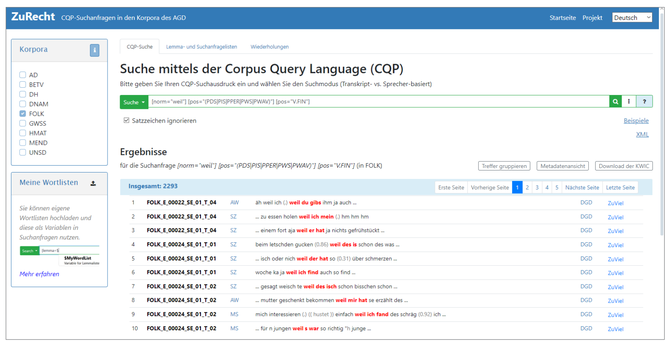
Nutzergruppenspezifische Zugänge zu mündlichen Korpora aus dem Archiv für Gesprochenes Deutsch: neue Tools, neue Forschungsperspektiven

Elena Frick / Henrike Helmer (Leibniz-Institut für Deutsche Sprache, Mannheim)

Projekt „ZuMult“, gefördert durch die DFG im LiS-Programm.
Kooperationspartner:
hzsk hamburger zentrum für sprachkorpora
Herder-Institut | UNIVERSITÄT LEIPZIG

ZuRecht

Zugang zur Recherche in Transkripten



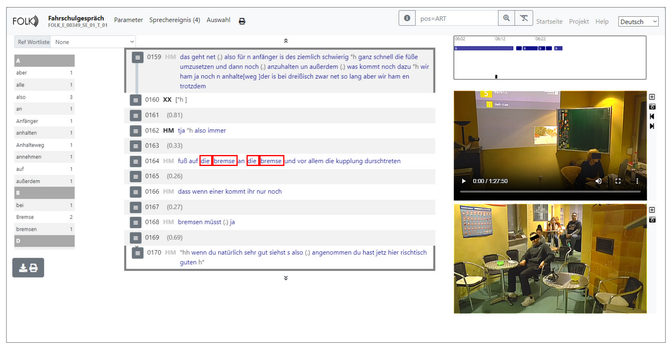
MTAS (Lucene)-basierte Suchmaschine

- Komplexe Suchanfragen nach zeitbasierten und sprecherübergreifenden Annotationen, nach Sprecherüberlappungen und Sprecherturns, nach unbestimmten Tokens, Pausen und anderen verbalen und non-verbalen Elementen;
- Suche mit Vokabellisten zu einem z.B. für den Sprachunterricht relevanten Thema;
- aktuell durchsuchbar sind 9 Korpora (Umfang: ca. 1.700 Stunden / ca. 12 Millionen Token)

Weitere Korpora des Archivs im Umfang von ca. 4.797 Stunden (davon transkribiert: 2.836 Stunden / 21.353.146 Token) können über die **Datenbank für Gesprochenes Deutsch (DGD)** durchsucht werden, URL: dgd.ids-mannheim.de)

ZuViel

Zugang zu Visualisierungselementen in Transkripten

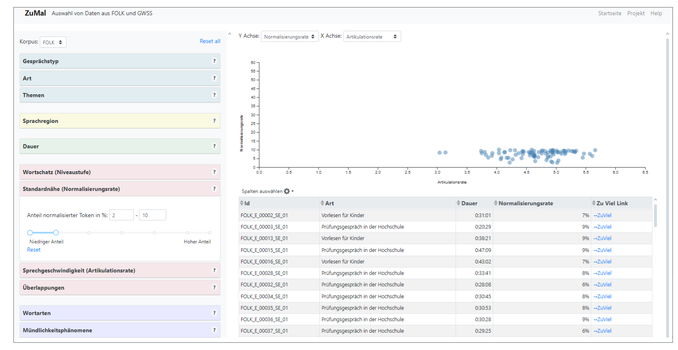


Transkriptbrowser mit synchronisierter Audio-/Videoanzeige

- Download ausgewählter Audio-/Video- und Transkriptpassagen
- Anzeige und Download von Wortlisten zu Transkripten
- Videountertitel
- Einstellbare Abspielgeschwindigkeit von Audio-/Videoaufnahmen
- Transcript Density Navigator
- Verschiedene Anzeigeoptionen von Transkripten (u.a. Visualisierung der schwierigkeitsbezogenen Phänomene in einzelnen Transkripten (z.B. Sprechgeschwindigkeit), was eine schnelle Beurteilung der Eignung des entsprechenden Korpusabschnittes für die Lehre erlaubt.)

ZuMal

Zugang zur Merkmalsauswahl von Gesprächen



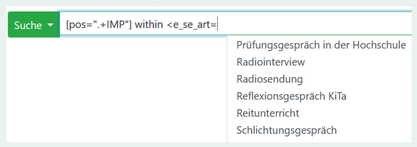
Filterung der einzelnen Interaktionen nach sprachdidaktisch relevanten und schwierigkeitsbezogenen Kriterien.

- Hierfür werden zum einen Kriterien genutzt, die auf die Metadaten der Sprechereignisse zurückgehen (z.B. Gesprächstyp, Thema, Sprachregion, Dauer des Gesprächs).
- Zum anderen wird auf Informationen zurückgegriffen, die mit digitalen Methoden automatisch aus den Korpusdaten berechnet werden können (z.B. Niveaustufenzugehörigkeit des enthaltenen Wortschatzes, Standardnähe, Sprechergeschwindigkeit, hoher/niedriger Anteil an Mündlichkeitsphänomenen)

SUCHE NACH WIEDERHOLUNGEN IN GESPROCHENEM DEUTSCH

Suchanfragen

- Anfragen mit der MTAS eigenen Variante der CQP-Suchanfragesprache
- Hilfe für Nutzende in Form von Auswahlmöglichkeiten über den Query Builder



Suchmodus

- Lexikalische Ebene, auf der Token miteinander verglichen werden. Es kann zwischen transkribierter, normalisierter und lemmatisierter Form gewählt werden.
- Zusätzlich können Nutzende über eine Schnittstelle zu **GermaNet** (Henrich und Hinrichs 2010) ihre Suche nach Wiederholungen auf Synonyme, Hyperonyme und Hyponyme ausweiten sowie ihre eigenen **Synonymlisten** hochladen und bei der Recherche verwenden.

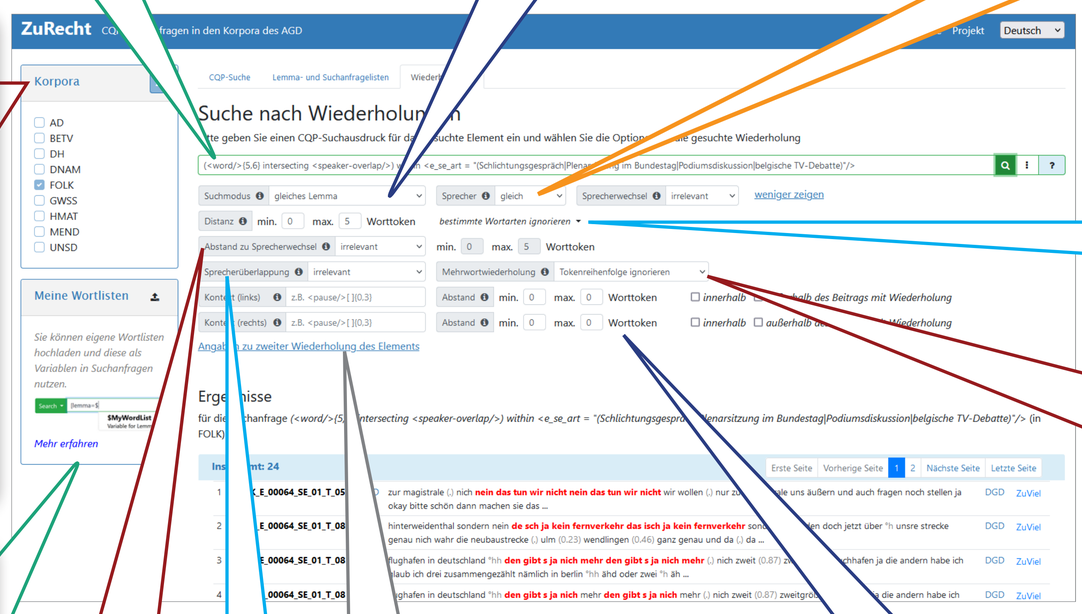
Sprecherspezifikation

- Self- und Other-Repetitions
- Bei der Suche nach Selbstwiederholungen kann zusätzlich angegeben werden, ob ein Sprecherwechsel zwischen dem gesuchten Element und der Wiederholung gewünscht ist.
- Bei der Suche nach Wiederholungen, die von anderen Sprechern realisiert werden, können Metadaten der Sprecher näher spezifiziert werden (z.B. das Geschlecht oder die Erstsprache).

Korpora des Archivs für Gesprochenes Deutsch (AGD)

z.B. FOLK (Forschungs- und Lehrkorpus Gesprochenes Deutsch):

- 350 Stunden Audio- und Videoaufnahmen,
- ca. 3,3 Millionen transkribierte Token
- Mehrebenenannotationen (normalisierte und lemmatisierte Formen, POS, phonetische Annotationen, Sprechgeschwindigkeit, Handlungsformate)



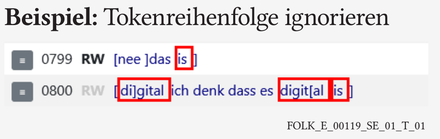
Distanz zwischen Wiederholungen

Bestimmte Wortarten wie etwa Artikel, Häsitationspartikel oder Interjektionen können vom Nutzer bei der Zählung der Tokenzahl zwischen den Wiederholungen ausgeschlossen werden.

Mehrwortwiederholungen

Es kann spezifiziert werden, ob die Reihenfolge der Token bei Mehrwortwiederholungen berücksichtigt oder ignoriert werden soll.

Beispiel: Tokenreihenfolge ignorieren



Nutzung eigener Wortlisten

Nutzende können eigene Wortlisten als Variablen speichern und in CQP-Abfragen einbauen, um effizient eine große Anzahl von Wörtern suchen zu lassen, ohne diese händisch ins Suchfeld eintragen zu müssen.

mehrfache Wiederholungen

Eine zweite Wiederholung der Token kann in einer separaten Form mit den gleichen Optionen konfiguriert werden.

Kontext

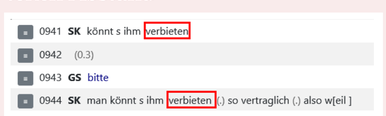
Ein CQP-Ausdruck kann verwendet werden, um ein Worttoken oder eine Worttokensequenz anzugeben, die direkt vor oder nach der Wiederholung vorkommen sollen. Dabei kann man festlegen, ob der angegebene Ausdruck innerhalb oder außerhalb des Beitrags mit Wiederholung vorkommen soll.

Beispiel: [norm="was"] [norm="für"], Distanz: min. 0, max. 1 Worttoken, innerhalb des Sprecherbeitrags

Abstand zum Sprecherwechsel

Angabe, mit welchem Tokenabstand zum vorherigen oder folgenden Sprecherwechsel Wiederholung gefunden werden soll.

Beispiel: nach dem Sprecherwechsel, min. 0, max. 5 Worttoken Abstand



Sprecherüberlappungen

Option zur Suche nach Wiederholungen innerhalb oder außerhalb von Überlappungen

Kontakt:
Elena Frick M.A.
Abteilung Pragmatik
Leibniz-Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim

Tel.: +49 621 1581-492
Fax: +49 621 1581-200
frick@ids-mannheim.de

Hausanschrift:
Leibniz-Institut für Deutsche Sprache
R 5, 6-13
68161 Mannheim

Tel.: +49 621 1581-0
Fax: +49 621 1581-200
info@ids-mannheim.de
www.ids-mannheim.de

Leibniz Gemeinschaft