Originalveröffentlichung in: Korpora Deutsch als Fremdsprache Jg. 3 (2023) Nr. 1, S. 1-12.
DOI: *https://doi.org/10.48694/kordaf.3721*

Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 3, Nummer 1 (Juli 2023)
ISSN 2749-4411

ZUGÄNGE ZU MÜNDLICHEN KORPORA FÜR DAF UND DAZ
Projekt, Datengrundlagen, technische Basis

Thomas Schmidt, Leibniz-Institut für Deutsche Sprache Mannheim
Christian Fandrych, Herder-Institut der Universität Leipzig
Elena Frick, Leibniz-Institut für Deutsche Sprache Mannheim
Matthias Schwendemann, Herder-Institut der Universität Leipzig
Franziska Wallner, Herder-Institut der Universität Leipzig
Kai Wörner, Zentrum für nachhaltiges Forschungsdatenmanagement Universität Hamburg

Abstract

Im vorliegenden Artikel wird ein Überblick über das von der DFG geförderte Projekt *Zugänge zu multimodalen Korpora gesprochener Sprache – Vernetzung und zielgruppenspezifische Ausdifferenzierung* (*ZuMult*) gegeben. Dabei wird zunächst auf die Sprachdaten und auf die technische Basis der Applikationen eingegangen, die dem Projekt zugrunde liegen. Im Anschluss werden die weiteren Beiträge in diesem Themenheft von *KorDaF* kurz vorgestellt. Übergeordnetes Thema von *ZuMult* ist die Verbesserung der Zugänglichkeit von digitalen mündlichen Sprachdaten für verschiedene Anwendungen und Zielgruppen, wobei der Fokus dieses Themenhefts auf Applikationen und Anwender:innen aus der Fremdsprachendidaktik und der DaF-/DaZ-Forschung und -Lehre liegt. Die einzelnen Beiträge beleuchten zentrale methodische und/oder technische Aspekte dieses Themas und beschreiben die Architektur und verschiedene prototypische Anwendungen, die das Projekt entwickelt hat.

Keywords: gesprochene Sprache; mündliche Korpora; Korpora in DaF/DaZ; Drei-Ebenen-Architektur; interne Ebene; konzeptuelle Ebene; externe Ebene; ZuMult-Datenmodell

Abstract

This article gives an overview of the DFG-funded project *Access to Multimodal Spoken Language Corpora - Networking and Target Group-Specific Differentiation* (*ZuMult*). It first presents the spoken language data and the technical foundation of the applications on which this project is based. This is followed by a short presentation of the contributions assembled in this thematic *KorDaF* issue. Their common theme consists in the enhancement of the accessibility of digital spoken language data for various applications and target groups, in particular applications and users from foreign language pedagogy and GFL/GSL research and teaching. The individual contributions highlight key methodological and/or technical aspects of this topic and describe the architecture and various prototypical applications which have been developed in the context of this project.

Keywords: spoken language; oral corpora; corpora in GFL/GSL; three-tier architecture; physical/internal level; logical/conceptual level; view/external level, ZuMult data model

> [We know] *far less about how best to support access to extended sessions of spontaneous speech. There is also a need for focussed 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 et al. 2005).

1. Ausgangspunkte

Mündliche Korpora bilden die empirische Basis zur Untersuchung vielfältiger Fragestellungen in der Linguistik (z.B. Gesprächsforschung, Soziolinguistik/Dialektologie, Phonetik/Phonologie, Korpuslexikographie), in der Sprachtechnologie (z.B. als Trainingsdaten für automatische

Spracherkennung und -synthese) und in weiteren wissenschaftlichen Disziplinen (z.B. qualitative Sozialforschung, Oral History Studies, Bildungsforschung, Psychologie). Der Aufbau solcher Korpora ist mit einem erheblichen organisatorischen, personellen und technischen Aufwand verbunden. Daher beschäftigen sich die betreffenden wissenschaftlichen Communities schon seit längerem mit der Frage, wie eine optimale Nutzung und Nachnutzung dieser Art von Forschungsdaten aussehen kann.

Die Bedingungen für die Arbeit mit mündlichen Korpora haben sich in den letzten fünfzehn Jahren grundlegend gewandelt. Die in diesem Zeitraum erfolgte Entwicklung von Korpustechnologie, die Etablierung von Standards und guten Praktiken, gezielte Initiativen zur Aufbereitung älterer Datenbestände, Projekte zum Aufbau neuer Korpora sowie die Einrichtung von Archiven und Distributionsplattformen haben dazu geführt, dass die sprachwissenschaftliche Forschung und Lehre nun erstmals in größerer Breite mit größeren Mengen mündlicher Daten arbeiten können. Damit erweitern sich auch die Möglichkeiten zum Einsatz mündlicher Korpora in der angewandten Linguistik.

Wie das einleitende Zitat feststellt, schließt sich an die Verfügbarkeit der Korpusdaten die Frage an, auf welche Weise diese für welche Nutzungsszenarien oder Gruppen von Nutzer:innen zugänglich gemacht werden, denn:

> 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 (Hunston 2022).

Im vorliegenden Themenheft stellen wir Lösungen vor, die im Rahmen des *ZuMult*-Projekts in diesem Sinne insbesondere für die Arbeit mit mündlichen Korpora im DaF-/DaZ-Kontext entwickelt wurden. Sie sind aber durchaus auch für andere Anwendungszwecke von Interesse.

2. Projekt

Das Projekt *Zugänge zu multimodalen Korpora gesprochener Sprache: Vernetzung und zielgruppenspezifische Ausdifferenzierung* (*ZuMult*)[1], dessen Ergebnisse wir in diesem Themenheft vorstellen, hatte sich vor der geschilderten Ausgangslage zum Ziel gesetzt, neue Zugänge zu bestehenden mündlichen Korpora zu schaffen, die gezielter an den Bedarfen bestimmter Nutzergruppen ausgerichtet sind, als dies bei ‚generischen' Korpusplattformen (wie beispielsweise der „Datenbank für Gesprochenes Deutsch", vgl. Schmidt 2017) der Fall ist. Für dieses Themenheft nehmen wir vor allem diejenigen Arbeiten des *ZuMult*-Projekts in den Blick, die auf DaF-/DaZ-Lehrende und die Verwendung von mündlichen Korpusdaten in der DaF-/DaZ-Praxis fokussieren.

ZuMult wurde zwischen 2018 und 2022 im Programm *Wissenschaftliche Literaturversorgungs- und Informationssysteme* (LIS) der Deutschen Forschungsgemeinschaft gefördert, Projektpartner waren das Leibniz-Institut für Deutsche Sprache in Mannheim (im Folgenden IDS), das Herder-Institut der Universität Leipzig sowie das Hamburger Zentrum für Sprachkorpora (im Folgenden HZSK) an der Universität Hamburg. Die Vorarbeiten zur zusätzlichen Annotation von Handlungssequenzen, die in das Teilprojekt *ZuHand* einflossen, wurden mit einem DAAD-Stipendium für Julia Kaiser finanziert. Die *ZuMult*-Projektpartner aus Hamburg und Mannheim brachten vor allem korpustechnologische Expertise aus dem Kontext der Entwicklung des Archivs für Gesprochenes Deutsch (AGD, vgl. Stift / Schmidt 2014), der Datenbank für Gesprochenes Deutsch (DGD, vgl. Schmidt 2017) und der Angebote des Hamburger Zentrums für

[1] https://zumult.org (27.02.2023).

Sprachkorpora (HZSK, vgl. Schmidt et al. 2011) ein. Die Leipziger Projektpartner erarbeiteten insbesondere die Konzepte für anwendungsorientierte sprachdidaktische Zugänge, überprüften deren Aussagekraft, wählten verschiedene Parameter aus, begleiteten die Umsetzung und überprüften deren Funktionalität. Ihre Expertise liegt in der sprachdidaktisch orientierten Korpuslinguistik und Mündlichkeitsdidaktik (vgl. etwa Fandrych et al. 2021). Neben den für die einzelnen Beiträge des Themenhefts aufgeführten Autor:innen haben auch Annette Portmann und Josip Batinić substanzielle Beiträge zur Projektarbeit geleistet.

Das Projekt *ZuMult* wurde mit dem Auslaufen der Förderung zum August 2022 abgeschlossen. Die im Rahmen des Projekts entwickelte Software wird aber weiter über den Programmbereich Mündliche Korpora des IDS zugänglich gemacht[2] und in Bezug auf Datenbasis und Funktionalität ausgebaut. Mit Version 2.16. der Datenbank für Gesprochenes Deutsch[3] wurden die *ZuMult*-Prototypen im Mai 2021 an diese Plattform angeschlossen und stehen somit über 16.500 registrierten DGD-Nutzer:innen zur Verfügung. Installationen von *ZuMult* an anderen Standorten und für andere Datentypen befinden sich zum Zeitpunkt des Erscheinens dieses Themenhefts in Planung.

3. Datengrundlagen

Die hier vorgestellten Zugänge wurden in erster Linie anhand zweier großer Korpora des gesprochenen Deutsch entwickelt: Zum einen ist dies das seit 2008 am IDS Mannheim aufgebaute *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (*FOLK*, vgl. Deppermann / Schmidt 2014; Schmidt 2014, 2016a, 2016b, 2018; Kaiser 2018), zum anderen das Korpus *Gesprochene Wissenschaftssprache kontrastiv* (*GeWiss*, vgl. Fandrych / Meißner / Slavcheva 2012; Meißner / Slavcheva 2014; Fandrych / Wallner 2023), das zunächst in einem gemeinsamen Projekt der Universität Leipzig, der Aston University in Birmingham und der Universität Wroclaw zwischen 2009 und 2013 erhoben und erschlossen wurde[4]. Darüber hinaus flossen auch weitere Korpora des Archivs für Gesprochenes Deutsch in die Entwicklung ein, insbesondere die Korpora *Deutsch Heute* (*DH*, vgl. Kleiner 2015), *Mennonitenplautdietsch in Nord- und Südamerika* (*MEND*, vgl. Kaufmann / Gorisch / Schmidt 2023) und *Deutsch in Namibia* (*DNAM*, vgl. Zimmer et al. 2020). Die Korpora des Hamburger Zentrums für Sprachkorpora (dort insbesondere das *Hamburg Map Task Corpus*, *HaMaTaC*, vgl. Hedeland / Schmidt 2012) dienten als Referenzpunkt für die Übertragbarkeit der im Projekt entwickelten Lösungen an andere Standorte.

Die genannten Korpora sind in der Literatur ausführlich beschrieben (siehe o.a. Referenzen). Wir beschränken uns daher für diesen Beitrag auf eine kurze Zusammenfassung der wichtigsten Eigenschaften von *FOLK* und *GeWiss*.

FOLK, das *Forschungs- und Lehrkorpus gesprochenes Deutsch* wird seit 2008 am Programmbereich „Mündliche Korpora“ des IDS aufgebaut. *FOLK* versteht sich als Referenzkorpus, in dem das gesprochene Deutsch in natürlichen Interaktionen in größtmöglicher Variationsbreite abgebildet werden soll. So enthält *FOLK* unterschiedlichste Interaktionstypen aus den Bereichen privater (z.B. private Telefongespräche, Tischgespräche), institutioneller (z.B. Unterrichtsstunden, berufliche Kommunikation) und öffentlicher Kommunikation (z.B. Fernseh-Talkrunden). Das Design und die Zusammensetzung von *FOLK* sind in Kaiser (2018) näher beschrieben. *FOLK* wird nach einem Workflow erschlossen, der neben der detaillierten Transkription auch mehrere Annotationsschritte umfasst und an dessen Ende Daten stehen, die sich optimal computergestützt

[2] http://zumult.ids-mannheim.de (27.02.2023).
[3] https://dgd.ids-mannheim.de (27.02.2023).
[4] In der Folgezeit wurde das Korpus durch Daten von weiteren Kooperationspartnern aus Sofia, Pisa und Jyväskylä ausgebaut, vgl. für eine Übersicht Fandrych / Wallner (2023).

verarbeiten und auswerten lassen (vgl. Schmidt 2017 und Westpfahl 2020). Aktuell (Februar 2023) umfasst *FOLK* 400 Gesprächsereignisse mit einer Gesamtdauer von über 336 Stunden, die auf Audio und zu einem großen Teil auch auf Video (teilweise aus mehreren Kameraperspektiven) aufgezeichnet wurden. Die zugehörigen Transkripte haben einen Gesamtumfang von über 3,2 Millionen Tokens.

GeWiss ist ein mehrsprachiges Vergleichskorpus der gesprochenen Wissenschaftssprache, das ausgewählte kommunikative Gattungen umfasst und in der ersten Projektphase (2009-2013) an drei Standorten entwickelt wurde (Leipzig/Deutschland, Birmingham/Großbritannien, Wrocław/Polen), in einer weiteren Phase kamen Daten aus Bulgarien (Sofia), Italien (Pisa) und Finnland (Jyväskylä) hinzu[5]. Als standortübergreifend relevante Gattungen wurden universitäre Prüfungsgespräche, studentische Referate und Expertenvorträge aus philologischen Fächern ausgewählt und für die Zwecke des Korpusaufbaus recht breit definiert (vgl. Fandrych et al. 2012). Das Korpus beinhaltet kommunikative Ereignisse in deutscher, englischer, italienischer und polnischer Sprache. Bei den Daten handelt es sich um ‚natürliche', d.h. nicht zum Zweck der Korpuserstellung elizitierte Sprachdaten. Neben L1-Daten enthält das Korpus auch L2-Produktionen für die Sprachen Deutsch und Englisch. Das Korpus umfasst 436 kommunikative Ereignisse mit ca. 1,2 Millionen Tokens und einer Gesamtlänge von 146 Aufnahmestunden; davon sind 276 kommunikative Ereignisse mit ca. 742.000 Tokens und einem Umfang von 92 Stunden deutschsprachig (vgl. ausführlicher Fandrych / Wallner 2023). Das Korpus eignet sich u.a. für den Vergleich von Gattungen in verschiedenen Sprachen oder auch an verschiedenen Standorten, für die Untersuchung von Gesprächsdaten von L2- und L1-Sprechenden sowie von Studierenden und Expert:innen. Die Daten liegen in transkribierter Form sowie auf Audio vor, die über die DGD abrufbare Version ist daneben durch weitere Aufbereitungs- und Annotationsschritte computerlinguistisch in vielfältiger Weise durchsuchbar (vgl. dazu auch Fandrych et al. in diesem Heft).

4. Technische Basis

Neben dem Zugangsaspekt hat *ZuMult* sich auch mit dem Aspekt der ‚Vernetzung' mündlicher Korpora auseinandergesetzt. Ziel war hier, eine Software-Architektur zu schaffen, die flexibel an unterschiedlichen Standorten eingesetzt und an lokale Gegebenheiten der Datenhaltung sowie verschiedene Nutzungsszenarien angepasst werden kann. Dies geschah auch vor dem Hintergrund der Beobachtung, dass existierende Korpusplattformen in aller Regel Insellösungen sind: „[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 2013: 154, vgl. dazu auch Batinić / Frick / Schmidt 2021).

Weil die einzelnen Beiträge sich auf den Zugangsaspekt konzentrieren, seien hier einleitend die wichtigsten Eigenschaften der Architektur und technischen Grundlagen kurz dargestellt.

(1) *ZuMult* arbeitet mit einer sogenannten Drei-Ebenen-Architektur. Diese unterscheidet zwischen:

- der internen Ebene (eng. *physical* oder *internal level*), die sich mit der physikalischen Speicherung von Daten befasst, z.B. in einer relationalen Datenbank, einem Datenrepositorium oder dem Dateisystem auf einem Server;

[5] Die in Finnland erhobenen Daten wurden 2018 integriert und sind ausschließlich über die Datenbank für gesprochenes Deutsch zugänglich; die anderen Daten können auch über das *GeWiss*-Portal abgerufen werden, vgl. https://gewiss.uni-leipzig.de/. Die unterschiedlichen Nutzungsmöglichkeiten von *GeWiss*, die mit den beiden Portalen verbunden sind, werden in Fandrych / Wallner (2023) näher beschrieben.

- der konzeptuellen Ebene (eng. *logical* oder *conceptual level*), in der die Objekte der Anwendung (im Fall von Korpora z.B. Audio- und Videoaufnahmen, Transkripte, Metadaten) in Datenstrukturen modelliert werden und die Operationen zur Verfügung stellt, die auf den Objekten ausgeführt werden (z.B. das Bilden eines Transkriptausschnitts, eine Suche auf Transkripten); die Gesamtheit der Datenstrukturen und -operationen werden über eine sogenannte API (Application Programming Interface) den externen Anwendungen zur Verfügung gestellt;
- der externen Ebene (eng. *view* oder *external level*), die die (typischerweise grafischen) Schnittstellen für Nutzer:innen umfasst, z.B. die Webseite im Browser, die ein Transkript anzeigt oder eine Webseite zum Eingeben und Auswerten einer Suchanfrage an ein Korpus.

Durch die konzeptionelle Trennung dieser drei Ebenen werden Software-Anwendungen flexibler: zum einen kann die Technologie auf der internen Ebene geändert werden (z.B. indem Daten von einer relationalen Datenbank ins Dateisystem übertragen werden), ohne dass konzeptuelle oder externe Ebene daran angepasst werden müssen. Zum anderen können auf der externen Ebene im Prinzip beliebig viele Anwendungen verschiedenen Typs (also neben Browser-Applikation z.B. auch Desktop-Anwendungen oder mobile Apps) entwickelt werden, die alle auf die gleiche konzeptionelle Ebene zugreifen können. Für die Beiträge in diesem Themenheft ist besonders der letzte Punkt wichtig: Die im Folgenden vorgestellten Prototypen *ZuMal*, *ZuRecht*, *ZuViel* und *ZuHand* sind Beispiele für verschiedene Anwendungen auf der externen Ebene, die alle auf eine gemeinsame Grundlage (die *ZuMult*-API) zurückgreifen.

(2) *ZuMult* stützt sich, wo immer möglich, auf Standards für die Datenrepräsentation, die in jüngerer Zeit auch in der Linguistik zunehmende Verbreitung erfahren haben. Für Audio- und Videoaufnahmen sind dies die Industrie-Standards PCM-WAV (Audio) und MPEG-4 (Video). Für strukturierte Textdaten (darunter fallen sowohl Transkripte mit ihren Annotationen als auch Metadaten zu Gesprächen und Sprecher:innen) wird generell davon ausgegangen, dass diese in XML vorliegen. Für Transkripte und Annotationen verwendet *ZuMult* den ISO-Standard „ISO 24624:2016 Language resource management — Transcription of spoken language“, der seinerseits auf den Richtlinien der Text Encoding Initiative (TEI) basiert. Dieser Standard ist interoperabel mit den Formaten der gängigsten Transkriptionseditoren (insbesondere *EXMARaLDA*, *FOLKER*, *ELAN*, *Praat* und *Transcriber*) und hat sich auch bei der Entwicklung anderer Anwendungen für linguistische Korpora bewährt (vgl. Hedeland / Schmidt 2022). Die Metadaten der konkret in den Anwendungen verwendeten Korpora folgen dem XML-Schema des Archivs für Gesprochenes Deutsch, zusätzlich ist *ZuMult* auf Metadaten eingestellt, die im XML-Format des *EXMARaLDA* Corpus Managers (*CoMa*) vorliegen. Weitere Standards, die in den Anwendungen zum Einsatz kommen, sind SVG (ein XML-basierter Standard zur Repräsentation von Vektor-Grafiken), XSL (ein XML-basierter Standard zur Transformation von XML-Daten) sowie VTT (ein Standard zur Kodierung von Untertiteln in Videos).

(3) In die Daten-Modellierung auf der konzeptuellen Ebene sind langjährige Erfahrungen eingeflossen, insbesondere aus der Entwicklung des *EXMARaLDA*-Systems (vgl. Schmidt / Wörner 2014), der Datenbank für Gesprochenes Deutsch, des HZSK-Korpusrepositoriums und des o.g. ISO-Standards für Transkripte gesprochener Sprache. Somit ist sichergestellt, dass die entwickelten Lösungen für vielfältige Typen mündlicher Korpora einsetzbar sind – also nicht nur für die hier im Fokus stehenden hauptsächlich deutschsprachigen Gesprächskorpora, sondern auch für Variationskorpora, Interviewkorpora und für Korpora in anderen Sprachen. Abbildung 1 stellt schematisch dar, welche Objekte eines Korpus in *ZuMult* modelliert werden und wie diese aufeinander bezogen sind.

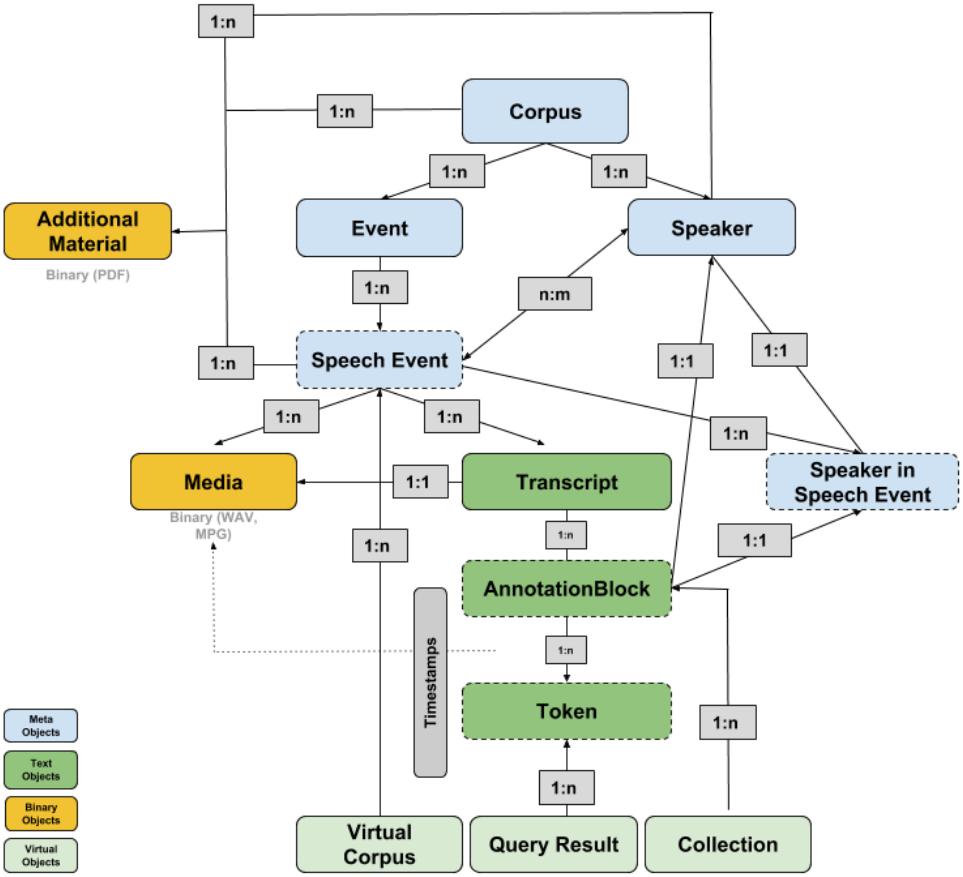

Abbildung 1
Datenmodell für *ZuMult*

(4) Bei der Implementierung der Anwendungen wurde darauf geachtet, möglichst weit verbreitete und etablierte Technologien zu verwenden, keine unnötigen Abhängigkeiten in den technischen Grundlagen zu schaffen und, wo immer möglich, auf vorhandenen Lösungen aufzubauen anstatt eigene von Grund auf neu zu entwickeln. Allgemein fiel die Wahl zur Implementierung der konzeptuellen Ebene daher auf das Java EE Framework, eine stabile Technologie, die seit Jahren von unzähligen Serveranwendungen verwendet wird. Für die Umsetzung der externen Ebene (d.h. der Webseiten) wurde mit jQuery, XSLT, HTML5 (inklusive der dort zur Verfügung stehenden Audio-/Video-Playback-Funktionalität) und Bootstrap gleichfalls auf bewährte und weit verbreitete technische Lösungen zurückgegriffen. Insbesondere die Verwendung von Bootstrap – eines Frameworks für die grafische Gestaltung von Webseiten, das auch bei vielen kommerziellen Webseiten zum Einsatz kommt – stellt sicher, dass Nutzer:innen mit den verwendeten Bedienungselementen bereits bis zu einem gewissen Grad vertraut sind, wenn sie mit einer der Anwendungen zu arbeiten beginnen.

(5) Auch die Query-Funktionalität, die in vielerlei Hinsicht ein zentrales Element von *ZuMult* darstellt, basiert auf Lösungen, die sich bereits in anderen Kontexten bewährt haben. So wurde die Suchkomponente der *ZuMult*-API mit Hilfe von MTAS (Multi Tier Annotation Search, vgl. Brouwer et al. 2016) umgesetzt. Es handelt sich um ein auf der Volltextsuch-Technologie Lucene basierendes, in Java implementiertes *Open-Source-Framework* zum Indizieren und Durchsuchen großer Sammlungen digitalisierter Texte mit vielfältigen Annotationen. Im *ZuMult*-Projekt wurde dieses Framework zum ersten Mal für die Suche in mündlichen Korpora eingesetzt (vgl. Frick / Schmidt

2020). Eine Besonderheit des MTAS-Frameworks ist sein JavaCC Parser, der eine formale CQP-basierte Suchanfragesprache (vgl. Evert et al. 2022) in Lucene-interne Suchanfragen übersetzt und somit einerseits effiziente Suchanfragen in einer speziell für Korpussuche entwickelten und Korpuslinguist:innen mehr oder weniger vertrauten Sprache ermöglicht und andererseits von der Geschwindigkeit und Skalierbarkeit der Lucene Technologie profitiert. Die vom MTAS verwendete Apache Lucene Programmbibliothek stellt eine weit verbreitete technische Lösung für Suchmaschinen dar, in der Linguistik und darüber hinaus, was eine solide Basis für Nachhaltigkeit bietet.

(6) Für den Zugriff auf die Funktionalität der konzeptuellen Ebene stellt *ZuMult* eine sogenannte REST-API als Programmierschnittstelle zur Verfügung. Damit können Entwickler:innen auf die Daten zugreifen, ohne sich mit den Details der Implementierung oder des Daten-Backends beschäftigen zu müssen. Dies erleichtert die Entwicklung neuer Anwendungen und die Integration von *ZuMult*-Funktionialität in andere Kontexte.

Zum Projektende wurde der *ZuMult* zugrunde liegende Code unter einer Open-Source-Lizenz auf der Plattform GitHub eingestellt und ist somit auch für andere Entwickler:innen für Anpassungen und Erweiterungen zugänglich.

5. Überblick über die Beiträge

Die einzelnen Beiträge dieses Themenhefts beschreiben jeweils methodische Überlegungen, die den einzelnen Anwendungsprototypen vorausgegangen sind, sowie deren Umsetzung als Web-Anwendungen für die Korpora *FOLK* und *GeWiss*.

Im Beitrag *ZuMal: Zielgruppenspezifische Gesprächsauswahl aus Korpora gesprochener Sprache* stellen Christian Fandrych, Cordula Meißner, Matthias Schwendemann und Franziska Wallner die in *ZuMult* entwickelten zentralen Auswahlwerkzeuge vor, die es ermöglichen, nach bestimmten Kriterien und Merkmalen gezielt Sprechereignisse für die weitere Nutzung auszuwählen. Hierfür wurden auf einer nutzerfreundlichen, adaptierbaren und mit verschiedenen Visualisierungen versehenen Oberfläche Filter- und Auswahlwerkzeuge angelegt, die miteinander kombinierbar sind und so die Recherche und gezielte Auswahl von Interaktionen für unterschiedliche Zwecke ermöglichen. Es finden sich Auswahlfilter, die auf den mit den Interaktionen erhobenen Metadaten basieren, daneben solche, die spezifischer aus fremdsprachendidaktischer Perspektive entwickelt wurden. Hierzu gehört die Möglichkeit, die Interaktionen der beiden Korpora nach bestimmten Schwierigkeitsindikatoren zu durchsuchen und entsprechend für didaktische Zwecke geeignete Sprechereignisse auszuwählen, ebenso wie Werkzeuge, die es erlauben, gezielt Interaktionen nach dem Auftreten wichtiger Mündlichkeitsphänomene sowie nach der Häufigkeit bestimmter Wortarten zu filtern. Der Beitrag erläutert dabei die diesen Auswahlmöglichkeiten jeweils zugrunde liegenden Konzepte, Vorannahmen und deren Operationalisierungen und zeigt praktisch und anhand von Beispielen, welche Anwendungsmöglichkeiten sich dabei ergeben.

Der Beitrag *ZuViel: Transkriptvisualisierung und Arbeiten mit Transkripten* von Thomas Schmidt, Matthias Schwendemann und Franziska Wallner beschreibt die mit dem Tool *ZuViel* geschaffenen Visualisierungsoptionen für Transkripte. Ein zentrales Anliegen von *ZuViel* besteht darin, Transkripte leicht lesbar und navigierbar zu machen und Nutzer:innen vielfältige Möglichkeiten anzubieten, die Visualisierung flexibel an eigene Bedürfnisse anzupassen. Im Beitrag wird erläutert, wie Nutzer:innen mit einem Transkript in *ZuViel* themengeleitet und -explorierend

interagieren können, und welche Anwendungsperspektiven sich mit diesem Tool für Nutzer:innen in didaktischen und forschungsbezogenen Kontexten ergeben.

Im Beitrag *ZuRecht: Neue Recherchemöglichkeiten in Korpora gesprochener Sprache für Gesprächsanalyse und Deutsch als Zweit- und Fremdsprache* stellen Elena Frick, Henrike Helmer und Franziska Wallner eine Benutzeroberfläche vor, die einen über bisherige Möglichkeiten deutlich hinausgehenden Zugriff auf Korpora des Archivs für gesprochenes Deutsch gestattet. Mithilfe komplexer Suchanfragen auf Transkripten gesprochener Sprache mit der speziell für die Korpusrecherche entwickelten Anfragesprache CQP lassen sich sowohl interaktional relevante als auch didaktisch motivierte Fragestellungen umfassend und effizienter als bisher bearbeiten. Im Beitrag wird dies anhand ausgewählter Suchanfragen aus der Gesprächsforschung und der Interaktionalen Linguistik sowie aus dem Kontext Deutsch als Fremd- und Zweitsprache demonstriert.

Der Beitrag *ZuHand: Zugang zu Handlungssequenzen und Themenausschnitten in einem qualitativ annotierten Subkorpus* von Julia Kaiser stellt dar, wie zielgruppenspezifische Annotationen eines (Teil-)Korpus über *ZuMult* nutzbar gemacht werden können. Das Tool *ZuHand* baut auf einem Annotations-Schema für Handlungen und Themen auf, das in qualitativer Korpusarbeit auf der Grundlage eines konversationsanalytisch ausgerichteten Ansatzes zur Vermittlung von Fremdsprachen entwickelt und auf einen Ausschnitt von *FOLK* angewandt wurde. *ZuHand* macht diese Annotationen unmittelbar für Sprachlehrende und -lernende nutzbar.

6. Fazit und Ausblick

Das Ziel von *ZuMult*, neue Zugangswege zu existierenden Korpora gesprochener Sprache, insbesondere für die DaF-/DaZ-Forschung und -Anwendung, zu entwickeln, hat sich als sehr anspruchsvoll erwiesen. Mit den hier vorgestellten Prototypen hoffen wir, diesem Ziel einen wesentlichen Schritt näher gekommen zu sein: Sie zeigen nicht nur das Potential, das authentische mündliche Daten für den DaF-/DaZ-Bereich bergen, sondern machen es bereits ein großes Stück weit in der Praxis nutzbar; hoffentlich auch für solche Anwender:innen, die sich dem Gegenstand ohne spezifische Expertise in korpuslinguistischer Methodik nähern. Auch wenn die Werkzeuge das Etikett ‚Prototyp' tragen, sind sie mit ihrer Einbettung in längerfristige Vorhaben des IDS, der dortigen Anbindung an die Datenbank für Gesprochenes Deutsch mit ihren über 16.500 registrierten Nutzer:innen und der Open-Source-verfügbaren Codebasis eine gute Grundlage für eine dauerhafte Weiterentwicklung und weitere Ausdifferenzierung.

Eine wesentliche Einschränkung ergibt sich derzeit aber aus der rechtlichen Autorisierung der Korpora *FOLK* und *GeWiss*: In den Einverständniserklärungen, die mit den aufgenommenen Personen getroffen wurden, wird die Nutzung der Daten auf akademische Forschung und Lehre beschränkt. Daher unterliegt auch die Arbeit mit *ZuMult* den Einschränkungen, die für die Korpora gelten. Das bedeutet zum gegenwärtigen Zeitpunkt, dass die Nutzung auf Hochschulen und Forschungseinrichtungen beschränkt ist – also etwa für DaF-/DaZ-Lehr- und Lernszenarien im Hochschul- und Wissenschaftsbereich möglich ist, leider aber nicht für entsprechende Angebote außerhalb des akademischen Bereichs. Ein datenschutzrechtlich auch für eine weitere Anwendung autorisiertes Korpus natürlicher Interaktionen bleibt ein Desiderat.

Über weitere Entwicklungen im Zusammenhang mit *ZuMult* werden wir auf absehbare Zeit über die Website https://zumult.org/ informieren.

Literatur und Ressourcen

Anthony, Laurence (2013): A critical look at software tools in corpus linguistics. In: *Linguistic Research* 30: 2, 141-161.

Batinić, Josip / Frick, Elena / Schmidt, Thomas (2021): Accessing spoken language corpora: an overview of current approaches. In: *Corpora* 16: 3, 417-445.

Brouwer, Matthijs / Brugman, Hennie / Kemps-Snijders, Marc (2016): MTAS: A Solr/Lucene based Multi-Tier Annotation Search solution. Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence. https://ep.liu.se/ecp/136/002/ecp17136002.pdf (27.02.2023).

Deppermann, Arnulf / Schmidt, Thomas (2014): Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik - Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2). In: Domke, Christine / Gansel, Christa (Hrsg.): *Korpora in der Linguistik - Perspektiven und Positionen zu Daten und Datenerhebung* [=Mitteilungen des Deutschen Germanistenverbandes 61: 1], 4-17.

Evert, Stephanie / CWB Development Team (2022): *The IMS Open Corpus Workbench (CWB) – CQP Interface and Query Language Manual. CWB Version 3.5*. https://cwb.sourceforge.io/files/CQP_Manual.pdf (27.02.2023).

Fandrych, Christian / Meißner, Cordula / Slavcheva, Adriana (2012): The GeWiss Corpus. Comparing Spoken Academic German, English and Polish. In: Schmidt, Thomas / Wörner Kai (Hrsg.): *Multilingual corpora and multilingual corpus analysis*. Amsterdam: John Benjamins, 319-337.

Fandrych, Christian / Meißner, Cordula / Wallner, Franziska (2021): Korpora gesprochener Sprache und Deutsch als Fremd- und Zweitsprache: Eine chancenreiche Beziehung. In: *Korpora Deutsch als Fremdsprache* 1: 2, 5-30. https://doi.org/10.48694/tujournals-76.

Fandrych, Christian / Wallner, Franziska (2023): Das GeWiss-Korpus: Neue Forschungs- und Vermittlungsperspektiven zur mündlichen Hochschulkommunikation. In: Deppermann, Arnulf / Fandrych, Christian / Kupietz, Marc / Schmidt, Thomas (Hrsg): *Korpora in der germanistischen Sprachwissenschaft: Mündlich, schriftlich, multimedial*. Berlin / Boston: De Gruyter, 129-160. https://doi.org/10.1515/9783111085708-007.

Goldman, Jerry / Renals, Steve / Bird, Steven / de Jong, Franciska / Federico, Marcello / Fleischhauer, Carl / Kornbluh, Mark / Lamel, Lori / Oard, Douglas W. / Stewart, Claire / Wright, Richard (2005): Accessing the Spoken Word. In: *International Journal on Digital Libraries*, 287-298.

Hedeland, Hanna / Schmidt, Thomas (2012): Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken German. In: Schmidt, Thomas / Wörner, Kai (Hrsg.): *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam: John Benjamins, 25-46.

Hedeland, Hanna / Schmidt, Thomas (2022): The TEI-based ISO Standard 'Transcription of spoken language' as an Exchange Format within CLARIN and beyond. In: Monachini, Monica / Eskevich, Maria (Hrsg.): *Selected Papers from the CLARIN Annual Conference 2021*. Linköping Electronic Conference Proceedings 189, 34-45. https://pdfs.semanticscholar.org/2245/ffc839b41bd78d65ae8ba4e4fdf10f46f2c7.pdf (27.02.2023).

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

Kaiser, Julia (2018): Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien. In: *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 19, 515-552. http://www.gespraechsforschung-online.de/fileadmin/dateien/heft2018/px-kaiser.pdf (27.02.2023).

Kaufmann, Göz / Gorisch, Jan / Schmidt, Thomas (2023): Das MEND-Korpus im Archiv für Gesprochenes Deutsch: Entstehung, Möglichkeiten, Grenzen. Erscheint in: Wolf-Farré, Patrick / Löff Machado, Lucas / Prediger, Angélica / Kürschner, Sebastian (Hrsg.): *Deutsche und weitere germanische Sprachminderheiten in Lateinamerika: Methoden, Grundlagen, Fallstudien*. Berlin: Lang.

Kleiner, Stefan (2015): „Deutsch heute“ und der Atlas zur Aussprache des deutschen Gebrauchsstandards. In: Kehrein, Roland / Lameli, Alfred / Rabanus, Stefan (Hrsg): *Regionale Variation des Deutschen. Projekte und Perspektiven*. Berlin u.a.: De Gruyter, 489-518.

Meißner, Cordula / Slavcheva, Adriana (2014): Das GeWiss-Korpus - ein Vergleichskorpus der gesprochenen Wissenschaftssprache des Deutschen, Englischen und Polnischen. Design und Aufbau. In: Fandrych, Christian / Meißner, Cordula / Slavcheva, Adriana (Hrsg.): *Gesprochene Wissenschaftssprache. Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron, 15-38.

Schmidt, Thomas (2014): Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD. In: *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 15, 196-233. http://www.gespraechsforschung-ozs.de/fileadmin/dateien/heft2014/px-schmidt.pdf (27.02.2023).

Schmidt, Thomas (2016a): Construction and Dissemination of a Corpus of Spoken Interaction - Tools and Workflows in the FOLK project. In: *Journal for Language Technology and Computational Linguistics* (Themenheft: Kupietz, Marc / Geyken, Alexander (Hrsg.): Corpus Linguistic Software Tools) 31: 1, 127-154.

Schmidt, Thomas (2016b): Good practices in the compilation of FOLK, the research and teaching corpus of spoken German. In: *International Journal of Corpus Linguistics* 21: 3, 396-418.

Schmidt, Thomas (2017): DGD – die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim. In: *Zeitschrift für germanistische Linguistik* 45: 3. Berlin / Boston: De Gruyter, 451-463.

Schmidt, Thomas (2018): Gesprächskorpora. In: Kupietz, Marc / Schmidt, Thomas (Hrsg.) (2018): *Korpuslinguistik*. Berlin / Boston: De Gruyter, 209-230.

Schmidt, Thomas / Wörner, Kai (2014): EXMARaLDA. In: Durand, Jacques / Gut, Ulrike / Kristoffersen, Gjert (Hrsg.): *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press, 402-419.

Schmidt, Thomas / Hedeland, Hanna / Lehmberg, Timm / Wörner, Kai (2011): Multilingual Corpora at the Hamburg Centre for Language Corpora. In: Hedeland, Hanna / Schmidt, Thomas / Wörner, Kai (Hrsg.): *Multilingual Resources and Multilingual Applications*. Proceedings of GSCL Conference 2011 Hamburg. Hamburg: Universität Hamburg, 227-233.

Stift, Ulf-Michael / Schmidt, Thomas (2014): Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In: Institut für Deutsche Sprache (Hrsg.): *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim: Institut für Deutsche Sprache, 360-375.

Westpfahl, Swantje (2020): *POS-Tagging für Transkripte gesprochener Sprache. Entwicklung einer automatisierten Wortarten-Annotation am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)*. Studien zur deutschen Sprache 83. Tübingen: Narr.

Zimmer, Christian / Wiese, Heike / Simon, Horst J. / Zappen-Thomson, Marianne / Bracke, Yannic / Stuhl, Britta / Schmidt, Thomas (2020): Das Korpus Deutsch in Namibia (DNam): Eine Ressource für die Kontakt-, Variations- und Soziolinguistik. In: *Deutsche Sprache* 48: 1, 210-232.

Biographische Notiz: Thomas Schmidt ist Software-Entwickler bei Musical Bits GmbH und linguisticbits.de. Er hat bis 2021 den Programmbereich „Mündliche Korpora“ und das Archiv für Gesprochenes Deutsch am Leibniz-Institut für Deutsche Sprache in Mannheim geleitet. Seine Forschungsinteressen liegen in der Korpuslinguistik und Korpustechnologie für mündliche Korpora sowie in der digitalen Lexikographie.

Kontaktanschrift:
Thomas Schmidt
Musical Bits GmbH
Nahestraße 28
55411 Bingen
Deutschland
thomas@linguisticbits.de

Biographische Notiz: Christian Fandrych ist Professor für Linguistik des Deutschen als Fremdsprache am Herder-Institut der Universität Leipzig. Schwerpunkte seiner Tätigkeit sind Wortbildung und Wortschatz des Deutschen, Grammatikvermittlung, Wissenschaftssprache, Text- und Gesprächslinguistik sowie Korpuslinguistik im Kontext des Deutschen als Fremd- und Zweitsprache.

Kontaktanschrift:
Christian Fandrych
Herder-Institut
Universität Leipzig
Beethovenstr. 15
04107 Leipzig
Deutschland
fandrych@uni-leipzig.de

Biographische Notiz: Elena Frick ist Computerlinguistin und wissenschaftliche Mitarbeiterin am Leibniz-Institut für Deutsche Sprache in Mannheim. Sie ist im Programmbereich “Mündliche Korpora” tätig und beschäftigt sich mit der Entwicklung digitaler Korpusanwendungen für sprachwissenschaftliche Forschung.

Kontaktanschrift:
Elena Frick
Leibniz-Institut für Deutsche Sprache
R5, 6-13
D-68161 Mannheim
Deutschland
frick@ids-mannheim.de

Biographische Notiz: Matthias Schwendemann ist wissenschaftlicher Mitarbeiter im Bereich Linguistik am Herder-Institut der Universität Leipzig. Seine Arbeitsschwerpunkte in Forschung und Lehre liegen in den Bereichen Lexikologie, Wissenschaftssprache und Erwerb und Entwicklung des Deutschen als Fremd- und Zweitsprache sowie der Analyse von Lernersprache.

Kontaktanschrift:
Matthias Schwendemann
Herder-Institut der Universität Leipzig
Beethovenstr. 15
04107 Leipzig
Deutschland
matthias.schwendemann@uni-leipzig.de

Biographische Notiz: Franziska Wallner ist wissenschaftliche Mitarbeiterin am Herder-Institut der Universität Leipzig. Ihre Forschungsschwerpunkte sind unter anderen das Deutsche als fremde Bildungs- und Wissenschaftssprache, die korpusbasierte Erforschung der gesprochenen Sprache, Mündlichkeitsdidaktik sowie die Nutzung von Korpora im Kontext von Deutsch als Fremd- und Zweitsprache.

Kontaktanschrift:
Franziska Wallner
Herder-Institut
Universität Leipzig
Beethovenstr. 15
04107 Leipzig
Deutschland
f.wallner@uni-leipzig.de

Biographische Notiz: Kai Wörner ist stellvertretender Leiter des Zentrums für nachhaltiges Forschungsdatenmanagement der Universität Hamburg. Zu seinen Hauptaufgaben gehören die langfristige Speicherung von Forschungsdaten und deren Kuratierung in Hinblick auf eine möglichst einfache Nachnutzbarkeit.

Kontaktanschrift:
Zentrum für nachhaltiges Forschungsdatenmanagement
Universität Hamburg
Monetastraße 4
20146 Hamburg
Deutschland
kai.woerner@uni-hamburg.de