [macht] det [das] ma [mal] ((Lachansatz))

mach[t] det [das] ma[l]

Korrekturen

Beitrag bearbeiten... Sprecherzuordnung ändern... Sprecherkürzel ändern...

ab|cd **OOV** %

Modus

☒ Wort ☐ Normal ☐ Lemma ☐ POS

	Wort	Normal	Lemma
1	oh		oh
2	tanzt		tanzen
3	mal	mal	mal
4	euren		euer
5	walzer	Walzer	Walzer
6	hier		hier
7	hm	hmhm	hmhm
8	hm	&	&
9	ja		ja
10	genau		genau
11	mach	macht	machen
12	det	das	die
13	ma	mal	mal

[12:18:19] Transkription C:\Users\bernd\Dropbox\work\2022_ORLEANS\FOLK_E_00039\FOLK_E_00039_SE_01_T_03_DF_01_c17.fin geöffnet.

Enrichissement: FOLK_E_0039 – Conversation de couple

Transcription	mach	det	ma
(fr.)	<i>faites</i>	<i>le</i>	<i>une-fois</i>
Normalisation	macht	das	mal
Lemma	machen	die	mal
POS	VVFIN	PDS	PTKMA

1. Transcription selon cGAT minimal en FOLKER
2. Normalisation orthographique en OrthoNormal
3. Lemmatisation et étiquetage POS avec TreeTagger
 - STTS 2.0 : Tagset et paramétrisation optimisés pour la langue parlée en interaction
 - Taux d'erreurs < 2% (lemmatisation) et <5% (POS)
 - Pas de correction manuelle

Enrichissement: FOLK_E_0039 – Conversation de couple

Transcription	mach	det	ma
(fr.)	<i>faites</i>	<i>le</i>	<i>une-fois</i>
Normalisation	macht	das	mal
Lemma	machen	die	mal
POS	VVFIN	PDS	PTKMA
Phonétique	[max]	[det]	[ma]

1. Transcription selon cGAT minimal en FOLKER
2. Normalisation orthographique en OrthoNormal
3. Lemmatisation et étiquetage POS avec TreeTagger
4. Annotation phonétique avec G2P
 - automatiquement par un service web de l'Archive Bavarien pour le signal vocal (BAS)
 - à partir de la transcription littéraire (approximation à la prononciation réelle)
 - avec syllabification des mots multi-syllabique

Enrichissement: FOLK_E_0039 – Conversation de couple

Transcription	mach	det	ma
(fr.)	<i>faites</i>	<i>le</i>	<i>une-fois</i>
Normalisation	macht	das	mal
Lemma	machen	die	mal
POS	VVFIN	PDS	PTKMA
Phonétique	[max]	[det]	[ma]
Vitesse	3.41 syll/s		

1. Transcription selon cGAT minimal en FOLKER
2. Normalisation orthographique en OrthoNormal
3. Lemmatisation et étiquetage POS avec TreeTagger
4. Annotation phonétique avec G2P
5. Calcul de la vitesse d'articulation
 - pour les contributions entières
 - syllabes par unités de temps (secondes)

Enrichissement: FOLK_E_0039 – Conversation de couple

Transcription	mach	det	ma
(fr.)	<i>faites</i>	<i>le</i>	<i>une-fois</i>
Normalisation	macht	das	mal
Lemma	machen	die	mal
POS	VVFIN	PDS	PTKMA
Phonétique	[max]	[det]	[ma]
Vitesse	3.41 syll/s		
Taux de norm.	100%		

1. Transcription selon cGAT minimal en FOLKER
2. Normalisation orthographique en OrthoNormal
3. Lemmatisation et étiquetage POS avec TreeTagger
4. Annotation phonétique avec G2P
5. Calcul de la vitesse d'articulation
6. Calcul du taux de normalisation
 - Proportion de mots où la forme transcrite n'égale pas la forme normalisée
 - Indicateur de dialecticité et d'informalité

Enrichissement: FOLK_E_0039 – Conversation de couple

0017 EL oh tanzt mal euren Walzer hier [hmhm]((lacht))
0018 NO [ja genau]
0019 (0.33)
0020 EL macht das mal ((Lachansatz))

G	
✓ ganz	1
✓ geben	1
✓ Geld	1
genau	1
H	
halt	1
✓ hier	1
hmhm	1
J	
✓ ja	4
K	

S	
✓ schön	1
✓ so	1
✓ suchen	2
T	
✓ tanzen	1
U	
✓ und	2
W	
Walzer	1
✓ wir	6
✓ wollen	1

1. Transcription selon cGAT minimal en FOLKER
2. Normalisation orthographique en OrthoNormal
3. Lemmatisation et étiquetage POS avec TreeTagger
4. Annotation phonétique avec G2P
5. Calcul de la vitesse d'articulation
6. Calcul du taux de normalisation
7. Calcul de mesures pour l'ensemble de la transcription
 - Vitesse d'articulation moyenne
 - Taux de normalisation total
 - Comparaison avec des listes de vocabulaire – indicateur de difficulté lexicale
 - Taux de chevauchement – indicateur d'interactivité
 - Fréquence relative de „mots de discours“ – indicateur d'interactivité, d'informalité, de difficulté lexicale

Encodage standardisé: ISO/TEI

```
58 <text xml:lang="de">
59   <timeline unit="s">
60     <when interval="3016.112" since="T_START" xml:id="T_START"/>
61     <when interval="3016.221" since="T_START" xml:id="TLI_21"/>
62     <when interval="3016.646" since="T_START" xml:id="TLI_22"/>
63     <when interval="3016.991" since="T_START" xml:id="TLI_23"/>
64     <when interval="3018.646" since="T_START" xml:id="TLI_24"/>
65     <when interval="3019.026" since="T_START" xml:id="TLI_25"/>
66     <when interval="3019.346" since="T_START" xml:id="TLI_26"/>
67     <when interval="3019.676" since="T_START" xml:id="TLI_27"/>
68     <when interval="3020.556" since="T_START" xml:id="TLI_28"/>
69     <when interval="3025.046" since="T_START" xml:id="TLI_29"/>
70   </timeline>
71   <body>
72     <annotationBlock end="TLI_22" start="TLI_21" who="EL" xml:id="c15"> [8 lines]
81     <pause dur="PT0.34S" end="TLI_23" rend="(0.34)" start="TLI_22" xml:id="p9"/>
82     <annotationBlock end="TLI_26" start="TLI_23" who="EL" xml:id="c17"> [19 lines]
102    <annotationBlock end="TLI_25" start="TLI_24" who="NO" xml:id="c18"> [9 lines]
112    <pause dur="PT0.33S" end="TLI_27" rend="(0.33)" start="TLI_26" xml:id="p10"/>
113    <annotationBlock end="TLI_28" start="TLI_27" who="EL" xml:id="c20">
114      <u xml:id="u_d337e1032">
115        <seg type="contribution" xml:id="seg_d337e1032">
116          <anchor synch="TLI_27"/>
117          <w lemma="machen" norm="macht" phon="max" pos="VVFIN" xml:id="w41">mach</w>
118          <w lemma="die" norm="das" phon="dst" pos="PDS" xml:id="w42">det</w>
119          <w lemma="mal" norm="mal" phon="ma" pos="PTKMA" xml:id="w43">ma</w>
120          <incident xml:id="n7">
121            <desc rend="(Lachansatz)">Lachansatz</desc>
122          </incident>
123          <anchor synch="TLI_28"/>
124        </seg>
125      </u>
126      <spanGrp subtype="time-based" type="speech-rate">
127        <span from="TLI_27" to="TLI_28">3.41</span>
128      </spanGrp>
129    </annotationBlock>
130    <pause dur="PT4.49S" end="TLI_29" rend="(4.49)" start="TLI_28" xml:id="p11"/>
131  </body>
132 </text>
133 </TEI>
```

- ISO 24624:2016 Language resource management — Transcription of spoken language
- Toutes les informations lisible par machine
- Basé sur les Recommandations pour l'encodage et l'échange de textes électroniques (TEI)
- Reliés à d'autres standards ISO de la famille "Gestion des ressources linguistiques"
- Interoperable avec ELAN, Transcriber, Praat, EXMARaLDA, FOLKER
- De plus en plus utilisé par les archives et les infrastructures digitales
- Autres outils: WebAnno/INCEpTION, WebLicht, BAS Services

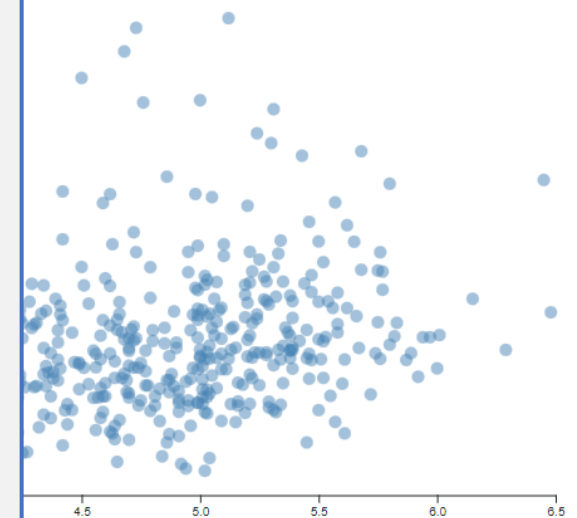
Korpus: FOLK Reset all

Gesprächstyp	?
Art	?
Themen	?
Sprachregion	?
Dauer	?
Wortschatz (Niveaustufe)	?
Standardnähe (Normalisierungsrate)	?
Sprechgeschwindigkeit (Artikulationsrate)	?
Überlappungen	?
Wortarten	?
Mündlichkeitsphänomene	?

Y Achse: Normalisierungsrate X Achse: Artikulationsrate

Enseignement de Langues

- „Il me faut un bon exemple“
- ...pour l'usage du marqueur de discours „ich sag mal“
- Pas trop difficile
 - avec un vocabulaire connu par les apprenants de niveau B1
 - où on ne parle pas trop vite
 - avec peu de chevauchements
 - assez proche à la prononciation standard
- Pour montrer et discuter/expliquer un extrait dans mon cours



Dauer	Normalisierungsrate	Zu Viel Link
0:59:45	22%	→ZuViel
0:31:01	7%	→ZuViel
0:20:29	9%	→ZuViel
1:26:47	23%	→ZuViel
1:15:31	28%	→ZuViel
0:41:39	26%	→ZuViel
1:28:23	21%	→ZuViel
0:56:14	23%	→ZuViel
0:43:04	26%	→ZuViel
0:47:56	17%	→ZuViel
3:16:53	17%	→ZuViel
0:40:03	20%	→ZuViel
0:38:21	9%	→ZuViel

FOLK_E_00008_SE_01	Unterrichtsstunde in der Berufsschule
FOLK_E_00009_SE_01	Unterrichtsstunde in der Berufsschule
FOLK_E_00010_SE_01	Spielinteraktion mit Kindern
FOLK_E_00011_SE_01	Spielinteraktion mit Kindern
FOLK_E_00012_SE_01	Spielinteraktion mit Kindern
FOLK_E_00013_SE_01	Vorlesen für Kinder

Korpus: FOLK Reset all

Gesprächstyp ?

Art ?

Themen ?

Sprachregion ?

Dauer ?

Wortschatz (Niveaustufe) ?

Standardnähe (Normalisierungsrate) ?

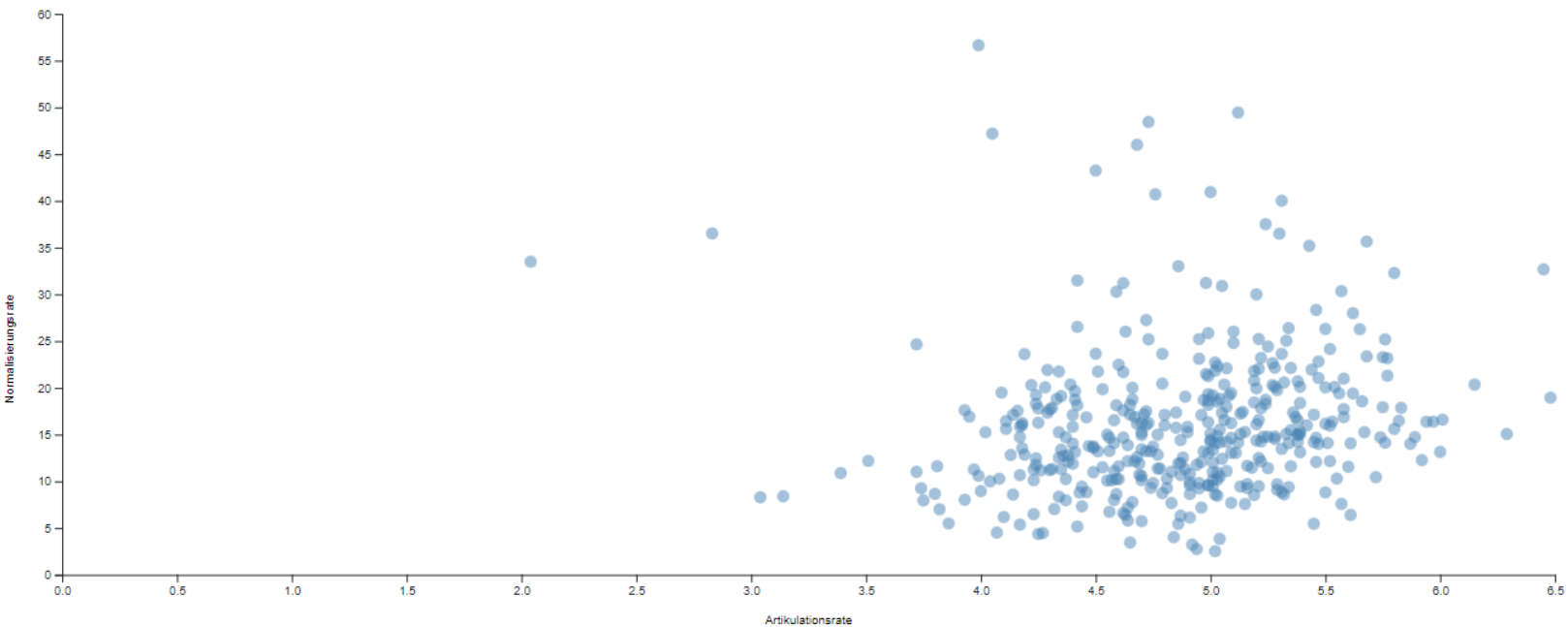
Sprechgeschwindigkeit (Artikulationsrate) ?

Überlappungen ?

Wortarten ?

Mündlichkeitsphänomene ?

Y Achse: Normalisierungsrate X Achse: Artikulationsrate



Spalten auswählen ⚙

Id	Art	Dauer	Normalisierungsrate	Zu Viel Link
FOLK_E_00001_SE_01	Unterrichtsstunde in der Berufsschule	0:59:45	22%	→ZuViel
FOLK_E_00002_SE_01	Vorlesen für Kinder	0:31:01	7%	→ZuViel
FOLK_E_00003_SE_01	Prüfungsgespräch in der Hochschule	0:20:29	9%	→ZuViel
FOLK_E_00004_SE_01	Unterrichtsstunde in der Berufsschule	1:26:47	23%	→ZuViel
FOLK_E_00005_SE_01	Unterrichtsstunde in der Berufsschule	1:15:31	28%	→ZuViel
FOLK_E_00006_SE_01	Unterrichtsstunde in der Berufsschule	0:41:39	26%	→ZuViel
FOLK_E_00007_SE_01	Unterrichtsstunde in der Berufsschule	1:28:23	21%	→ZuViel
FOLK_E_00008_SE_01	Unterrichtsstunde in der Berufsschule	0:56:14	23%	→ZuViel
FOLK_E_00009_SE_01	Unterrichtsstunde in der Berufsschule	0:43:04	26%	→ZuViel
FOLK_E_00010_SE_01	Spielinteraktion mit Kindern	0:47:56	17%	→ZuViel
FOLK_E_00011_SE_01	Spielinteraktion mit Kindern	3:16:53	17%	→ZuViel
FOLK_E_00012_SE_01	Spielinteraktion mit Kindern	0:40:03	20%	→ZuViel
FOLK_E_00013_SE_01	Vorlesen für Kinder	0:38:21	9%	→ZuViel

Korpus: FOLK ↕ Reset all

Gesprächstyp ?

Art ?

Themen ?

Sprachregion ?

Dauer ?

Wortschatz (Niveaustufe) ?

Wortschatzliste: GOETHE_B1 ↕

Deckung in %: 80 - 100 80

niedrig ↔ hoch

Reset

Standardnähe (Normalisierungsrate) ?

Anteil normalisierter Token in %: 5 - 13

Niedriger Anteil ↔

Reset

Sprechgeschwindigkeit (Artikulationsrate) ?

Anzahl artikulierter Silben pro Sekunde: 2 - 4

niedrig ↔ hoch

Reset

Überlappungen ?

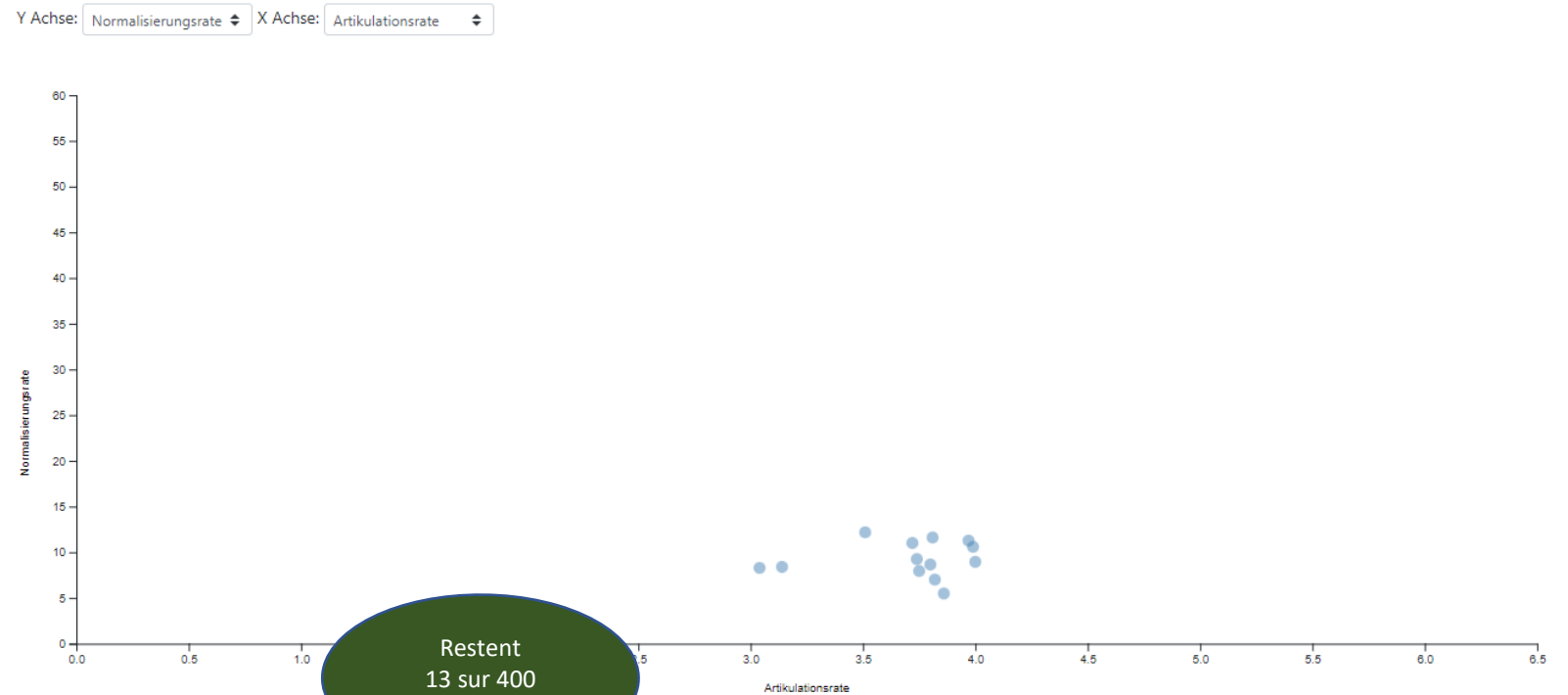
Wortarten ?

Mündlichkeitsphänomene ?

Avec un vocabulaire connu par les apprenants de niveau B1

Assez proche à la prononciation standard

Où on ne parle pas trop vite



Spalten auswählen ⚙

Id	Art	Dauer	Normalisierungsrate	Zu Viel Link
FOLK_E_00013_SE_01	Vorlesen für Kinder	0:38:21	9%	→ZuViel
FOLK_E_00014_SE_01	Vorlesen für Kinder	0:31:11	12%	→ZuViel
FOLK_E_00016_SE_01	Vorlesen für Kinder	0:43:02	7%	→ZuViel
FOLK_E_00017_SE_01	Vorlesen für Kinder	0:14:16	11%	→ZuViel
FOLK_E_00185_SE_01	Sprachbiografisches Interview	0:42:04	9%	→ZuViel
FOLK_E_00364_SE_01	Beratungsgespräch Studienbeihilfe	0:08:17	11%	→ZuViel
FOLK_E_00379_SE_01	Gitarrenunterricht	0:32:14	8%	→ZuViel
FOLK_E_00395_SE_01	Telefongespräch	0:25:57	5%	→ZuViel
FOLK_E_00413_SE_01	Sprachförderung in der KiTa	0:14:08	8%	→ZuViel
FOLK_E_00421_SE_01	Telefongespräch	0:23:01	12%	→ZuViel
FOLK_E_00430_SE_01	Telefongespräch	0:45:11	9%	→ZuViel
FOLK_E_00431_SE_01	Telefongespräch	0:26:24	8%	→ZuViel
FOLK_E_00438_SE_01	Telefongespräch	1:20:08	11%	→ZuViel

Korpora

- ☐ DH
☒ FOLK
☐ GWSS
☐ MEND
☐ DNAM
☐ BETV
☐ HMAT

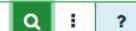
CQP-Suche

Lemma- und Suchanfragelisten

Wiederholungen

Suche mittels der Corpus Query Language (CQP)

Bitte geben Sie Ihren CQP-Suchausdruck ein und wählen Sie den Suchmodus (Transkript- vs. Sprecher-basiert)

Suche ☒ Satzzeichen ignorierenmarqueur de
discours
„ich sag mal“[Beispiele](#)[XML](#)

Ergebnisse

für die Suchanfrage `[norm="ich"]([lemma="sagen"]([norm="ich"]?)[norm="(ein=)?mal"])` (in FOLK)

Treffer gruppieren

Metadatenansicht

Download der KWIC

Insgesamt: 173

Erste Seite Vorherige Seite 1 2 3 4 5 Nächste Seite Letzte Seite

1	FOLK_E_00039_SE_01_T_02	NO	... n kopf jejang (.) ich sach ma (.) wenn de wörklich	DGD	ZuViel
2	FOLK_E_00040_SE_01_T_02	JM	nochma tausend nei °h ich sag mo check (0.41) check (0.64) °hh (2.39) °hhh (0.44) isch ...	DGD	ZuViel
3	FOLK_E_00020_SE_01_T_01	HM	... net irgendwie °hhh ha ich sag mal friedrich ist praktisch ...	DGD	ZuViel
4	FOLK_E_00021_SE_01_T_01	SK	aber wa s ich sach ma özil özil sechs ...	DGD	ZuViel
5	FOLK_E_00001_SE_01_T_01	RG	ma dann ja (0.4) isch sach ma die geberspannung dann ...	DGD	ZuViel
6	FOLK_E_00004_SE_01_T_01	GS	... wollte ma ja ich sach ma unabhängig von fachkompetenze ...	DGD	ZuViel
7	FOLK_E_00007_SE_01_T_01	JK	halt ziemlich (0.36) ins ((schneuzt sich)) ich sag mal (0.41) es war halt ...	DGD	ZuViel
8	FOLK_E_00007_SE_01_T_01	RZ	äh ((hustet)) ja (0.34) grundgleichung ich sag mal so ja wird ...	DGD	ZuViel
9	FOLK_E_00007_SE_01_T_01	GS	ne (0.22) fähigkeiten haben (.) ich sach ma so (0.3) geh weg ...	DGD	ZuViel

137 occurrences
pour le terme de
recherche CQP

- Moteur de recherche MTAS, basé sur Lucene, avec support pour le format ISO/TEI et la langue de recherche CQP
- Recherches simples sont simples, recherches complexes sont possibles (et complex...)
- Recherche CQP > Concordance KWIC – mais aussi utilisable pour d'autres formes de recherche (e.g. liste de mots)

FOLK Unterrichtsstunde in der Berufsschule FOLK_E_00007_SE_01_T_01 Parameter Sprechereignis (2) Auswahl

Ref Wortliste None

A

als 1

B

bekannt 1

Beruf 1

D

d 1

E

eigentlich

erlernen

es

G

gar 1

gehen 1

H

halt 2

0205 JK (.) n der beruf (.) zu erlernen war uns eigentlich als ziel so bekannt und es war gar nich so °h es ging halt ziemlich

0206 (0.36)

0207 JK ins

0208 XM ((schneuzt sich))

0209 JK ich sag mal

0210 (0.41)

0211 JK es war halt viel umfangreicher

0212 (0.26)

0213 GS hmhm

0214 (0.55)

0215 GS äh

Pour montrer et discuter/expliciter un extrait dans mon cours

Transcription

« Density Viewer »

Lecteur audi

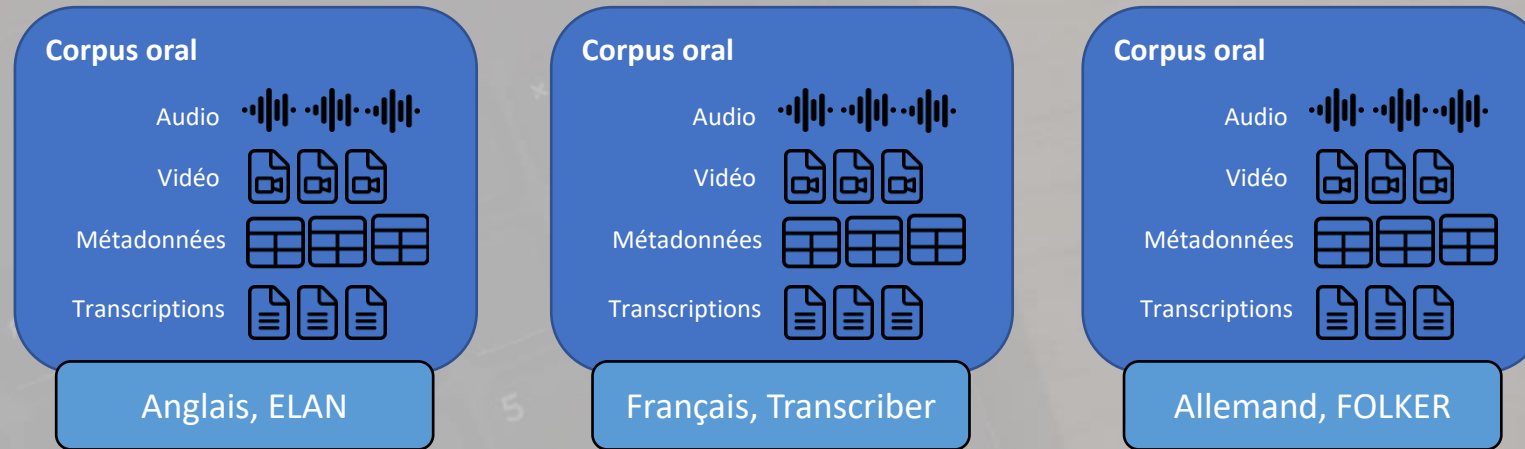
04:12 04:22

0:00 / 1:28:23

pos=ART

Startseite Projekt Help Deutsch

- ZuViel: Visualisation de transcriptions, interaction avec la transcription
- Synchronisation bidirectionnelle entre audio/vidéo, transcription, « Density viewer »
- Configurable, p.ex.:
 - afficher formes transcrites vs. formes normalisées / des sous-titres pour la vidéo
 - mettre en évidence les formes présentes dans une liste de mots
 - Ralentir/accélérer la lecture de l'audio/du vidéo
- Définir et télécharger un extrait en différents formats



Variation en contenu
et en réalisation
technique

Modèle de données
complet,
manipulation d'objets

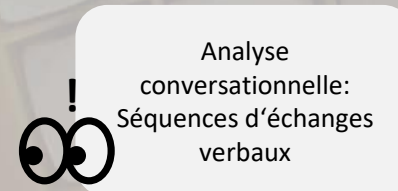
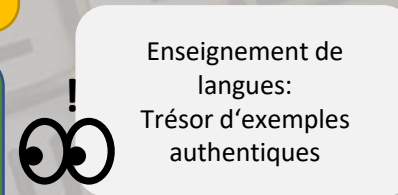
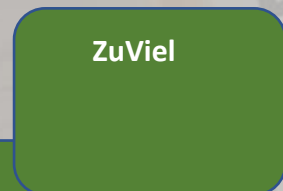
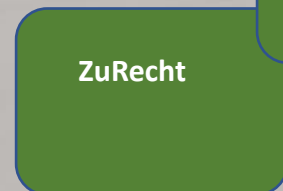
Variation en
perspective /
méthodologie de
recherche /
d'application



Variation en contenu
et en réalisation
technique



Modèle de données
complet,
manipulation d'objets



Variation en
perspective /
méthodologie de
recherche /
d'application

Préparation d'un corpus de rencontre ESLO - ZuMult

Trois entretiens d'ESLO: ESLO_ENT_1001, 1005 et 1013

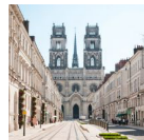
- Conversion Transcriber > ISO/TEI
- Tokenisation – mots et non-mots (pauses etc.)
- Annotation
 - Normalisation – de façon exemplaire: ouais > oui et quelques autres formes très fréquentes
 - Lemmatisation / POS-Tagging avec TreeTagger et le tagset développé par le projet « Traitement de Corpus Oraux en Français » (TCOF)
 - Annotation phonétique par G2P pour le français
- Calcul de la vitesse d'articulation
- Métadonnées dans un fichier COMA (EXMARaLDA CorpusManager)
- Instance de ZuMult local sur mon ordinateur

Demonstration: ESLO dans ZuMult

Enquêtes SocioLinguistiques à Orléans (ESLO)

Demo ZuMult

ZuViel: Visualisation de transcriptions



ESLO_ENT_1001

Discussion en face à face entre un chercheur et un locuteur témoin à partir d'une trame d'entretien. Enregistrée le 20/01/2010 par Olivier Baude.

Transcription



ESLO_ENT_1005

Discussion en face à face entre un chercheur et une locutrice témoin à partir d'une trame d'entretien. Enregistrée le 01/03/2010 par Olivier Baude.

Transcription



ESLO_ENT_1013

Discussion en face à face entre une chercheuse et un locuteur témoin à partir d'une trame d'entretien. Enregistrée le 01/03/2010 par Pauline Philardeau.

Transcription

ZuRecht: Recherche CQP

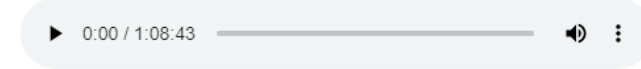
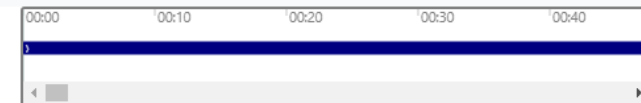
Saisissez une expression de recherche CQP

Mode recherche

Expression CQP, p.ex. [norm="je"][norm="cherche"]



<http://localhost:8080/DGDRESTTest/eslo-demo/eslo-demo.jsp>



Ref Wortliste None

adresse	1
adresser	1
adulte	1
adéquat	1
Afpa	3
africain	1
agacer	1
agencement	1
ah	115
aider	2
ailleurs	2
aimer	22
aise	1
Ajaccio	11
algue	1
alimentation	1



- 0001 **ch_OB1** ((nontrans)) alors
- 0002 **BV1** marche bien au moins votre truc euh
- 0003 **ch_OB1** là ?(0.54) ouais ça tourne
- 0004 **BV1** d'accord
- 0005 **ch_OB1** ça marche ouais ouais qu -qu 'on je vois bien l 'inquiétude qu 'on soit pas obligé de refaire ça euh tous les jours quoi
- 0006 **BV1** non c'est pas ça c'est parce que vous vouliez tester donc euh
- 0007 **ch_OB1** ouais ouais ben là là quand vous parlez c'est impeccable
- 0008 **BV1** Ok
- 0009 **ch_OB1** en termes de de de niveau d'enregistrement non mais c'est pas mal fait ces petits (0.56) ces petits choses -là ça m'aide bien sinon euh (0.90) sinon je pourrais me tromper (1.03) ((bb)) je vais commencer peut-être tout tout simplement sur euh sur une question sur euh bah depuis quand vous habitez euh là dans l -dans le quartier par exemple
- 0010 **BV1** (0.37) ici ça fait euh (0.74) un an un peu plus d -entre un an et un an et demi
- 0011 **BV1** avant j'étais euh un peu plus bas
- 0012 **ch_OB1** oui et euh et à Orléans c'est v -
- 0013 **BV1** Orléans ça fait deux ans et demi un truc comme ça (0.41) ouais deux ans et demi
- 0014 **ch_OB1** d'accord
- 0015 **ch_OB1** et avant Orléans ?
- 0016 **BV1** Blois enfin à côté de Blois
- 0017 **ch_OB1** ah ouais (0.77) Blois euh là c'est euh et avant Blois c'est que vous êtes né de né à Blois ou ah de Blois ou de Blois ou de près de Blois ?
- 0018 **BV1** non je suis natif de Blois
- 0019 **BV1** hm na -bah natif de Blois après j'ai été euh (0.81) à je sais plus à vingt kilomètres de Blois et après j'ai toujours habité à (0.60) à Blois euh enfin (0.41) s -à dix minutes de Blois sud quoi
- 0020 **ch_OB1** ouais
- 0021 **BV1** voilà
- 0022 **ch_OB1** d'accord vers (0.58) quand j -qu 'est -ce que je connais par là euh euh Villerbon non ?tout ça non non c'est au nord ça non
- 0023 **BV1** après Chambord Vineuil
- 0024 **BV1** ah non c'est au nord

ZuRecht: Recherche CQP

Saisissez une expression de recherche CQP

Mode recherche ▼

`(([norm="à"][norm="la"])?[norm="limite"])`



1	ESLO2_ENT_1005_C	EH5	... en euh à Orléans sud limite Olivet parce qu ' on habitait ...	ZuViel
2	ESLO2_ENT_1005_C	EH5	... en a de mauvaises mais à la limite même si on regarde une ...	ZuViel
3	ESLO2_ENT_1005_C	EH5	... de mauvaises mais à la limite même si on regarde une ...	ZuViel
4	ESLO2_ENT_1013_C	ZD13	... fait pas parce que euh à la limite euh ((bb)) y a pas cette ...	ZuViel
5	ESLO2_ENT_1013_C	ZD13	... parce que euh à la limite euh ((bb)) y a pas cette ...	ZuViel

Résumé et perspectives

- ZuMult comme architecture flexible pour accéder aux corpus oraux
 - Niveau supplémentaire d'abstraction / de standardisation
 - Enrichissement automatique des données: normalisation, lemmatisation, étiquetage POS, annotation phonétique, calcul de mesures diverses
 - Outils prototypiques:
 - ZuMal : Sélection de données
 - ZuRecht : Recherche en CQP
 - ZuViel : Visualisation de transcription
- Testé avec les corpus de l'Archive de l'Allemand Parlé (AGD) et FOLK
- Développement continu à l'IDS et en Open Source
- Prochaine station: Texas German Dialect Corpus, UT Austin



<https://linguisticbits.de>
thomas@linguisticbits.de
<http://www.exmaralda.org>
<http://zumult.ids-mannheim.de>

ZuRecht: Recherche CQP

Saisissez une expression de recherche CQP

Mode recherche ▼ [norm="merci"]

1	ESLO2_ENT_1001_C	ch_OB1	bon d ' accord si euh merci bien bah y a pas ...
2	ESLO2_ENT_1001_C	ch_OB1	y a pas de soucis merci bien bah j ' ai tout ...
3	ESLO2_ENT_1001_C	ch_OB1	... on peu terminer là euh merci beaucoup je vais arrêter euh ...
4	ESLO2_ENT_1005_C	ch_OB1	accord ((rire)) ((i)) d ' accord bah écoutez merci beaucoup merci vraiment pour cet ...
5	ESLO2_ENT_1005_C	ch_OB1	... accord bah écoutez merci beaucoup merci vraiment pour cet euh cet ...
6	ESLO2_ENT_1013_C	ZD13	... non c ' est tout hm merci Pauline et ben écoute merci ...
7	ESLO2_ENT_1013_C	ch_PP6	merci Pauline et ben écoute merci beaucoup au revoir ((rire)) ((nontrans)) ...