IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

58. Jahrestagung
des Leibniz-Instituts für Deutsche Sprache

KORPORA IN DER GERMANISTISCHEN SPRACHWISSENSCHAFT

MÜNDLICH, SCHRIFTLICH, MULTIMEDIAL

15.–17. MÄRZ 2022

ABSTRACTS

Leibniz Gemeinschaft

Stand: 7. März 2022
Termine und Inhalte der Vorträge können sich kurzfristig ändern.
Bitte beachten Sie auch die entsprechenden Hinweise während der Jahrestagung.

Gestaltung: Norbert Cußler-Volz (IDS/Öffentlichkeitsarbeit)
© 2022 Leibniz-Institut für Deutsche Sprache, Mannheim.

KORPORA IN DER GERMANISTISCHEN SPRACHWISSENSCHAFT

MÜNDLICH, SCHRIFTLICH, MULTIMEDIAL

Die Tagung gibt einen Überblick zu aktuellen Entwicklungen der Erschließung und Nutzung von Korpora in der germanistischen Linguistik und darüber hinaus. Dabei steht im Vordergrund, wie bekannte und neue Korpora für die Untersuchung verschiedenster linguistischer Fragestellungen, z.B. der Lexikografie, der Gesprächsforschung, des Spracherwerbs oder der historischen Sprachwissenschaft, genutzt werden können. Im Einzelnen geht es um:

- Korpusangebote und Korpusdesign
- Forschungssoftware (Annotation, Erschließung, Auswertung, Visualisierung) für die Arbeit mit Korpora
- Korpusaufbereitung
- den Zusammenhang von Korpusaufbereitung und Nutzungsmöglichkeiten bzw. Forschungsfragestellungen
- ethische und rechtliche Aspekte der Korpussammlung, -aufbereitung, -bereitstellung und -nutzung

Diese Fragen werden im Kontext wissenschaftstheoretischer Überlegungen zur Frage des Nutzens von Korpora für die linguistische Erkenntnisbildung behandelt. Es werden dabei nicht nur klassische Schrift- und Tonkorpora, sondern auch Korpora, die Daten aus anderen und multiplen Medialitäten beinhalten, vorgestellt werden, insbesondere Videokorpora und *social media*-Korpora. Eine weitere Dimension sind Vergleichskorpora mehrerer Sprachen oder Medialitäten (mündlich vs. schriftlich) sowie diachrone (Vergleichs-)Korpora und der Blick auf nicht-deutschsprachige Korpusangebote.

DIE VORTRÄGE DER JAHRESTAGUNG:

DIENSTAG, 15. MÄRZ 2022, 13.45 UHR

Das Gesamtkonzept des Deutschen Referenzkorpus DeReKo – Vom Design bis zur Verwendung und darüber hinaus

Marc Kupietz / Harald Lüngen / Nils Diewald (IDS)

Aufgabe und Ziel des Deutschen Referenzkorpus DeReKo ist es, eine allgemeine Forschungsdatengrundlage für das IDS selbst und für die synchron arbeitende germanistische Linguistik insgesamt dauerhaft zu sichern. Zu diesem Zweck wird DeReKo seit 1968 am IDS aufgebaut und laufend stichprobenartig um ein möglichst breit gefächertes Spektrum des aktuellen deutschen Schriftsprachgebrauchs erweitert. LinguistInnen und, sofern möglich, auch Forschende aus angrenzenden Disziplinen sollen durch DeReKo in die Lage versetzt werden, sich für eine große Bandbreite an Fragestellungen und Sprachdomänen geeignet stratifizierte Sub-Korpora zu definieren, mithilfe derer sie bestehende Hypothesen zuverlässig testen und interessante neue Hypothesen gewinnen können.

Wir wollen in unserem Beitrag die sich durch die einschlägige Thematik der IDS-Jahrestagung bietende Gelegenheit nutzen, um einen Blick hinter DeReKo's Kulissen zu gewähren. Zum einen werden wir dabei die strategische Konzeption des Korpusausbauprojekts im Kontext seiner angrenzenden Projekte beleuchten. Dabei werden wir nach einer kurzen Rekapitulation von DeReKo's Grundlagen und Designprinzipien näher auf Erweiterungsstrategien eingehen und unseren Ansatz vorstellen, DeReKo mit begrenzten Ressourcen für einerseits möglichst viele, andererseits aber auch für innovative und anspruchsvolle Anwendungen nutzbar zu machen. Am Beispiel der Tokenisierung und der (Re-)Konstruktion und Modellierung von Metadaten werden wir außerdem exemplarisch Herausforderungen diskutieren, die häufig bei der Kompromissfindung zwischen unterschiedlichen linguistischen Anforderungen und allgemeinen wissenschaftlichen Maximen vor dem Hintergrund ökonomischer und technischer Notwendigkeiten entstehen. Dies erscheint uns relevant, da solche Zielkonflikte und notwendigen Kompromisslösungen oft schwer dokumentierbar sind und bei der Korpusnutzung unserer Erfahrung nach selten erahnt werden und daher einen oft nicht bemerkten aber potenziell starken Einfluss auf Untersuchungsergebnisse haben können. Zudem zeigt sich dabei, wie sehr auch bei solchen vermeintlichen Detailentscheidungen, die im Zuge der Erstellung großer Allzweck-Korpora häufig gefällt werden müssen, heterogene sprachwissenschaftliche, technische und infrastrukturelle Aspekte miteinander in Wechselwirkung stehen und welche Konsequenzen bzgl. der Verwendbarkeit damit verbunden sind.

Neben diesem Blick hinter die DeReKo-Kulissen wollen wir auch über aktuelle Neuerungen bei DeReKo und seinem Umfeld berichten. Diese betreffen zum einen substanzielle Erweiterungen von DeReKo selbst, wie z.B. im Bereich IBK und Social-Media das in Nottingham entwickelte NottDeuYTSch-Korpus mit YouTube-Kommentaren und ein

am IDS entwickeltes und kontinuierlich erweitertes Twitter-Korpus sowie im Bereich der Fachsprache das in Greifswald und Leipzig aufwändig entwickelte Geschriebene Ingenieurwissenschaftliche Korpus Gingko. Zum anderen betreffen diese auch neue Verwendungsmöglichkeiten von DeReKo hinsichtlich vergleichender und kontrastiver Forschung im Kontext der Initiative Europäisches Referenzkorpus (EuReCo) und der Durchführung komplexer, mehrgliedriger, reproduzierbarer und replizierbarer Analysen mit Hilfe der Korpusanalyseplattform KorAP sowie seiner Bibliotheken für die Programmiersprachen R und Python.

DIENSTAG, 15. MÄRZ 2022, 14.40 UHR

Das ZDL-Regionalkorpus: Design, lexikographische Nutzung und Korpusergänzungen

Andreas Nolda / Alexander Geyken / Adrien Barbaresi (BBAW Berlin)

In der 2015 erschienen Publikation über das Variantenwörterbuch formulieren Bickel, Hofer und Suter (2015: 544) fünf Desiderata, die ein Korpus als Datengrundlage für die lexikographische Beschreibung diatopischer Variation im Standarddeutschen idealerweise erfüllen sollte:

1. Die Textbasis müsse gezielt nach national und regional zuordenbaren Texten abgesucht werden können, und die potenziellen Varianten müssten räumlich zugeordnet werden können.
2. Das Korpus sollte möglichst nur neuere und neueste standardsprachliche Texte enthalten.
3. Das Korpus sollte groß genug sein, um auch bei selteneren Varianten aussagekräftige Treffermengen zu liefern.
4. Das Korpus sollte linguistisch getaggt sein.
5. Absolute und relative Frequenzen einer Variante in den Vollzentren des Deutschen sollten automatisiert erhoben werden können.

„Ein linguistisches Korpus, das alle diese Wünsche erfüllt," stellen Bickel, Hofer und Suter fest, „gibt es zur Zeit [2015] noch nicht."

In unserem Beitrag stellen wir das *ZDL-Regionalkorpus* (Nolda, Barbaresi und Geyken 2021) vor, das 2020 am Zentrum für digitale Lexikographie der deutschen Sprache (ZDL) erstmals verfügbar gemacht wurde und seitdem fortlaufend aktualisiert und ausgebaut wird. Aufgrund seines Designs und seiner Funktionalität strebt das ZDL-Regionalkorpus die Erfüllung der o.g. Desiderata an. Das ZDL-Regionalkorpus umfasst über 31 Millionen Zeitungsartikel aus Lokal- und Regionalressorts deutschsprachiger Zeitungen (Stand Dez. 2021). Es deckt den Zeitraum seit 1993 ab und wird monatlich aktualisiert. Alle Zeitungsartikel sind gemäß dem Arealkonzept (https://www.dwds.de/d/

regionalangaben) annotiert. Die Texte sind lemmatisiert und mit Wortarten versehen. Die über die Webseite des „Digitalen Wörterbuch der deutschen Sprache“ (DWDS, www.dwds.de) bereitgestellte Benutzerschnittstelle ermöglicht die spezifische Suche in Arealen und stellt Werkzeuge zur frequenzbasierten tabellarischen und kartographischen Darstellung zur Verfügung.

Die Daten des ZDL-Regionalkorpus sind der Ausgangspunkt für die lexikographische Beschreibung der diatopischen Variation im Digitalen Wörterbuch der deutschen Sprache. Bislang wurden etwa 1000 regional markierte Stichwörter mit Hilfe des ZDL-Regionalkorpus neu beschrieben. Wir zeigen anhand ausgewählter Beispiele, für welche Phänomene das ZDL-Regionalkorpus gut geeignet ist, gehen aber ebenso auf die Grenzen des Korpus ein.

Im dritten Teil des Beitrags zeigen wir anhand einiger ausgewählter Beispiele, wie das ebenfalls im ZDL erstellte Web-Monitorkorpus als ergänzende Ressource für ZDL-Regionalkorpus eingesetzt werden kann. Das Web-Monitorkorpus vereinigt über 500 reichweitenstarke Quellen des deutschen Sprachraums. Er wird täglich aktualisiert und weist derzeit eine Größe von über 1,1 Milliarden Tokens auf (https://www.dwds.de/d/korpora/webmonitor).

Literatur:

Barbaresi, Adrien und Alexander Geyken (2020). Die Webkorpora im DWDS – Strategien des Korpusaufbaus und Nutzungsmöglichkeiten. In Deutsch in Sozialen Medien: Interaktiv – multimedial – vielfältig, hg. v. Konstanze Marx, Henning Lobin und Axel Schmidt, Institut für Deutsche Sprache: Jahrbuch 2019, Berlin: de Gruyter, 345–348.

Bickel, Hans, Lorenz Hofer und Sandra Suter (2015). Variantenwörterbuch des Deutschen (VWB) – neu: Dynamik der deutschen Standardvariation aus lexikografischer Sicht. In Regionale Variation des Deutschen: Projekte und Perspektiven, hg. v. Roland Kehrein, Alfred Lameli und Stefan Rabanus, Berlin: de Gruyter, 541–562.

Geyken, Alexander, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand und Lothar Lemnitzer (2017). Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). Zeitschrift für germanistische Linguistik, 45(2), 327–344.

Nolda, Andreas, Adrien Barbaresi und Alexander Geyken (2021). Das ZDL-Regionalkorpus: Ein Korpus für die lexikografische Beschreibung der diatopischen Variation im Standarddeutschen. In Deutsch in Europa: Sprachpolitisch, grammatisch, methodisch, hg. v. Henning Lobin, Andreas Witt und Angelika Wöllstein, Institut für Deutsche Sprache: Jahrbuch 2020, Berlin: de Gruyter, 317–322.

DIENSTAG, 15. MÄRZ 2022, 15.20 UHR

Czech National Corpus: long-term language mapping at your service

Michal Křen (Karls-Universität Prag)

Introduction

The mission of the Czech National Corpus (CNC) is a long-term, extensive and continuous mapping of the Czech language. For this purpose, CNC systematically collects, processes and annotates language data to produce large general-purpose corpora of Czech with the aim to offer a diverse and high-quality language resources for empirical research. In addition to corpus compilation, CNC is also very active in the development of specialized analytical tools in the form of web-based user applications that enable effective and intuitive work with the language data.

This contribution aims to give an overview of the CNC project (LM2018137; funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures) and to present the main achievements and challenges in the domains of corpus compilation and application development.

Corpus compilation

Most of the CNC corpora can be characterized as traditional, with emphasis on well-defined composition and reliable metadata. The main data collection areas:

- Contemporary printed Czech: annually updated SYN-series corpora (current overall size 4.7G words, i.e. tokens not including punctuation) that include five 100M-word representative corpora published every five years that cover consecutive time periods;
- Contemporary spoken Czech: spontaneous informal conversations covered by the ORTOFON corpus (2.1M words) and semi-formal monologues covered by the ORATOR corpus (1.2M words);
- Contemporary web Czech: ONLINE monitor corpus of Czech internet media (more than 6G words) updated on a daily basis and complemented by the NET corpus of semi- formal internet communication (discussion forums and blogs; 176M words).

Other data collection areas include multilingual InterCorp parallel corpus (1.5G words), DIAKORP corpus of historical Czech (3.5M words) and DIALEKT corpus of Czech dialects (223k words).

Application development

Currently, there are thirteen CNC-developed web applications that view corpus data from various perspectives. All the applications are integrated into the CNC research portal at http://www.korpus.cz/ together with many-faceted user support and related

services. This contribution concentrates on selected applications that can be attractive also for linguists who work with languages other than Czech:

- KonText: general-purpose concordancer that supports various corpus types, including spoken and parallel corpora;
- Word at a glance: user-friendly word profiles based entirely on corpus data;
- Calc: corpus calculator divided into modules that cover typical statistical tasks commonly encountered in corpus research.

CNC has an established user community of more than 8,000 registered active users from the Czech Republic and abroad. In 2021, the average number of user queries exceeded 4,000 per day.

DIENSTAG, 15. MÄRZ 2022, 16.15 UHR

Korpora zur deutschen Sprache in Österreich. System- und soziolinguistische Perspektiven

Alexandra N. Lenz (Österreichische Akademie der Wissenschaften / Universität Wien)

Im Fokus des Vortrags steht die österreichische Sprachkorpus-Landschaft, die gleich über mehrere alte und jüngste Schätze verfügt, die es teilweise noch zu heben bzw. – in Abhängigkeit von der jeweiligen Spezifik der Korpora – aufzubereiten gilt. Die Diskussion wird sich auf vier ausgewählte Korpora konzentrieren, die im Hinblick auf die angekündigten Perspektiven der System- versus/und Soziolinguistik unterschiedliche Potentiale und Herausforderungen mit sich bringen:

1. Das Korpus des Wörterbuchs bairischer Mundarten in Österreich (WBÖ), das schriftlich fixierte Dialektdaten (v. a.) aus der ersten Hälfte des 20. Jahrhunderts umfasst, die größtenteils digitalisiert in einer xml-Datenbank (TEI) zur Verfügung stehen (s. Stöckle 2021).
2. Das Korpus Österreichische Dialektaufnahmen im 20. Jh. (ÖDiA20Jh), das sich aus ca. 530 Stunden Dialektaufnahmen inkl. Metadaten zusammensetzt, die v. a. in den 1950er bis 1980er Jahren flächendeckend in Österreich gesammelt wurden, und das aktuell an der Österreichischen Akademie der Wissenschaften umfangreiche Aufbereitungsstufen durchläuft (s. Lenz 2020).
3. Das Austrian Media Corpus (AMC), das in sich die gebündelte Presselandschaft Österreichs der letzten 20 Jahre vereinigt und – im Gegensatz zu den beiden oben genannten Korpora – eine Ausgangsbasis für standardschriftsprachliche Analysen darstellt (s. Ransmayr, Mörth & Ďurčo 2017).
4. Das Korpus des SFB „Deutsch in Österreich (DiÖ). Variation – Kontakt – Perzeption“, das eine Fülle von (medial) schriftlichen und mündlichen Sprachdaten umfasst, die system- wie soziolinguistische Perspektivierungen der areal-hori-

zontalen wie vertikal-sozialen Dimension der deutschen Sprache in Österreich ermöglichen (s. Lenz 2018).

Im Vortrag werden die Genese der verschiedenen Korpora und die Forschungsziele und -intentionen, die ihrem Aufbau, ihrer Struktur und (bisherigen) Aufbereitung zugrunde lagen bzw. liegen ebenso zur Sprache kommen wie auch ethisch-rechtliche Aspekte ihrer Verfügbarmachung. In einem zweiten Teil werden konkrete linguistische Analysen exemplarisch herangezogen, um Nutzungsmöglichkeiten und -grenzen im Hinblick auf system- und soziolinguistische Forschungsfragen zu illustrieren und reflektieren.

Literatur:

Lenz, Alexandra N. (2018): The Special Research Programme „German in Austria. Variation – Contact – Perception". In: Yearbook Sociolinguistica 32, 269–277.

Lenz, Alexandra N., Ludwig Maximilian Breuer, Christian Huber, Benjamin Fischer & Bernhard Graf (2020): „Österreichische Dialektaufnahmen im 20. Jahrhundert" – Zur Genese, Aufbereitung und wissenschaftlichen Nutzung eines einmaligen Sprachkorpus. In: Jahrbuch des Phonogrammarchivs 10, 128–140.

Ransmayr, Jutta, Karlheinz Mörth & Matej Ďurčo (2017): AMC (Austrian Media Corpus) – Korpusbasierte Forschungen zum österreichischen Deutsch. In: Wolfgang Ulrich Dressler & Claudia Resch (Hg.): Digitale Methoden der Korpusforschung in Österreich (Veröffentlichungen zur Linguistik und Kommunikationsforschung 30). Wien: Verlag der Österreichischen Akademie der Wissenschaften, 27–38.

Stöckle, Philipp (2021): Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In: Lenz, Alexandra N. & Philipp Stöckle (Hg.): Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts. Unter Mitarbeit von Angela Bergermayer, Andreas Gellan, Sabine Wahl, Eva-Marina Wahlmüller und Patrick Zeitlhuber (Zeitschrift für Dialektologie und Linguistik – Beihefte 181). Stuttgart: Steiner, 11–46.

MITTWOCH, 16. MÄRZ 2022, 9.30 UHR

Das Forschungs- und Lehrkorpus Gesprochenes Deutsch: Zum Nutzen eines großen annotierten Korpus gesprochener Sprache für interaktionslinguistische Fragestellungen

Arnulf Deppermann (IDS) / Silke Reineke (IDS) /
Thomas Schmidt (IDS, Universität Basel)

Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, Schmidt 2018) ist ein wachsendes Korpus von Audio- und Videoaufnahmen des gesprochenen Deutsch, das seit 2012 am IDS wissenschaftsöffentlich über die Datenbank für Gesprochenes Deutsch verfügbar gemacht wird (dgd.ids-mannheim.de). FOLK umfasst umfangreich dokumentierte Aufnahmen natürlicher Interaktionsereignisse aus allen Sektoren der Gesellschaft (privater und öffentlicher Bereich, Institutionen aus den Bereichen Bildung, Medizin,

Politik, Verwaltung, Wirtschaft, usw.). Sämtliche Daten sind nach den den GAT2-Konventionen für Minimaltranskripte (Selting et al. 2009) erschlossen und mit reichhaltigen linguistischen und gesprächsstrukturellen Annotationen versehen. Im Vortrag werden wir diskutieren, welche Prinzipien für die Kopuszusammensetzung und -weiterentwicklung in Hinblick auf regionale Verteilung, SprecherInnenmerkmale und im Korpus repräsentierte Interaktionstypen leitend sind. Wir präsentieren, welche linguistischen, gesprächsstrukturellen und soziologischen Auswertungen auf der Basis von Metadaten, Annotationen und eines strukturierten Datenmodells für die Transkripte datenbankseitig unterstützt werden. Wir stellen Typen von Untersuchungsinteressen vor, für die das Korpus im Bereich der Konversationsanalyse, der interaktionalen Linguistik, der Soziolinguistik und der Gesprochene-Sprache-Forschung nutzbar ist. Anhand einer Untersuchung des Formats *was heißt X?* zeigen wir sodann Schritt für Schritt exemplarisch, wie das Korpus für interaktionslinguistische Fragestellungen genutzt werden kann. Wir demonstrieren, wie datenbankbasierte Korpusanalyse, manuelle Auswertung und interaktionslinguistische Analyse einander ergänzen können und wie FOLK zur Erstellung von multimodalen Kollektionen von Videoausschnitten einer sprachlichen Praktik genutzt werden kann.

Literatur:

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

Selting, Margret et al. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung Online 10, S. 353-402.

MITTWOCH, 16. MÄRZ 2022, 10.10 UHR

Ein Korpus für die vergleichende Interaktionsforschung: Das ‚Parallel European Corpus of Informal Interaction' (PECII)

Jörg Zinken / Laurenz Kornfeld / Uwe-A. Küttner (IDS)

Wir stellen ein neues Parallelkorpus sozialer Alltagsinteraktionen vor: Das *Parallel European Corpus of Informal Interaction*, PECII. Das Korpus wird zurzeit am Leibniz-Institut für Deutsche Sprache aufgebaut, mit Unterstützung der Leibniz-Gemeinschaft und externer Projektpartner. Es beinhaltet Videoaufnahmen von Interaktionen in den folgenden Sprachen: Deutsch, Englisch, Italienisch, Polnisch; in geringerem Maße Finnisch und Französisch. Die Hauptidee des Korpus ist es, Videodaten sozialer Interaktion in verschiedenen Sprachen bereit zu stellen, die gleichzeitig (1) naturalistisch sind, d.h., die soziale Ereignisse erfassen, die nicht extra für die Zwecke der Aufnahme stattfinden; und die (2) vergleichbar sind, d.h., wir kontrollieren bestimmte Parameter der erhobenen Daten. So soll das Korpus eine neuartige Ressource für die sprach- und situationsübergreifende vergleichende Forschung von Sprachgebrauch undsozialer Interaktion werden. Folgende Interaktionstypen finden sich gegenwärtig im Korpus:

- Familienfrühstücke: Familien mit zwei oder mehr Kindern (im Alter zwischen 5 und 14 Jahren) frühstücken gemeinsam am Wochenende
- Autofahrten: Drei Erwachsene (Freunde oder Verwandte) fahren gemeinsam zu einem weiter entfernten Ziel (‚über Land')
- Brettspiele: Vier Erwachsene (Freunde oder Verwandte) spielen gemeinsam ein rundenbasiertes Brettspiel

Alle Ereignisse werden mit zwei Kameras gefilmt. Das Korpus soll nach Fertigstellung der wissenschaftlichen Öffentlichkeit über eine Datenbank zugänglich gemacht werden. Im Vortrag stellen wir die Konzeption des Korpus, besondere Herausforderungen bei der Datenerhebung und die von uns gewählten Lösungen vor, und zeigen exemplarisch das Potential des Korpus für sprach- und situationsvergleichende Forschungen zum alltäglichen Sprachgebrauch.

Literatur:

Rossi, Giovanni; Mondada, Lorenza; Sorjonen, Marja-Leena; Weidner, Matylda; Zinken, Jörg (2018): PECII: A parallel corpus for comparative research. ICCA. Loughborough University. Loughborough, UK, 13.07.2018.

Küttner, U.-A.; Kornfeld, L.; Mack, C.; Mondada, L.; Rogowska, J., Rossi, G.; Sorjonen, M.-L.; Weidner, M.; & Zinken, J. (in prep.): Introducing "The Parallel European Corpus of Informal Interaction" (PECII). New Avenues for Comparative Interactional Linguistic Research.

MITTWOCH, 16. MÄRZ 2022, 11.05 UHR

Korpora für die Diskursanalyse

Marcus Müller (TU Darmstadt)

Für die Diskursforschung sind Korpora konstitutiv, seit Michel Foucault (1972: 10) die Bildung „kohärenter und homogener" Textkorpora als eines der methodologischen Probleme der „neuen Geschichtswissenschaft" ausgewiesen hat. In der Germanistischen Linguistik hat man sich oft auf den Operationalisierungsvorschlag Dietrich Busses und Wolfgang Teuberts (1994: 4) berufen, Diskurse „im forschungspraktischen Sinn" als „virtuelle Textkorpora" zu verstehen. Ebenfalls konstitutiv für die Diskursforschung ist es aber, Kontexte des Sprachgebrauchs in die Analyse einzubeziehen, z.B. thematische, situationale oder soziale (Müller 2012). Es geht also gleichermaßen um Serialität und Kontextsensitivität von Daten – zwei operationale Prinzipien, die in der Methodologie eher auseinanderstreben und in der praktischen Arbeit jeweils aufs Neue zusammenzubringen sind: Da Textkorpora bekanntlich nichts anderes als distributionelle Information über die Ausdrucksseite sprachlicher Zeichen (Gries 2009: 1226) bieten, müssen sie zusätzlich mit Kontextinformation angereichert werden. Daher ist die Frage nach angemessenen und notwendigen Repräsentationsformaten von Kontexten beim Korpusdesign, bei der Datenmodellierung und beim Preprocessing

besonders wichtig. Eine große Rolle bei der korpuslinguistischen Kontextualisierung spielen auch Annotationen, insbesondere solche, die auf eigenen Tagsets beruhen, in einem ersten Schritt manuell durchgeführt werden und dann ggf. auf größere Datensätze angewendet werden sollen (z.B. Becker, Bender & Müller 2020; Müller, Bartsch & Zinn 2021).

Vor diesem Hintergrund thematisiert der Vortrag an ausgewählten Beispielen den Zusammenhang von Korpusaufbereitung, Datenanreicherung und Nutzungsszenarien im Kontext des Discourse Lab (https://www.discourselab.de), das an der TU Darmstadt und der Universität Heidelberg betrieben und in linguistischen und interdisziplinären Forschungs- und Lehrprojekten genutzt wird.

Literatur:

Becker, M., Bender, M. & Müller, M. (2020): Classifying Heuristic Textual Practices in Academic Discourse: A Deep Learning Approach to Pragmatics. In: International Journal of Corpus Linguistics (4)2020, pp. 426–460. https://doi.org/10.1075/ijcl.19097.bec

Busse, D. & Teubert, W. (1994): Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In: Busse, D., Hermanns, F. & Teubert, W. (Hgg.), Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik, 10–28. Opladen: VS Verlag für Sozialwissenschaften.

Foucault, M. (1972): The Archaeology of Knowledge. New York: Pantheon Books. [L'Archéologie du savoir, Paris : Gallimard, 1969].

Gries, S. Th. (2009): What is Corpus Linguistics?. Language and Linguistics Compass, 3: 1225–1241. doi:10.1111/j.1749-818X.2009.00149.x.

Müller, M. (2012): Vom Wort zur Gesellschaft: Kontexte in Korpora: Ein Beitrag zur Methodologie der Korpuspragmatik. In: Felder, E., Müller, M. & Vogel, F. (Hgg.), Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen, 33–82.

Müller, M., Bartsch, S. & Zinn, J.O. (2021): Communicating the unknown. An interdisciplinary annotation study of uncertainty in the coronavirus pandemic. In: International Journal of Corpus Linguistics 26/4, 498–531. https://doi.org/10.1075/ijcl.21096.mul

Internationales **DoktorandInnen-Netzwerk** des IDS

POSTERSESSION

Die Dativalternation in der Geschichte des Neuhochdeutschen: Eine historische und korpusbasierte Untersuchung

Evi van Damme (Gent, Belgien)

Das Forschungsprojekt befasst sich mit der Entwicklung der Dativalternation (z.B. *Er verkaufte dem Kloster zwei Äcker* vs. *Er verkaufte die Äcker an das Kloster*) im Neuhochdeutschen seit 1650 und gliedert sich in zwei Teiluntersuchungen: Neben der Durchsicht historischer sprachwissenschaftlicher Quellen liegt der Fokus auf der quantitativen und qualitativen Analyse historischer Korpusbelege.

Variation des segmentalen Merkmals Quantität auf höheren Ebenen der phonetischen Manifestation aufgrund gewählter Textsorten im Deutschen, Italienischen und Polnischen

Aleksandra Molenda (Breslau, Polen)

Das Dissertationsprojekt setzt Untersuchungen der Vokal- sowie Konsonantenquantität auf suprasegmentaler und prosodischer Ebene voraus. Da das Merkmal Quantität trotz unterschiedlichen phonologischen Status in den analysierten Sprachen systematisch realisiert wird und als ein prosodisches Phänomen nicht ausgeschlossen ist, soll seine Funktion in realen mündlichen Texten eruiert werden.

Wie verhält sich das Englische im deutschen Schriftsystem? Graphematische Zweifelsfälle bei Integrationsprozessen und ihre Steuerungsfaktoren

Marco Gierke (Mannheim)

Wächst zusammen, was zusammengehört? Anhand von Bildungen aus dem Englischen wird das Konditionierungsgefüge graphematischer Zweifelsfälle im Schriftsystem des Deutschen untersucht. Ziel ist es, den Einfluss längenhierarchischer Aspekte bei der Gestaltung der Kompositionsfuge im Zusammenspiel mit anderen, beispielsweise prosodischen Steuerungsfaktoren bemessen zu können.

Die orthographische Normalisierung deutschsprachiger Kontaktvarietät-Daten

Margaret Blevins (Austin, Texas, USA)

Dieses Dissertationsprojekt befasst sich mit der orthographischen Normalisierung deutschsprachiger Kontaktvarietäten. Es wird ein Multiebenen-System zur orthographischen Normalisierung deutschsprachiger Kontaktvarietäten vor (Normalisierung + Sprach-Tagging) vorgeschlagen. Diese Studie konzentriert sich vorerst auf die Daten des *Texas-German-Dialect-Projects* (Boas et al. 2010).

Zur Adaptabilität von Zeigegesten in der sozialen Interaktion

Mojenn Schubert (Mannheim)

Das Forschungsprojekt adressiert die Frage, inwiefern im gestischen Verhalten Interagierender eine Orientierung an interaktional bedingten Voraussetzungen und Erfordernissen deutlich wird. Diese Interaktionsorientierung wird am Beispiel von Zeigegesten in drei Fallstudien herausgearbeitet. Hierfür werden ihre Aufrechterhaltung, ihre gestische Transformation sowie ihre interpersonale Verwendung analysiert.

Sprachliche Manifestation von Widerstand in der psychodynamischen Psychotherapie und deren kommunikative Bearbeitung

Carolina Fenner (Mannheim)

Die kumulative Dissertation entsteht im Rahmen des interdisziplinären DFG-Projekts „Sprachliche Manifestation von Widerstand in der psychodynamischen Psychotherapie". In zwei Artikeln werden jeweils ein Widerstandsphänomen und eine Form des Widerstandsmanagements linguistisch und sequenzanalytisch untersucht. In einem dritten Artikel wird ein weiteres Widerstandsphänomen im longitudinalen Verlauf gesprächsanalytisch betrachtet.

Veränderungen in der sprachlichen Raumkonstruktion/ Ortsherstellung im Terrorismusdiskurs

Petr Kuthan (Brünn, Tschechien/Würzburg)

Im Zentrum des diskurs- und raumlinguistisch verankerten Forschungsprojekts stehen Fragen nach lexikalischen und argumentativen Konstanzen und musterhaften Veränderungen in der sprachlichen Raumkonstruktion nach einem terroristischen Anschlag. Dazu wird ein Textkorpus aus regionalen und überregionalen Tages- und Wochenzeitungen sowie aus Online-Artikeln in quantitativen und qualitativen Analysen untersucht.

Deontisches Sprechen im Deutschen und Italienischen. Eine komparative interaktionslinguistische Betrachtung sprachlicher Modalität in informellen Settings

Christina Mack (Mannheim)

Anhand eines zweisprachigen (Dt./It.) Video-Korpus wird untersucht, wie aus interaktionalen Prozessen heraus deontische Bedeutung entsteht. Neben verbaler Modalität berücksichtigt die Analyse auch die multimodale Ausgestaltung von Handlungssequenzen. Besonders interessiert dabei das Spannungsfeld zwischen verschiedenen Ressourcen von Einzelsprachen und universellen interaktionalen Aufgaben.

Intervenieren im Alltag. Eine konversationsanalytische Untersuchung

Laurenz Kornfeld (Mannheim)

Das Konfrontieren einer anderen Person auf Basis einer Regel oder Norm ist häufig zu beobachten. In Brettspielen beispielsweise werden Spieler:innen teilweise mit Regelverletzungen konfrontiert und können auf die Regel aufmerksam machen und intervenieren, aber auch außerhalb von Brettspielen lassen sich diese Momente identifizieren. In der Dissertation werde ich u.a. sprachstrukturelle Aspekte als auch Merkmale der Wissens- und Kompetenzzuschreibung untersuchen.

Interaktionale Herstellung von Autonomie und Autorität in informellen Interaktionen im deutsch-polnischen Sprachvergleich

Jowita Rogowska (Mannheim)

Meine Dissertation entsteht im Rahmen des Projekts *NoRM-aL*. Sprachvergleichend gehe ich der Frage nach, wie Sprecher*innen in alltäglichen Momenten ihre Autonomie und ihre Autorität lokal aushandeln, wie diese deontischen Handlungen verbal und multimodal gestaltet werden und wie sie sich in Sequenzen interaktional auswirken. Die Datengrundlage bilden polnische und deutsche Aufnahmen aus dem Korpus PECII.

Individuelle linguistische Variabilität in sozialen Medien: Ein multimediales Korpus

Tatjana Scheffler / Hannah Seemann / Lesley-Ann Kern
(Ruhr-Universität Bochum)

Soziale Medien wie Blogs oder Netzwerke wie Twitter stellen heute einen wichtigen Ort der informellen Textproduktion dar und eignen sich aufgrund ihres hohen Grades an Digitalisierung hervorragend als Grundlage für korpuslinguistisches Arbeiten. Digitale Korpora sozialer Medien sind als Schnittmenge zwischen konzeptioneller Mündlichkeit und schriftsprachlicher Ausarbeitung (Koch & Oesterreicher, 1985, 2007) von besonderer Bedeutung für die Untersuchung von Fragestellungen diverser linguistischer Teildisziplinen. Wir berichten über ein neues Social-Media-Korpus, das es erlaubt, sprachliche Daten derselben Sprecher:innen über zwei soziale Medien hinweg miteinander zu vergleichen und somit Rückschlüsse auf die individuelle linguistische Variabilität innerhalb sozialer Medien zu ziehen.

Zur Datensammlung wurde eine Liste von Elternblogger:innen verwendet, die sowohl auf Twitter aktiv sind als auch eigene Blogs pflegen. Zu dieser Liste wurden drei weitere bekannte Elternblogger:innen hinzugefügt. Die Tweets dieser Nutzer:innen zwischen Oktober 2016 und Februar 2017 wurden mit Hilfe des Python-Pakets *tweepy* über die Twitter-API extrahiert. Über die Twitter-Biographie wurde auf den Blog der jeweiligen Person zugegriffen und der Text via RSS-Feed oder über Feedreader ausgelesen. Es entsteht jeweils ein Dokument mit Blogposts der Nutzer:innen und ein Dokument mit den gesammelten Tweets.

Im Anschluss an die Datensammlung wurden alle Blogger:innen kontaktiert und hatten die Möglichkeit, der Verwendung der Daten zu widersprechen (Opt-Out-Verfahren). Erhielten wir keine Rückmeldung, wurde von Zustimmung ausgegangen und die Daten wurden verwendet, nicht kontaktierbare Personen wurden mit ihren Daten entfernt. Alle Texte wurden manuell anonymisiert, sodass Rückschlüsse auf Einzelpersonen nicht länger möglich sind. Dafür wurden Personen- und Ortsnamen, Kontaktdaten, Usernamen sowie Links durch Platzhalter ersetzt. Um sicherzustellen, dass sowohl Blogposts als auch Tweets einer Person zugeordnet werden können, erhielten alle Nutzer:innen eine anonymisierte Zahlenkennung. Das Korpus beinhaltet Blogposts und Tweets von 44 Elternblogger:innen, insgesamt 468 Blogposts (ca. 360.000 Token) und 81.440 Tweets (ca. 1,2 Mio Token). Das Korpus liegt als Textdokument, tokenisiert und in xml-Format vor. Es kann für wissenschaftliche Forschungen genutzt werden und ist online verfügbar.[1]

Die parallele Sammlung von Blogposts und Tweets ermöglicht eine Untersuchung der intra-Autor:innen- und inter-Medien-Variation. Unsere These ist es, dass die

1) http://staff.germanistik.rub.de/digitale-forensische-linguistik/forschung/textkorpus-sprachliche-variation-in-sozialen-medien/

Autor:innen ihre Sprachverwendung an das jeweilige Medium anpassen und sich somit Unterschiede in der Nutzung linguistischer Merkmale zwischen Blogs und Tweets einer Person finden lassen. Begründet wird die Wahl linguistischer Merkmale abgesehen von der persönlichen Präferenz durch die Affordanzen eines Mediums, die bestimmte Formen sprachlichen Ausdrucks überhaupt erst ermöglichen (Gibson, 2014). Dieser Aspekt spielt, genau wie weitere nicht-linguistische Faktoren, bei der Wahl des entsprechenden Registers bei der Kommunikation über soziale Medien eine essenzielle Rolle.

Diese These haben wir bisher v.a. anhand der Nutzung von Modalpartikeln und Intensivierern untersucht. Die Analyse dieser beiden linguistischen Phänomene ist im Kontext unseres Korpus von Interesse, da Intensivierer und Partikeln häufig der konzeptionellen Mündlichkeit zugeordnet werden (Lütten, 1979), Blogs und Tweets allerdings schriftliche Medien darstellen. Im Rahmen unserer bisherigen Untersuchungen wurden verschiedene Modalpartikeln und der Intensivierer so annotiert. Wir stellen fest, dass sowohl ein Unterschied in der Häufigkeit und Art der Nutzung dieser Ausdrücke zwischen den Medien erkennbar ist, aber dass einzelne Personen auch bestimmte idiosynkratische Nutzungspräferenzen haben.

Literatur:

Gibson, J. J. (2014). The Theory of Affordances. In *The Ecological Approach to Visual Perception*. Taylor & Francis Group.

Koch, P., & Oesterreicher, W. (2007). Schriftlichkeit und kommunikative Distanz. *Zeitschrift für germanistische Linguistik*, 35(3). https://doi.org/10.1515/zgl.2007.024

Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe—Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36, 15–43. https://doi.org/10.15496/publikation-20410

Lütten, J. (1979). Die Rolle der Partikeln *doch*, *eben* und *ja* als Konsensus-Konstitutiva in gesprochener Sprache. In H. Weydt (Hrsg.), *Die Partikeln der deutschen Sprache* (S. 30–38). de Gruyter.

Deutsche Gebärdensprache zugänglich und nutzbar machen: Das Öffentliche DGS-Korpus und das Digitale Wörterbuch der Deutschen Gebärdensprache

Reiner Konrad / Julian Bleicken / Calvin Khan / Amy Isard / Gabriele Langer / Anke Müller / Marc Schulder (Universität Hamburg)

Das DGS-Korpus-Projekt, ein Langzeitprojekt der Akademie der Wissenschaften in Hamburg, das am Institut für Deutsche Gebärdensprache und Kommunikation Gehörloser der Universität Hamburg durchgeführt wird, hat seit September 2019 ca. 50 Stunden annotierter Gespräche online gestellt. Das Referenzkorpus der DGS, für das 560 Std. Videomaterial zur Verfügung stehen, wird zum Projektende 2023 fertigge-

stellt, Vorabeinträge des korpusbasierten **Digitalen Wörterbuchs der Deutschen Gebärdensprache (DW-DGS)** sind bereits veröffentlicht.

Im Öffentlichen DGS-Korpus wird das Videomaterial in drei verschiedenen Portalen präsentiert, um den unterschiedlichen Bedarfen und Interessen der Nutzer:innen zu entsprechen (Jahn et al. 2018, Hanke et al. 2020): **MEINE DGS** für die Sprachgemeinschaft, die sich für die Kultur Gehörloser und deren Sprache interessiert, **MEINE DGS – annotiert** für alle, die ein bestimmtes Forschungsinteresse oder bereits Erfahrung mit annotierten Gebärdensprachdaten haben, und ab Release 4 **MEINE DGS – ANNIS**. Damit steht auch für die DGS ein in der Korpuslinguistik bekanntes Korpussuchwerkzeug zur Verfügung (Krause 2019, Krause et al. 2016). Bei der Auswahl der Gespräche lag der Schwerpunkt auf Inhalten, die die sozialen und kulturellen Aspekte der Sprachgemeinschaft widerspiegeln sowie die Erfahrungen Gehörloser als sprachliche Minderheit in einer hörenden Mehrheitsgesellschaft.

MEINE DGS präsentiert die Videos mit deutschen Untertiteln sowie einigen Filtern (nach Region, Altersgruppe und Gesprächsformaten) und einer Suchmöglichkeit (Schlagwörter). MEINE DGS – annotiert enthält Online-Transkripte mit Annotationen (Übersetzung, Glossen zur Lemmatisierung der darin enthaltenen Gebärden (Types) und Mundbilder bzw. Mundgestik), eine Types-Liste mit Tokens als KWIC-Konkordanz sowie einen Sachindex der Schlagwörter. Zu jedem Transkript können sowohl die Videos als auch die Annotations- und Metadaten heruntergeladen werden inkl. OpenPose-Daten für computergestützte automatisierte Auswertung (Schulder/Hanke 2020). Das Datenset als solches sowie alle darin enthaltenen Transkripte und Types-Einträge sind über DOIs eindeutig adressierbar. Mit MEINE DGS – ANNIS kann nach Gebärden (Glossen), deren Form (HamNoSys, Hanke 2004) und Kontext sowie nach Text (Übersetzung, Mundbild) gesucht werden. Die verschiedenen Suchkriterien können sowohl untereinander als auch mit Metadaten kombiniert werden.

Im DW-DGS werden die Gebärden als Videos und Micons (moving icons) repräsentiert. Die Datengrundlage der lexikographischen Analyse ist das gesamte DGS-Korpus (Langer et al. 2018a/b). Die Wörterbucheinträge fassen die Ergebnisse zusammen und geben den Nutzer:innen einen strukturierten Überblick u.a. zu Lesarten, Kollokationen und semantischen Präferenzen sowie Formvarianten und Variation nach Region und Altersgruppen. Hinzu kommen authentische Verwendungsbeispiele aus dem Öffentlichen DGS-Korpus, die die Gebärde im Kontext zeigen.

DW-DGS und Öffentliches DGS-Korpus sind in mehrfacher Weise miteinander verlinkt. Die Nutzer:innen können von den Informationen im Wörterbuch zu den Videos und annotierten Daten im Öffentlichen DGS-Korpus gelangen als auch von den Types des Öffentlichen DGS-Korpus zu den bereits vorhandenen Wörterbuch-Einträgen (Müller et al. 2020).

In der Präsentation zeigen wir die vielfältigen Zugriffsmöglichkeiten auf das Wörterbuch sowie die Inhalte von Release 4 des Öffentlichen DGS-Korpus, indem wir auf verschiedene Nutzungsszenarien eingehen, vom interessierten Laien zur DGS-Lehre, vom Studium zur Forschung, von DGS-Muttersprachler:innen zu DGS-Lernenden.

Links:

ANNIS: https://corpus-tools.org/annis/

DGS-Korpus-Projekt: http://dgs-korpus.de

Digitales Wörterbuch DGS – Deutsch http://dw-dgs.de/

MEINE DGS: http://meine-dgs.de

MEINE DGS – annotiert: http://ling.meine-dgs.de

Literatur:

Hanke, Thomas (2004): HamNoSys. Representing sign language data in language resources and language processing contexts. In: Streiter, Oliver / Vettori, Chiara (eds.): Proceedings of the LREC 2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication. Paris: ELRA, S. 1-6. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/04001.pdf (letzter Zugriff: 2022-02-18).

Hanke, Thomas / Konrad, Reiner / Jahn, Elena / Schulder, Marc (2020): Extending the Public DGS Corpus in Size and Depth. In: Efthimiou, Eleni / Fotinea, Stavroula-Evita / Hanke, Thomas / Hochgesang, Julie A. / Kristoffersen, Jette / Mesch, Johanna (eds.): Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. Language Resources and Evaluation Conference (LREC 2020), Marseille, 11-16 May 2020, S. 75-82. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/20016.pdf (letzter Zugriff: 2022-02-18).

Jahn, Elena / Konrad, Reiner / Langer, Gabriele / Wagner, Sven / Hanke, Thomas (2018): Publishing DGS corpus data: Different Formats for Different Needs. In: Bono, Mayumi / Efthimiou, Eleni / Fotinea, Stavroula-Evita / Hanke, Thomas / Hochgesang, Julie / Kristoffersen, Jette / Mesch, Johanna / Osugi, Yutaka (eds.): Workshop Proceedings. 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. Language Resources and Evaluation Conference (LREC), Miyazaki, Japan, 12 May 2018. Paris: ELRA, S. 83-90. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/18018.pdf (letzter Zugriff: 2022-02-18).

Krause, Thomas (2019): ANNIS: A graph-based query system for deeply annotated text corpora. Berlin (Diss., HU zu Berlin). DOI: 10.18452/19659.

Krause, Thomas / Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. Digital Scholarship in the Humanities, 31(1). pp. 118-139. DOI: 10.1093/llc/fqu057.

Langer, Gabriele / Müller, Anke / Wähl, Sabrina (2018a): Queries and Views in iLex to Support Corpus-based Lexicographic Work on German Sign Language (DGS). In Bono, Mayumi / Efthimiou, Eleni / Fotinea, Stavroula-Evita / Hanke, Thomas / Hochgesang, Julie / Kristoffersen, Jette / Mesch, Johanna / Osugi, Yutaka (eds.): Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 12.5.2018. Paris, Frankreich: European Language Resources Association (ELRA), S. 107-114. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/18026.pdf (letzter Zugriff: 2022-02-18).

Langer, Gabriele / Müller, Anke / Wähl, Sabrina / Bleicken, Julian (2018b): Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How. In: Čibej, Jaka et al. (Hrsg.):

Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (EURALEX 2018), Ljubljana, Slowenien, 17.-21.7.2018. Ljubljana, Slowenien: Ljubljana University Press, S. 483-497. URL: https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2933-1 (letzter Zugriff: 2022-02-18).

Müller, Anke / Hanke, Thomas / Konrad, Reiner / Langer, Gabriele / Wähl, Sabrina (2020): From Dictionary to Corpus and Back Again – Linking Heterogeneous Language Resources for DGS. In Efthimiou, Eleni / Fotinea, Stavroula-Evita / Hanke, Thomas / Hochgesang, Julie A. / Kristoffersen, Jette / Mesch, Johanna (eds.): Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. 12th International Conference on Language Resources and Evaluation (LREC 2020), Marseille, Frankreich, 16.5.2020. Paris, Frankreich: European Language Resources Association (ELRA), S. 157-164. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/20025.pdf (letzter Zugriff: 2022-02-18).

Schulder, Marc, & Hanke, Thomas. (2020, May 11). OpenPose in the Public DGS Corpus (Version 2). DOI: 10.25592/uhhfdm.1866.

Erhebung, Aufbereitung und (kollaborative) Nutzung des Korpus „Deutsch in Österreich. Variation – Kontakt – Perzeption“

Jakob Bal / Johanna Fanta-Jende / Katharina Korecky-Kröll / Katharina Kranawetter / Markus Pluschkovits / Rita Stiglbauer / Rebecca Stocker / Florian Tavernier / Anja Wittibschlager (Universität Wien)

Der Spezialforschungsbereich (FWF F60) „Deutsch in Österreich“ (im Folgenden „DiÖ“, siehe für einen Überblick Lenz 2018 oder Budin et al 2018) gliedert sich in drei wissenschaftliche Task-Cluster („Variation“, „Kontakt“, „Perzeption), die von einem technischen Task-Cluster („Forschungsplattform“) unterstützt werden. Auf Basis der kollaborativen Projektarbeit ist ein umfassendes Sprachkorpus entstanden, dessen Methoden und Tools im Rahmen der Tagung vorgestellt werden sollen. Im Fokus stehen dabei die linguistischen und technischen Zugänge zu Datenerhebung, -aufbereitung und -analyse sowie die Zugänglichkeit für projektexterne Anwendungen und die Übertragbarkeit auf zukünftige Forschungsvorhaben (alle Tools werden auf Open-Source-Basis entwickelt und stehen in den GitHub-Repositorien des DiÖ-Projekts zur Verfügung, siehe DiÖ 2021).

Die linguistische Grundlage der Präsentation stellt ein umfassendes gesprochensprachliches Korpus dar, das sich an den Kernthemen der Teilprojekte PP03 „Sprachrepertoires und Varietätenspektren“ und PP08 „Standardvarietäten aus Perspektive der perzeptiven Variationslinguistik“ orientiert. Zwischen 2016 und 2019 wurden ca. 570 Stunden Audiomaterial von rund 150 Personen unterschiedlichen Alters und Geschlechts aus ganz Österreich aufgezeichnet. Um möglichst viel über die inter- und intraindividuelle Variation in Österreich zu erfahren, wurden die Personen mit einem breiten Mix an Erhebungsmethoden konfrontiert: Kontrollierte Settings um-

fassten mündliche Übersetzungsaufgaben aus dem Dialekt in das Standarddeutsche und umgekehrt, multimodale Sprachproduktionsexperimente im Standard und im Dialekt sowie Bildbenennungs- und Vorleseaufgaben (mit Fokus auf dem standardnahen Pol). Freie Settings zielten auf die Spontansprache ab: Forscher*innen führten mit den Gewährspersonen soziolinguistische Interviews, um den standardnahen Spontansprachgebrauch sowie die Spracheinstellungen in einem formellen Kontext zu elizitieren. Demgegenüber sollten Freundesgespräche, in denen Gespräche der Gewährspersonen mit jeweils einer befreundeten Person aus demselben Ort erhoben wurden, dialektnahe spontansprachliche Register und Spracheinstellungen in einem informellen Setting abdecken.

Für den Umgang mit dieser heterogenen Datengrundlage wurden mehrere technische Tools im Rahmen des DiÖ-Projekts entwickelt. Die Transkriptionssoftware SpeechHouse (bisheriger Arbeitstitel transcribe, Graf et al. in Vorbereitung) ist ein plattformunabhängiges, multi-user taugliches und token-basiertes Transkriptionstool. Das Tool orientiert sich an den Prinzipien des User-Centered Designs (vgl. Chammas / Quaresma / Mont'Alvão 2015) und ermöglicht Nutzer*innen ein flüssiges, zeitaligniertes und gleichzeitiges Transkribieren von Sprachdaten. Es unterstützt außerdem die Transkription auf verschiedenen Tiers, um Ansprüchen unterschiedlicher linguistischer Systemebenen (z.B. phonetische Transkription, lautorientierte Transkription für Syntax usw.) nachzukommen. Darauf folgt die linguistische Annotation in der projekteigenen Datenbank, die ebenfalls auf die unterschiedlichen Forschungsinteressen abzielt.

Für den anschließenden Schritt der Datenauswertung soll ein sich in Entwicklung befindliches Kartierungstool (Alpha-Fassung mit Hilfe des Vue.js Framework, vgl. You 2020) vorgestellt werden. Dieses dient einerseits den Linguist*innen des DiÖ-Projekts als Werkzeug zur Visualisierung des annotierten georeferenzierten Sprachmaterials im Raum und soll andererseits einen niederschwelligen Zugang zum Korpus für ein interessiertes Lai*innenpublikum ermöglichen. Dabei steht der explorative Zugang im Vordergrund, der es Nutzer*innen erlauben soll, mit den Karten dynamisch zu interagieren und aktuelle Fragen der Variationslinguistik in Österreich auf Basis der regionalen Verortung und einer entsprechenden Verknüpfung zu den Audiobeispielen kennenzulernen.

Schließlich soll mit Abschluss des Projekts 2023 eine geplante Online-Forschungsplattform das gesamte Datenmaterial allen Interessierten öffentlich zugänglich gemacht werden. Die Präsentation bei der Methodenmesse bietet also einen Vorgeschmack auf die zu erwartenden umfassenden Möglichkeiten der Datenerfassung, -aufbereitung und -auswertung zum Deutschen in Österreich.

Literatur:

Budin, Gerhard, Stephan Elspaß, Alexandra N. Lenz, Stefan Michael Newerkla & Arne Ziegler (2018). Der Spezialforschungsbereich „Deutsch in Österreich (DiÖ). Variation – Kontakt – Perzeption". Zeitschrift für germanistische Linguistik 46(2). https://doi-org.uaccess.univie.ac.at/10.1515/zgl-2018-0017 [Abruf 12. September 2021].

Chammas, Adriana / Quaresma, Manuela / Mont'Alvão, Cláudia. 2015. „A Closer Look on the User Centred Design". Procedia Manufacturing 2015/3, 5397–5404.

DiÖ. 2021. Github Repositorien des Spezialforschungsbereich Deutsch in Österreich. O.O. https://github.com/german-in-austria (Abruf 12. September 2021).

Graf, Arnold, Ludwig Maximilian Breuer, Tahel Singer & Markus Pluschkovits. in Vorbereitung. Transcribe: a web-based linguistic transcription tool. O.O.

Lenz, Alexandra N. (2018). The Special Research Programme "German in Austria: Variation – Contact – Perception". Sociolinguistica 32(1), 269-277.

You, Evan (2020): Vue.Js. The Progressive JavaScript Framework. URL: https://vuejs.org [Abruf 1. Februar 2021].

Korpora zur Fußballlinguistik – Eine multilinguale Forschungsressource zur Sprache der Fußballberichterstattung

Simon Meier-Vieracker (TU Dresden)

Im Projekt *fussballlinguistik.de* werden seit einigen Jahren multilinguale Korpora mit Texten der internetbasierten Fußballberichterstattung wie etwa Spielberichte, Taktikanalysen und v.a. Liveticker aufgebaut und über eine webbasierte Plattform (www.fussballlinguistik.de/korpora) der Forschungsöffentlichkeit zugänglich gemacht (Meier 2017, 2020). Inzwischen kann in vollständig morphosyntaktisch annotierten Texten in vierzehn europäischen Sprachen (deutsch, Schweizer Standarddeutsch, englisch, norwegisch, niederländisch, französisch, italienisch, spanisch, portugiesisch, polnisch, tschechisch, russisch, griechisch, ungarisch) im Gesamtumfang von rund 78 Millionen Tokens recherchiert werden. Damit dokumentiert die Ressource in bislang einmaliger Weise den Schriftsprachgebrauch in der thematischen Domäne der Fußballberichterstattung. Vor allem die durch konzeptionelle Mündlichkeit gekennzeichnete, aber gleichwohl redaktionelle Textsorte Liveticker ist damit zum ersten Mal in größerem Umfang für korpuslinguistische Zwecke dokumentiert. Auch wenn die Fußballsprache seit vielen Jahren intensiv beforscht wird, mussten Forschende bislang entweder eigene Datensammlungen erstellen oder auf die thematisch unspezifischen Pressekorpora zurückgreifen, die natürlich auch Fußballtexte enthalten. Mit der neuen Ressource können aber viel fokussierter und zugleich auf einer breiten empirischen Basis die Eigenschaften dieser Sondersprache auf unterschiedlichen sprachlichen Ebenen untersucht werden.

Insbesondere die lexikologisch und phraseologisch orientierte Forschung zur Fußballsprache (Schmidt 2009; Burkhardt 2014), aber auch an syntaktischen Eigenheiten interessierte Forschung (Jürgens 2009), findet durch die Korpora eine nützliche empirische Basis, auch und gerade in sprachvergleichender Perspektive. Die Analysesoftware CQPweb (Hardie 2012) erlaubt nicht nur die bequeme Suche nach Belegen und ihrer kontextuellen Einbettung, sondern ermöglicht mit den implementierten statistischen Auswertungsmethoden wie Kollokationen und Keywords auch datengeleitete Explorationen. Die verfügbaren Metadaten, etwa zu Textsorte, Datum oder Wettbewerb, erlauben zudem die Erhebung von bedingten Verteilungen (Stefanowitsch 2020), so

dass auch diachrone Entwicklungen oder mediale Prägungen (etwa live vs. non-live) untersucht werden können.

Die Korpora zur Fußballlinguistik bieten aber nicht nur für korpuslinguistisch versierte Forschende eine nützliche Ressource, sondern können aufgrund der globalen Popularität des Fußballs auch in didaktischen Settings eingesetzt werden. In Frage kommen einerseits (fremd)sprachdidaktische Anwendungen, um die Besonderheiten der medialen Fußballsprache auch sprachkontrastiv zu vermitteln (Burkhardt 2010). Andererseits sind methodendidaktische Zugänge möglich, um am Beispiel der populären und lebensweltlich relevanten Fußballsprache korpuslinguistische Fragestellungen und Techniken zu vermitteln und *corpus literacy* (Mukerjee 2004) zu fördern. Ergänzt wird die Korpusressource durch ein Online-Tutorial (https://ali.gsw.tu-dresden.de/tutorial.html). Auf dem Blog *fussballlinguistik.de* werden zudem kleinere Forschungsergebnisse in popularisierender Form publiziert, welche die Nutzungsmöglichkeiten der Korpora veranschaulichen.

Aus Urheberrechtsgründen können in der Analysesoftware nur Ausschnitte der Texte angezeigt werden, ein Download der ganzen Datensätze ist nicht möglich. In einem GitHub-Repositorium werden jedoch Scraping-Skripte angeboten, mit denen ein Teil der Korpora *from the scratch* selbst erstellt werden kann, so dass dann auch avanciertere korpuslinguistische Zugänge wie etwa die Berechnung von Word Embedding-Modellen möglich sind.

Literatur:

Burkhardt, Armin (2010): Der deutsche Fußball und seine Sprache. Ein Beitrag zur Fach- und Sondersprachendidaktik im DaFUnterricht. In: Brünner Hefte zu Deutsch als Fremdsprache 3 (2), S. 149–166.

Burkhardt, Armin (2014): Vom Runden und vom Eckigen, in das es hinein muss. Zu den Bezeichnungen von Ball und Tor in der deutschen Fußballsprache. In: Muttersprache 2014 (4), S. 290–301.

Hardie, Andrew (2012): CQPweb — combining power, flexibility and usability in a corpus analysis tool. In: International Journal of Corpus Linguistics 17 (3), S. 380–409. doi:10.1075/ijcl.17.3.04har.

Jürgens, Frank (2009): Syntaktische Formen bei der Fußballberichterstattung. In: Burkhardt, Armin/Schlobinski, Peter (Hg.): Flickflack, Foul und Tsukahara: der Sport und seine Sprache. Mannheim: Dudenverlag. S. 160–174.

Meier, Simon (2017): Korpora zur Fußballlinguistik – eine mehrsprachige Forschungsressource zur Sprache der Fußballberichterstattung. In: Zeitschrift für germanistische Linguistik 45 (2), S. 345–349. doi:10.1515/zgl-2017-0018.

Meier, Simon (2020): Blogs, Bots & Co. – Public Humanities in den Sozialen Medien. In: Marx, Konstanze/Lobin, Henning/Schmidt, Axel (Hg.): Deutsch in Sozialen Medien. Interaktiv – multimodal – vielfältig. (= Jahrbuch des Instituts für Deutsche Sprache 2019). Berlin, Boston: De Gruyter. S. 359–362. doi:10.1515/9783110679885-021.

Mukherjee, Joybrato (2004): Bridging the Gap between Applied Corpus Linguistics and the Reality of English Language Teaching in Germany. In: Connor, Ulla/Upton, Thomas A. (Hg.): Applied Corpus Linguistics: A Multidimensional Perspective. Leiden: Brill. S. 239–250. doi:10.1163/9789004333772.

Schmidt, Thomas (2009): The Kicktionary – a multilingual lexical resource of football language. In: Multilingual FrameNets in Computational LexicographyMethods and Applications. Berlin, Boston: De Gruyter Mouton. doi:10.1515/9783110212976.1.101.

Stefanowitsch, Anatol (2020): Corpus linguistics: A guide to the methodology. (= Textbooks in Language Sciences). Berlin: Language Science Press.

LeiKo: Ein Vergleichskorpus für Leichte Sprache

Sarah Jablotschkin / Heike Zinsmeister (Universität Hamburg)

Mit dem Konzept „Easy-to-read" werden Teilsysteme natürlicher Sprachen bezeichnet, welche durch eine systematische Reduktion auf den Ebenen Lexik und Syntax entstehen und den Zugang zu geschriebenen Informationen für Erwachsene mit geringen Lesekompetenzen gewährleisten. Im Deutschen gibt es „Leichte Sprache", welche sich nach spezifischen linguistischen und typografischen Regeln richtet, und die weniger restringierte „einfache Sprache". Beide Varianten erhalten im akademischen sowie nicht-akademischen Diskurs vermehrt Aufmerksamkeit – nicht zuletzt dank der im Jahr 2009 in Deutschland ratifizierten UN-Behindertenrechtskonvention (UN-BRK). Diese fordert neben der baulichen Barrierefreiheit auch Barrierefreiheit in Bezug auf Informationen und Medien (vgl. United Nations 2006 Art. 21). Im Zuge der Implementierung der UN-BRK in deutsche Gesetzgebung wurden Regelwerke für Leichte Sprache entwickelt (BITV 2.0, 2011; BMAS 2014), welche wiederum zum Gegenstand linguistischer Beschreibungen wurden (Bredel & Maaß 2016).

Bisher ist jedoch nicht ausreichend validiert worden, inwiefern die vereinfachten Strukturen die Verständlichkeit tatsächlich verbessern und welche Auswirkungen sie auf übergeordnete linguistische Ebenen, wie zum Beispiel die Textebene, haben. Dass die stark regulierte und vereinfachte Syntax Einfluss auf Mittel der Textverknüpfung hat, wird in Beispiel (1) deutlich. Um eine Subordination zu vermeiden und trotzdem die Anordnung *Folge – Ursache* zu erlauben, wird hier für die kausale Verknüpfung der Propositionen der Adverbkonnektor *nämlich* eingesetzt, welcher im Standarddeutschen eher selten ist (Breindl & Walter 2009: 39) und aufgrund seiner funktionalen/pragmatischen Eigenschaften nicht in allen Kontexten mit anderen Kausalkonnektoren austauschbar ist.

(1) Am Flughafen von Hannover sind Flüge ausgefallen.
Eine Start·bahn vom Flughafen war nämlich kaputt. (NDR 2018)

Zur Bearbeitung solcher Desiderata tragen wir mit LeiKo bei, einem Vergleichskorpus mit Nachrichtentexten in Leichter und einfacher Sprache. Die frei verfügbare Ressource wurde systematisch zusammengestellt sowie linguistisch annotiert und kann für linguistische und computerlinguistische Forschung genutzt werden. LeiKo besteht aus 216 Nachrichten- und Zeitungstexten (50.452 Tokens) und ihren Metadaten. Entsprechend der Websites, auf denen die Texte ursprünglich veröffentlicht wurden und die sich teilweise an unterschiedlichen Regelwerken orientieren, besteht LeiKo aus vier Subkorpora. Alle Texte sind tokenisiert, lemmatisiert, wortarten- und dependenz-

annotiert. Zusätzlich sind die Texte mit Koreferenzannotationen angereichert. Ein Pilotkorpus von 40 Texten wurde manuell korrigiert und mit Diskursrelationen gemäß den Penn-Discourse-Treebank-Guidelines (PDTB Research Group 2008, Webber et al. 2019) sowie typografischen und textstrukturellen Annotationen angereichert. Ein Großteil der Annotationen ist bereits in ANNIS (Krause & Zeldes 2016) durchsuch-bar.

Auf der Methodenmesse beschreiben wir die Subkorpora von Leiko mit Annotationsstatistiken und präsentieren zwei Beispielstudien, die mögliche Anwendungsbereiche demonstrieren. Das Korpus ist frei verfügbar auf Zenodo (DOI:10.5281/zenodo.3626763).

Literatur:

Bredel, Ursula & Christiane Maaß (2016): Leichte Sprache. Theoretische Grundlagen. Orientierung für die Praxis. Mannheim: Duden-Verlag.

Breindl, Eva & Maik Walter (2009): Der Ausdruck von Kausalität im Deutschen. Eine korpusbasierte Studie zum Zusammenspiel von Konnektoren, Kontextmerkmalen und Diskursrelationen. Mannheim: Institut für Deutsche Sprache.

Bundesministerium für Justiz und Verbraucherschutz (2011): Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung -- BITV 2.0). Zuletzt geändert: 21.05.2019.

Bundesministerium für Arbeit und Soziales (BMAS) (2014): Leichte Sprache. Ein Ratgeber. In Zusammenarbeit mit dem Netzwerk Leichte Sprache.

Krause, Thomas & Amir Zeldes (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: Digtial Scholarship in the Humanities 2016 (31).

NDR (2018): Beton kaputt: Flugzeuge könne nicht starten. Nachricht vom 25.07.2018. URL: https://www.ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache/Beton-kaputt-Flugzeuge-koennen-nicht-starten,flughafenhannover154.html

PTDB Research Group (2008): The PDTB 2.0 Annotation Manual. Institute for Research in Cognitive Science, University of Pennsylvania.

United Nations (2006): Übereinkommen über die Rechte von Menschen mit Behinderungen.

Webber, Bonnie, Rashmi Prasad, Alan Lee & Aravind Joshi (2019): The Penn Discourse Treebank 3.0 Annotation Manual.

Das SALAMMBO-Korpus: experimentelle Sprach-, Atmungs- und Bewegungsdaten für die Untersuchung von sprachlichen Effekten körperlicher Aktivität

Heather Weston / Susanne Fuchs
(Leibniz-Zentrum Allgemeine Sprachwissenschaft)

Gesprochene Sprache findet in nahezu allen Lebensbereichen statt. Welche Wechselwirkung gibt es zwischen Sprechen, Atmungs- und Bewegungsabläufen? Das neu erstellte Korpus leistet einen Beitrag zum Verständnis von gesprochener Sprache im Rahmen von körperlichen Tätigkeiten, indem dieser Frage experimentell nachgegangen wird. Neben akustischen Sprachdaten beinhaltet das Korpus zusätzlich weitere Datentypen: Atmungsdaten (mit Respiratory Inductance Plethysmography) und Bewegungsdaten (mit Motion Capture).

Insgesamt wurden ca. 17 Stunden gesprochener Sprache von 48 weiblichen Muttersprachlerinnen des Standarddeutschen aufgenommen. Im Korpus sind drei Sprachaufgaben erfasst: angehaltene Vokale, gelesener Text und spontansprachliche Äußerungen. Angehaltene Vokale werden als Untersuchungsgegenstand und Diagnostikum in der klinischen Forschung zu funktionalen Stimmstörungen eingesetzt. Gelesene Texte ermöglichen detaillierte Analysen stimmlicher und sprecherischer Veränderungen bei gleichbleibendem Sprachmaterial. Die spontansprachlichen Äußerungen erlauben zusätzlich die Untersuchung von linguistischen Strukturen und (Atem-)Pausensetzung. Diese Sprachaufgaben wurden von den Teilnehmerinnen unter verschiedenen körperlichen Belastungen durchgeführt: neben einer Kontrollbedingung ohne Belastung wurden zwei experimentelle Bedingungen mit je leichter und moderater körperlicher Anstrengung aufgenommen.

Das Korpus wurde in erster Linie konzipiert, um phonetischen Fragen nachzugehen. Gleichzeitig dienen die Sprachaufgaben, insbesondere die Spontansprache, der Untersuchung diverser Fragestellungen aus anderen sprachwissenschaftlichen Disziplinen. Die aufbereiteten Daten, die im Rahmen des DFG-Projektes SALAMMBO („Spoken Language in Motions") erhoben wurden, werden über die CLARIN-D Infrastruktur verfügbar gemacht.

In diesem Beitrag werden die Eckdaten des SALAMMBO-Korpus präsentiert, insbesondere wie Atmungs- und Bewegungsdaten in die linguistische Analyse eingebunden werden können. Des Weiteren wird auf die Nachnutzung der experimentellen Daten eingegangen, z.B. werden verschiedene Szenarien für die Sekundäranalyse aufgezeigt.

Das Referenzkorpus Altdeutsch. Ein tiefenannotiertes Mehrebenenkorpus für Forschung und Lehre

Stephanie Jandt / Gohar Schnelle / Phuong Thao Tran
(Humboldt-Universität zu Berlin)

Das Referenzkorpus Altdeutsch (ReA, aktuelle Version 1.1, Donhauser 2015) ist eintiefenannotiertes Korpus der gesamten althochdeutschen und altsächsischen Textüberlieferung und als diachrones Sprachstufenkorpus Teil des Verbunds ‚Deutsch Diachron Digital – Referenzkorpora zur deutschen Sprachgeschichte'. Die einzelnen Textüberlieferungen sind auf Basis handschriftennaher Editionen in Form von Subkorpora oder – bei kleineren Texten – in Form von Dokumenten repräsentiert, wobei die zahlreichen lateinabhängigen Texte (bspw. Tatian, Isidor, Murbacher Hymnen) gemeinsam mit der lateinischen Vorlage aufgenommen wurden. Im ReA werden insgesamt 24 Subkorpora repräsentiert, die ungefähr 535.000 Token[1] umfassen. Die Gesamtanzahl setzt sich u.a. aus den Token der althochdeutschen Texte (ca. 241.000 Token[2]) und der lateinischen Textgrundlagen (ca. 132.000 Token[3]) zusammen. Als tiefenannotiertes Mehrebenenkorpus enthält das ReA eine Primärtextebene („edition") und eine Normalisierungsebene („text"), die durch eine Segmentierung der historischen Schreibung eine Grundlage für die weiteren linguistischen Annotationen bildet (siehe Abbildung 1).

edition	nu	noh	uuerchum	deononte	uueraltí	kelaubuliugan		cote
text	nu	noh	uuerchum	deononte	uueraltí	kelaubu	liugant	cote
lemma	nu	noh	werk	dionon	weralt	gilouba	liogan	got
posLemma	ADV	ADV	NA	VV	NA	NA	VV	NA
pos	ADV	ADV	NA	VVPS	NA	NA	VVFIN	NE
inflectionClassLemma			A_NEUT	WK2	I_FEM	O_FEM	ST2A	A_MASC
inflectionClass			A_NEUT	WK2	I_FEM	O_FEM	ST2A	A_MASC
inflection			PL_DAT	P_POS_MASC_PL_NOM_ST	SG_DAT	SG_DAT	IND_PRES_PL_3	SG_DAT
lang	goh	goh	goh	goh	goh	goh	goh	goh
clause	CP					CF_U_M		
document	(Althochdeutsche) Benediktinerregel							
chapter	1							
translation	nun	noch	Werk, Tat	(be)dienen, aufwarten	Zeit; Zeitalter, Welt	Glaube(nsbekenntnis)	lügen	Gott

Abbildung 1: Getrenntschreibung auf text-Ebene von nicht-lemmakonformer Zusammenschreibung auf **edition**-Ebene. Die Ebene **posLemma** und **inflectionClassLemma** enthalten Information zu den Lemmata *gilouba* und *liogan*. Die Annotation auf der Ebene **pos**, **inflectionClass** und inflection enthalten Analysen zum Beleg im Gebrauchskontext.

Den Kern der grammatischen Annotationen bilden lexikalische („lemma", „translation", „lang", „posLemma"), morphologische („inflectionClassLemma", „inflection"), syntaktische („clause", „pos") und textstrukturelle („document", „verse", „chapter", „line", „rhyme") Ebenen. Relevante Informationen der jeweiligen Texte werden in den Metadaten annotiert: zeitliche und dialektale Einordnung, Textform und inhaltliche Einordnung, Verhältnis zu einer lateinischen Vorlage u.e.m.

1) https://korpling.org/annis3/?id=1a8aee45-fc94-4d04-8c09-d05470d9b63a
2) https://korpling.org/annis3/?id=38c73ca5-3d49-46a7-8475-5f87ace1cd96
3) https://korpling.org/annis3/?id=9da36a0a-cedc-431e-93a9-225337f95444

Eine Besonderheit des Korpus ist der komplementäre analytische Bezug der grammatischen Annotationsebenen: „posLemma“ und „inflectionClassLemma“ enthalten Informationen zum Lemma, „pos“, „inflectionClass“, „inflection“ beziehen sich auf den Gebrauch des Lemmas im Kontext (siehe Abbildung 2). Durch die Verknüpfung lemmabezogener und auf den belegbezogenen Annotationen kann gebrauchsbasierte Variation leicht identifiziert werden, z.B. Substantivierungen. Das besondere Potential, das diese analytischen Ebenen für die forschungsrelevanten Suchanfragen liefern, möchten wir gern in unserem Videobeitrag präsentieren.

edition	thorfti	,	sô	thar	êr	ênig	thegno
text	thorfti	,	sô	thar	êr	ênig	thegno
lemma	thurban		so	thar	er	enig	thegan
posLemma	VM	$,	KO	ADV	ADV	ADJ	NA
pos	VMFIN	$,	KOUS	ADV	ADV	NA	NA
inflectionClassLemma	PRPR					A,O	A,I_MASC
inflectionClass	PRPR					A,O	A_MASC
inflection	SUBJ_PAST_SG_3					POS_MASC_SG_NOM_0	PL_GEN
lang	osx		osx	osx	osx	osx	osx
clause	CF_I_O		CF_I_Adv_Mod				
document	Heliand						
chapter	III						
verse	178						
rhyme	S1b						S2
translation	Veranlassung haben, brauchen, dürfen		wie	da, dort	vor, eher, früher	irgendein(er)	der Mann des Gefolges, Dienstmann

Abbildung 2: Auf der **posLemma**-Ebene wurde das Lemma *enig* Adjektiv annotiert. Diese Information bezieht sich auf das zugewiesene Lemma. Aus dem gebrauchbasierten Kontext auf der text-Ebene hat *enig* auf der **pos**-Ebene den Wert eines Nomen Appelativum.

Das ReA wird als Ressource bereits vielseitig in der Forschung (u.a. Catasso 2021, Coniglio et al. 2021, Flick 2020, Jäger 2016, Luxner 2020, Mittmann 2020, Petrova 2020, Schnelle 2018, Zeige 2019) und in der sprachhistorischen universitäten Lehre (Schnelle 2019, Schnelle et al. 2021) genutzt. So stellt z.B. das ‚Lesekorpus Altdeutsch‘ (Mitmann & Plate 2019) eine auf philologische Textarbeit ausgerichtete Visualisierung bereit. Das seit Juli 2021 laufende Projekt „D4“ (Deutsch-Diachron-Digitale Didaktik, Gohar Schnelle) erstellt auf Basis des Referenzkorpus Lern- und Forschungsmaterialien für die Universität, darunter Tutorials für die Korpusnutzung, Anleitungen für die Einbindung von Korpora in den sprachhistorischen Unterricht sowie Visualisierungen empirischer Daten (siehe Abbildung 3).

Das ReA ist auf dem Laudatio-Repositorium (Guescini & Odebrecht 2020) veröffentlicht und archiviert. Alle Referenzkorpora des DDD-Verbunds sind über die Such- und Visualisierungs-plattform ANNIS (Krause 2019) weltweit kostenfrei und browserbasiert zugänglich. Annotation und Abfrage folgen gemeinsamen Standards (z.B. HiTS Dipper et al 2013). Die Plattform bietet eine Abfrage einfacher und verschränkter Suchparameter, eine transparente Visualisierung der Suchergebnisse eine Editionsansicht der Texte, sowie Analyse und Exportfunktionen.

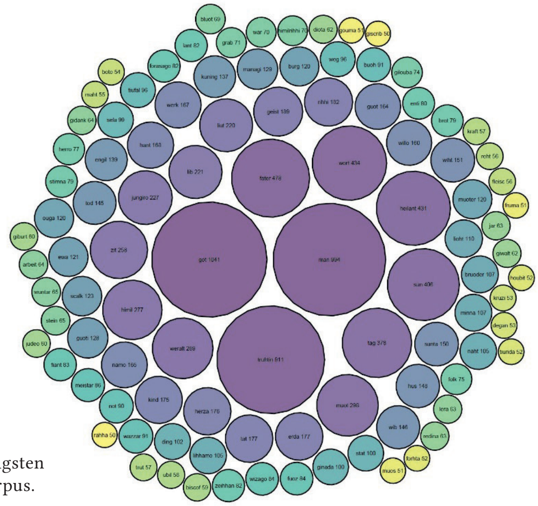

Abbildung 3: Wordbubbles der 100 häufigsten Nomen und deren Frequenz im ReA-Korpus.

Durch die Tiefenannotation und die umfassenden Metadaten bietet das RA einen spannenden Einblick in die Variabilität und Wandelbarkeit sprachlicher Strukturen und gemeinsam mit den Referenzkorpora des DDD-Verbunds ein Fenster in die deutsche Sprachgeschichte. Darüber hinaus macht es auch die Bearbeitung zahlreicher empirischer Forschungsfragen aus Sprach- und Literaturwissenschaft und anderen textbezogenen Wissenschaften möglich.

Literatur:

Catasso, Nicholas (2021): How theoretical is your (historical) syntax? Towards a typology of Verb-Third in Early Old High German. Journal of Comparative Germanic Linguistics 24/1, S. 1–48.

Coniglio, Marco; C. De Bastiani; R. Hinterhölzl; T. Weskott (2021): In the right mood, in the right place: on mood and verb placement in Old Germanic subordinate clauses. Journal of Historical Syntax 5, S.1-27.

Dipper, Stefanie; Karin Donhauser; Sonja Linde; Stefan Müller; Klaus-Peter Wegera (2013): "HiTS: Ein Tagset Für Historische Sprachstufen Des Deutschen." Journal for Language Technology and Computational Linguistics 28 (1), 85–137.

Donhauser, Karin (2015): Das Referenzkorpus Altdeutsch. Das Konzept, die Realisierung und die neuen Möglichkeiten. In: Jost Gippert und Ralph Gehrke (Hg.): Historical Corpora. Challenges and Perspectives. Tübingen: Narr,S 35–49.

Donhauser, Karin; Gippert, Jost; Lühr, Rosemarie: ddd-ad. Version 1.1. Online verfügbar unter http://hdl.handle.net/11022/0000-0007-C9C7-6 .

Flick, Johanna. 2020. Die Entwicklung des Definitartikels im Althochdeutschen Eine kognitiv-linguistische Korpusuntersuchung. Language Science Press. (Empirically Oriented Theoretical Morphology and Syntax 6)

Guescini, Rolf; Odebrecht, Carolin (2020): Laudatio Repository - Long-term Access and Usage of Deeply Annotated Information: Digitale Medien, Computer- und Medienservice der Humboldt-Universität zu Berlin.

Jäger, Agnes (2016): Vergleichskasus im Althochdeutschen. In: Neri, Sergio/Schuhmann, Roland/Zeilfelder, Susanne (Hrsg.): „dat ih dir it nu bi huldi gibu". Linguistische, germanistische und indogermanistische Studien. Rosemarie Lühr gewidmet. Wiesbaden: Reichert, S. 193–208.

Krause, Thomas (2019): ANNIS: A graph-based query system for deeply annotated text corpora. Dissertation. Humboldt-Universität zu Berlin, Berlin.

Luxner, Bernhard (2020): Die althochdeutschen Adjektive auf -aht(i)/-oht(i). Eine diachron- und synchron-vergleichende Untersuchung. Dissertation. Friedrich-Schiller-Universität Jena

Mittmann, Roland (2020): Zur althochdeutschen Zeit- und Dialektgliederung. Eine computergestützte Untersuchung auf Grundlage der textlichen Überlieferung. Hamburg: Baar. (Studien zur historisch-vergleichenden Sprachwissenschaft 15)

Mittmann, Roland; Ralf Plate (2019): Das ‚Referenzkorpus Altdeutsch' als Lesekorpus. Grammatisch annotierte und mit Wörterbüchern verknüpfte Texte für Lehre und Selbststudium, in: Bleier, Roman/Fischer, Franz/Hiltmann, Torsten/Viehhauser, Gabriel/Vogeler, Georg (Hg.): Digitale Mediävistik (Das Mittelalter 24/1), 173–187. https://titus.uni-frankfurt.de/lea/

Petrova, Svetlana (2020): Embedded V2 in Old High German. A corpus study. In Rethinking Germanic V2, Rebecca Woods und Sam Wolfe (Hgg.), 555–573. Oxford: Oxford University Press.

Schnelle, Gohar (2019): Corpus based teaching of Old High German. An experience report. Digital Approaches to Teaching Historical Languages (DAtTeL), Humboldt Universität zu Berlin, Jacob-und-Wilhelm-Grimm-Zentrum – Auditorium, Berlin, March 28-29, 2019. Vortrag.

Schnelle, Gohar (2018): Funktional bedingte Variation in der Evangelienharmonie Otfrids von Weißenburg. Eine methodische Annäherung an eine variationistische korpusbasierte Registerstudie des Althochdeutschen. Masterarbeit an der Humboldt-Universität zu Berlin. https://doi.org/10.18452/19250

Schnelle, Gohar; Svetlana Petrova; Birgit Herbers (2021): Das Referenzkorpus Altdeutsch (ReA) in der universitären Lehre. Sprachgeschichte und Bildung. 13. Jahrestagung, Pädagogische Hochschule Tirol, 20. bis 22. September 2021. Vortrag.

Zeige, Lars Erik (2019): Die Präpositionalgruppe im ältesten Deutsch. Korpus, Grammatik, Darstellung. 2. Bd. Habilitationsschrift, Humboldt-Universität zu Berlin.

Das *Berlin Dialogue Corpus* (BeDiaCo) und das *Corpus of Non-Native Addressee Register* (CoNNAR) – themen- und aufgabenorientierte spontansprachliche Dialoge in direkter und videobasierter Kommunikation

Malte Belz / Bianca Sell / Robert Lange / Megumi Terada / Christine Mooshammer / Anke Lüdeling (Humboldt-Universität zu Berlin)

1. Forschungsgegenstand

Das *Berlin Dialogue Corpus* enthält spontane gesprochensprachliche Dialoge in verschiedenen Situationen und dient der Grundlagenforschung in Bezug auf phonetisch-phonologische, morphologische, syntaktische, prosodische, pragmatische und interaktionslinguistische Phänomene spontaner gesprochener Sprache. Auch Forschung zu intra- und interindividueller Variation und registerspezifischen Fragestellungen sind möglich.

Das *Corpus of non-native addressee register* (CoNNAR) wurde mit der Fragestellung entwickelt, ob Sprecher*innen sich in ihrem Sprechstil an die tatsächlichen oder angenommenen Bedürfnisse ihrer Gesprächspartner*innen anpassen (Bell, 1984), wenn sie mit erwachsenen Lerner*innen kommunizieren (andere Begriffe sind *foreigner talk*, *foreigner-directed speech* oder Xenolekt, vgl. Bradlow and Bent, 2002; Roche, 1998). Auch dieses Korpus kann für weitere Fragestellungen herangezogen werden.

2. Korpus BeDiaCo

BeDiaCo v.2 (Belz et al. 2021a) enthält 150.000 Token von 36 Sprecher*innen in zwei Subkorpora mit insgesamt 11,2 h spontansprachlicher Dialoge (reine Artikulationszeit) und vorgelesener Wortlisten. Das Subkorpus BeDiaCo-main ($BeDiaCo_m$) enthält 46.000 Token von 16 SprecherInnen (10 männlich, 6 weiblich) zwischen 18 und 31 Jahren ($\bar{x}$ = 24,1, s = 4,2) mit 3 h Artikulationszeit, wobei sich die GesprächspartnerInnen vorher nicht kannten und in acht *face-to-face*-Dialogen miteinander sprechen. Das Subkorpus BeDiaCo-videocall ($BeDiaCo_v$) enthält 104.000 Token von 20 SprecherInnen (10 männlich, 10 weiblich) zwischen 19 und 32 Jahren ($\bar{x}$ = 25,7, s = 3,8) mit 8 h Artikulationszeit, die einander kennen (Mitbewohner*innen, Geschwister, Partner*innen/Eheleute). Sie wurden in zwei Bedingungen durchgeführt: face-to-face und mit dem Videokonferenztool Zoom.

In beiden Subkorpora wird der freie Dialog themenorientiert mit einer Frage des Experimentleiters nach der Qualität des Mensaessens ($BeDiaCo_m$), oder dem Thema Berlin bzw. Traumreise ($BeDiaCo_v$) begonnen und dauert zwischen 10 und 15 min, wobei die Versuchspersonen schnell dazu übergehen, über selbstgesetzte Themen zu sprechen. Die aufgabenbasierten Dialoge nutzen ins Deutsche übersetzte Diapix-Aufgaben (Baker & Hazan 2011, Van Engen et al. 2010) mit zwei fast identischen Bildern, in welchen die Versuchspersonen Unterschiede finden müssen. Zusätzlich sind Wortlisten für alle 36 Versuchspersonen enthalten, in denen ein Trägersatz *Sage X bitte* ein zweisilbiges Wort X enthält, welches auf der ersten Silbe betont ist. Die Wörter ent-

halten alle Monophthonge des Deutschen und die beiden Reduktionsvokale [ə ɐ] in der zweiten unbetonten Silbe. Das Experiment ist folgendermaßen aufgebaut: Wortliste, Diapix, freier Dialog, Diapix, Wortliste. Die Versuchspersonen saßen in einer schallisolierten Kabine (mit Hörsprechgarnitur *beyerdynamics Opus 54*) bzw. in der Zoombedingung (mit Kopfhörern und Stabmikrofonen) zusätzlich in einem Büro, welches an das Labor angrenzt. Das Experiment dauerte maximal eine Stunde und wurde mit 10 bzw. 11 Euro kompensiert.

2.1 Annotation

Die Annotation ist in einem Mehrebenenmodell aufgebaut, mit dem akustischen Signal aligniert und ausführlich dokumentiert (Belz et al. 2021b). Enthalten sind manuell erstellte und automatisch alignierte Ebenen (z.B. die Transliteration mithilfe von WebMAUS, vgl. Kisler et al. 2017), automatisch erstellte Ebenen (z.B. Normalisierung, Lemmatisierung und Wortarten) sowie manuell erstellte und alignierte Ebenen (z.B. zu Füllpartikeln und Intonationsphrasen). Die Annotation erfolgt in Praat (Boersma & Weenink 2019).

2.2 Zugang und Wiederverwendung

Die pseudonymisierten Daten sind für wissenschaftliche Zwecke über das Medienrepositorium der Humboldt-Universität zu Berlin verfügbar. Das Korpus kann bspw. mit *emuR* (Winkelmann et al. 2018) in R analysiert werden. Eine solche vorerstellte Datenbank wird in einer neuen Version zur lokalen Verwendung bereitgestellt. Das Korpus wird für die Arbeit an neuen Forschungsfragen kontinuierlich verbessert und kollaborativ mit neuen Ebenen annotiert. Dies erfolgt über eine vom Computer- und Medienservice der Humboldt-Universität zu Berlin bereitgestellten Versionierungssoftware (GitLab). Weitere Annotationsebenen von Dritten können nach Rücksprache mit den HerausgeberInnen in das Korpus aufgenommen werden.

3. Korpus CoNNAR

CoNNAR besteht aus zwei Subkorpora, die insgesamt 40 Versuchspersonen sowie 16 instruierte Confederates enthalten. Die Versuchspersonen (L1 Deutsch, 20 männlich, 20 weiblich, zwischen 18 und 40 Jahren) durchlaufen das Experiment jeweils zweimal – einmal mit einem*r Nicht-Muttersprachler*in (L1 Englisch, Proficiency in Deutsch B1/B2 oder C1, Alter zwischen 18 und 40 Jahren) und einmal mit einem*r deutschen Muttersprachler*in, der*die dem*der anderen Confederate in Alter und Geschlecht entspricht. Das Experiment dauert 60–90 Minuten und wird mit 11 Euro kompensiert.

Das Subkorpus CoNNAR-videocall ($CoNNAR_v$, ca. 110.000 Token) besteht aus 20 Versuchspersonen (je 10 weiblich und männlich, Alter 20–38, $\bar{x}$ = 26, s = 4,5) und 8 Confederates (je 4 weiblich und männlich, Alter 20–27, $\bar{x}$ = 22,9, s = 2, 4 Muttersprachler*innen und 4 Lerner*innen), die in getrennten Räumen (die VP in einer schallisolierten Kabine, der Confederate im angrenzenden Büro) sitzen und über das Videokonferenztool Zoom verbunden sind. Zoom wurde nur zur Kommunikation verwendet, die Audioaufnahmen erfolgte über Stabmikrofone. Das Experiment dauert jeweils etwa eine Stunde und wurde mit 11 € vergütet. Das Experimentdesign besteht aus Wortliste, freie Konversation, Diapix, Diapix, Wortliste, und ist damit ähnlich zu BeDiaCo auf-

gebaut, mit kleinen Unterschieden. In der freien Konversation unterhalten sich die Teilnehmer*innen 8 Minuten lang über ein Thema ihrer Wahl. In der Diapixaufgabe sollen die Teilnehmer*innen in 8 Minuten möglichst viele der 12 Unterschiede finden. Der Confederate bittet bei dem zweiten Diapix nach 3–4 gefundenen Unterschieden um eine klarere Aussprache.

Das Subkorpus CoNNAR-face-to-face ($\text{CoNNAR}_{\text{f2f}}$) befindet sich in der Erhebung. Geplant sind auch hier 20 Versuchspersonen und 8 Confederates, die gemeinsam in einer schallisolierten Kabine sitzen. Das Experimentdesign wird außerdem erweitert um eine Bildbeschreibungsaufgabe (die Versuchspersonen sehen ein Bild und lesen eine Bildbeschreibung vor bzw. beschreiben es selbst in einem Satz so, dass der*die Confederate das passende Bild aus 4 ähnlichen Bildern auswählen kann) und eine kognitive Belastung (vor den Diapixaufgaben sehen die Versuchspersonen jeweils eine 3×3-Matrix mit unterschiedlich komplex angeordneten Punktmustern, die nach der Diapixaufgabe reproduziert werden muss, vgl. De Neys and Schaeken, 2007).

3.1 Aufbereitung und Annotation

Die erhobenen Daten werden in Praat (Boersma and Weenink, 2019) transliteriert und in WEB-Maus (Kisler et al., 2017) um segmentale Annotationen erweitert. Die Annotation ist in einem Mehrebenenmodell aufgebaut und mit dem akustischen Signal aligniert. Mithilfe eines selbst erstellten Skripts werden Normalisierungen und POS-Tags ergänzt.

3.2 Zugang und Wiederverwendung

Die pseudonymisierten Daten werden nach Abschluss von Version 1 für wissenschaftliche Zwecke zur Verfügung gestellt. Das Korpus kann für die Arbeit an neuen Forschungsfragen kontinuierlich verbessert und kollaborativ mit neuen Ebenen annotiert werden (GitLab). Weitere Annotationsebenen von Dritten können nach Rücksprache mit den HerausgeberInnen in das Korpus aufgenommen werden.

Acknowledgements

Wir danken Melina Pfundstein, Alina Zöllner und Lea-Sophie Adam für die Hilfe bei der Erstellung und Annotation von BeDiaCo. BeDiaCo wurde gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1412, 416591334 und die Medienkommission des Akademischen Senats der Humboldt-Universität zu Berlin. CoNNAR wird wurde gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1412, 416591334.

Literatur:

Baker, Rachel & Valerie Hazan. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods* 43(3). 761–770. https://doi.org/10.3758/s13428-011-0075-y.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2).145–204.

Belz, Malte, Christine Mooshammer, Alina Zöllner & Lea-Sophie Adam. 2021a. *Berlin Dialogue Corpus (BeDiaCo): Version 2*. Medien-Repositorium. https://rs.cms.hu-berlin.de/phon.

Belz, Malte, Alina Zöllner, Megumi Terada, Robert Lange, Lea-Sophie Adam & Bianca Sell. 2021b. *Dokumentation und Annotationsrichtlinien für das Korpus BeDiaCo v2*. https://doi.org/10.5281/zenodo.4593351.

Boersma, Paul & David Weenink. 2019. *Praat: doing phonetics by computer [Computer program]*. http://www.praat.org/.

Bradlow, Ann R & Tessa Bent. 2002. The clear speech effect for non-native listeners. *J. Acoust. Soc. Am.* 112(1). 13.

De Neys, Wim & Walter Schaeken. 2007. When People Are More Logical Under Cognitive Load: Dual Task Impact on Scalar Implicature. *Experimental Psychology* 54(2). 128–133. https://doi.org/10.1027/1618-3169.54.2.

Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347. https://doi.org/10.1016/j.csl.2017.01.005.

Roche, Jörg. 1998. Variation in Xenolects (Foreigner Talk). *Sociolinguistica* 12(1). 117–139.

Van Engen, Kristin J., Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim & Ann R. Bradlow. 2010. The Wildcat Corpus of Native-and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles. *Language and Speech* 53(4). 510–540. https://doi.org/10.1177/0023830910372495.

Winkelmann, Raphael, Klaus Jaensch, Steve Cassidy & Jonathan Harrington. 2018. *emuR: Main Package of the EMU Speech Database Management System*.

Die Korpora IS, ISW und ISZ der AGD/DGD

Anne Betten (Universität Salzburg) / Carolina Flinz (Universität Mailand) / Simona Leonardi (Universität Genua)

Die Korpora IS, ISW und ISZ der AGD/DGD enthalten heute 216 Interviews mit 209 deutschsprachigen Emigranten der 1. Generation in Israel (IS und ISW) und 100 Interviews mit 66 Sprechern der 2. Generation (ISZ) – insgesamt ca. 517 Stunden, aufgenommen zwischen 1989 und 2019 von Anne Betten und weiteren Interviewerinnen. Ursprüngliches Ziel war die Dokumentation der Bewahrung eines schriftnahen **Bildungsbürgerdeutsch** („Weimarer Deutsch"), noch rund 60 Jahre nach der Emigration.

Die Aufnahmen wurden als **spontane narrative autobiographische Interviews** durchgeführt (meist mit einer Person, gelegentlich auch mit Ehepartnern oder Freunden). **Thematisch** bieten sie u.a. Material für Oral History (u.a. Farges 2020) und deutsch-jüdische Kulturgeschichte, **linguistisch** für ein breites Spektrum an Untersuchungsinteressen, z.B.:

Neuere Sprachgeschichte: Hochsprachlicher und regionaler Wortschatz bzw. regionale dt. Umgangssprachen zu Beginn des 20. Jhs., inklusive Pommern, Ostpreußen, Schlesien, Böhmen, Mähren, Bukowina u.a.m.; ferner Sprachkritik, v.a. anhand metasprachlicher Kommentare zum heutigen Deutsch.

Migrationslinguistik: Interferenz und Zusammenspiel mit den neuen Sprachen (Deutsch – Hebräisch – Englisch) in der 1. Generation, Zusammenhänge von Sprachkompetenz und sozial-psychologischen Faktoren v.a. in der 2. Generation (s. Betten 2016a).

Gesprächs- und Erzählanalyse: Bisher standen Untersuchungen von Textsorten, Interaktion, Emotion im Vordergrund (vgl. Leonardi et al. 2016, Betten 2017). Erzähltechnische Verfahren wurden mehrmals analysiert (z.B. im Sammelband Leonardi et al. 2016). Als Erzählungen aus der eigenen Lebens- und Familiengeschichte eignen sich die Interviews für Untersuchungen zum Wechselspiel zwischen Erinnerungsprozess, aus dem Gedächtnis (re)konstruierten Inhalten und deren Verbalisierung (s. etwa Betten 2008; Leonardi 2016). Mehrfachinterviews mit verschiedenen Interviewerinnen in Zeitabständen bis zu 20 Jahren ermöglichen Analysen zum Einfluss der interpersonellen Beziehung, aber auch der Zeit auf Kommunikationsverhalten und Sprachgestaltung („Alterssprache", s. Betten 1998) sowie auf Veränderungen in der Erinnerung und der Bewertung des Erlebten (Betten 2016b). Ein Teilkorpus von ISZ wurde speziell für die Untersuchung von Wiedererzählungen in spontanen Narrativen erhoben (Luppi i.E.). Multimodale Analysen anhand der (wenigen) Videoaufnahmen wurden bislang nur ansatzweise erprobt (Koesters Gensini/D'Alesio 2017).

Emotions- und Traumaforschung: Die Narrative thematisieren oft einschneidende, meist mit antisemitischen Maßnahmen und Angriffen zusammenhängende Erlebnisse. An solchen Passagen lassen sich die Beziehungen zwischen Trauma, erinnerten sowie reaktivierten Emotionen und Sprache nachverfolgen, die mit (gesprächs)linguistischen Mitteln analysiert werden können (s. Leonardi et al. 2016; Häußinger 2021).

Chronotopische Beziehungen: In der neueren Forschung wird die Verschränkung der zeitlichen mit der räumlichen Dimension im Erzählen betont: Ermittlungen und Analysen der Orte im Korpus lassen sich mit narratologischen Forschungen im Zusammenhang mit Migration und Exil verknüpfen (u.a. Schwitalla 2012; Betten 2013; Hasslauer 2016); dies steht im Fokus des Projekts *Orte und Erinnerung: Eine Kartographie des Israelkorpus* (Thüne 2020; Leonardi et al. i.E.; s. https://kartografiedesisraelkorpus.wordpress.com).

Quantitativ-qualitative Ansätze: In jüngster Zeit wurden quantitativ-qualitative Ansätze mit Hilfe korpuslinguistischer Tools an den Korpora erprobt. Im Fokus der Untersuchungen stand zum einen die **Emotionsthematisierung** (Emotionswortschatz, Intensitätspartikeln, emotionsdeklarative Formeln, vgl. Brambilla/Flinz 2019, Flinz 2019, Flinz/Moroni 2020; Flinz i.E.) und zum anderen **Ortsnennungen**. Orte – sowohl im Sinne von geographischen Bestimmungen als auch von ‚sozialen Räumen' und ‚Plätzen' einer Stadt – sind auch Schwerpunkt des o.g. Projektes *Orte und Erinnerung*. Als Pilotstudie wurden unterschiedliche Ortsangaben automatisch und manuell annotiert. Die Ergebnisse haben gezeigt, dass ein Zusammenspiel von automatischen und manuellen Verfahren unumgänglich ist, da mit der automatischen Annotation nur partiell Orte aufgefunden werden, Lexeme in Wortverbindungen fehlerhaft markiert und Bedeutungsfacetten nicht unterschieden werden (Flinz/Ruppenhofer 2021).

Literatur:

Ständig aktualisierte Bibliographie zum Korpus: https://www.zotero.org/groups/2219390/israelkorpus/library

Betten Anne (1998): Ist „Altersstil" in der Sprechsprache wissenschaftlich nachweisbar? Überlegungen zu Interviews mit 70- bis 100jährigen Emigranten. In Reinhard Fiehler und Caja Thimm (eds.), *Sprache und Kommunikation im Alter.* Opladen / Wiesbaden, 131–142.

Betten Anne (2008): Schöne und schwere Gedanken an Lublin. In Max Stebler (ed.), *Nicht nur ein Grund für Dankbarkeit.* Festschrift für Jerzy Jeszke. Lublin, 11–24.

Betten Anne (2013): Die erste Reise zurück nach Deutschland: Thematische Fokussierung und Perspektivierung in Erzählungen jüdischer Emigranten. In Arnulf Deppermann und Martin Hartung (eds.), *Gesprochenes und Geschriebenes im Wandel der Zeit. Festschrift für Johannes Schwitalla.* Mannheim, 115–144.

Betten Anne (2016a): Zusammenhänge von Sprachkompetenz, Spracheinstellung und kultureller Identität – am Beispiel der 2. Generation deutschsprachiger Migranten in Israel. In Leonardi, Thüne und Betten (eds.), 353–381.

Betten Anne (2016b): Familiales Gedächtnis und individuelle Erinnerung. Zum Umgang mit traumatischen Erfahrungen in der 1. und 2. Generation deutsch-jüdischer Migranten in Israel. In Leonardi, Thüne und Betten (eds.), 85–121.

Betten Anne (2017): Plädoyer für Themen- und Textsortenvariation in sprachbiographischen Interviews. Am Beispiel von Aufnahmen mit deutschsprachigen Emigranten in Israel. In Ingrid Schröder und Carolin Jürgens (eds.), *Sprachliche Variation in autobiographischen Interviews: Theoretische und methodische Zugänge* (Sprache in der Gesellschaft 35). Frankfurt/M., 167–198.

Brambilla Marina und Flinz Carolina (2019): Orte und entgegengesetzte Emotionen (LIEBE und HASS) in einem Korpus biographischer Interviews (Emigrantendeutsch in Israel – Wiener in Jerusalem). *Studi Germanici* 15/16, 165–187.

Farges Patrick (2020): *Le Muscle et l'Esprit. Masculinités germano-juives dans la post-migration: Le cas des yekkes en Palestine / Israël après 1933.* Bruxelles et al.

Flinz Carolina (2019): Multiword Units and N-Grams Naming FEAR in the Israel-Corpus. In Gloria Corpas Pastor & Ruslan Mitkov (eds.), *Computational and Corpus-Based Phraseology. Third International Conference, Europhras 2019 (Malaga, Spain, September 25–27, 2019, Proceedings).* Cham, 86–98.

Flinz Carolina (i.E.): Emotions and their Relation to Places of the Migration Trajectory. Experiential Declarative Formulas in the Corpus Emigrantendeutsch in Israel: Wiener in Jerusalem (ISW). In Patricia Ronan & Evelyn Ziegler (eds.), *Language and Identity in Migration Contexts.* Oxford.

Flinz Carolina und Moroni Manuela (2020): Die Verwendung von ‚ganz' bei der Thematisierung von Emotionen im Korpus „Emigrantendeutsch in Israel: Wiener in Jerusalem". *Ricognizioni. Rivista di lingue, letterature e culture moderne* 7(13) https://www.ojs.unito.it/index.php/ricognizioni/article/view/4472

Flinz Carolina und Ruppenhofer Josef (2021): Auf dem Weg zu einer Kartographie: automatische und manuelle Analysen am Beispiel des Korpus ISW. *Sprachreport* 1, 44–50.

Haßlauer Steffen (2016): Fluchterlebnisse und ihr sprachlicher Ausdruck. Untersuchungen zu

Agency, Emotionen und Perspektivierung in den Erzählungen zweier jüdischer Emigrantinnen. In Leonardi, Thüne und Betten (eds.), 201–230.

Häußinger Barbara (2021): Vom Sprechen und Schweigen. Zur Darstellung lebensweltlicher Brüche und Verlusterfahrungen in den narrativen Interviews des Israelkorpus. Lorella Bosco und Marella Magris (eds.), *Studi Germanici – Quaderni dell'AIG* 3 (Il non detto / Das Ungesagte), 163–184.

Koesters Gensini Sabine E. & D'Alesio Veronica (2017): Tra il detto e il non detto: l'espressione delle emozioni nelle narrazioni di Dov Zuriel (17.12.1925-30.8.2014). In Sabine E. Koesters Gensini & Maria Francesca Ponzi (eds.), *La lingua emigrata. Ebrei tedescofoni in Israele: studi linguistici e narratologici* (Studi e Ricerche 63). Roma, 109–140.

Leonardi Simona (2016): Erinnerte Emotionen in autobiographischen Erzählungen. In Leonardi, Thüne und Betten (eds.), 1–45.

Leonardi Simona, Thüne Eva-Maria und Betten Anne (eds.) (2016): *Emotionsausdruck und Erzählstrategien in narrativen Interviews: Analysen zu Gesprächsaufnahmen mit jüdischen Emigranten.* Würzburg.

Leonardi Simona, Costa Marcella, Koesters Gensini Sabine E. und Schettino Valentina (eds.) (i.E.): *Orte und Erinnerung: Eine Kartographie des Israelkorpus.* Rom.

Luppi Rita (i.E.): Recounting Central Biographical Experiences: An Analysis of Retellings in Interviews with Second Generation German Speakers in Israel. In Patricia Ronan & Evelyn Ziegler (eds.), *Language and Identity in Migration Contexts.* Oxford.

Schwitalla Johannes (2012): Raumorganisation in Weg-Erzählungen. In Christoph Schubert und Theresa Pham (eds.), *RaumTexte – TextRäume. Sprachwissenschaftliche Studien zur Verortung im Diskurs.* Berlin, 69–112.

Thüne Eva-Maria (2020): Zurück nach Hannover. Ehemalige MigrantInnen jüdischer Herkunft erzählen. In Janina Behr, François Conrad, Stephan Kornmesser und Kristin Tschernig (eds.), *Schnittstellen der Germanistik. Festschrift für Hans Bickes.* Berlin u.a., 323–345.

Regionalsprachliche Korpora des Deutschen – Repositorien des Forschungszentrums Deutscher Sprachatlas

Hanna Fischer / Brigitte Ganswindt / Georg Oberdorfer
(Forschungszentrum Deutscher Sprachatlas, Marburg)

Um Forschungsdaten einerseits für die wissenschaftliche Community sowie die interessierte Öffentlichkeit verfügbar zu machen und sie andererseits in eine Langzeitarchivierung zu überführen, wurde in den letzten Jahren das Management dieser Ressourcen und damit auch der Umgang mit Daten im wissenschaftlichen Alltag stärker fokussiert. Darunter fällt auch der Aufbau von Repositorien, der als wichtige Aufgabe von empirisch arbeitenden Forschungseinrichtungen gelten kann.

Auch das Forschungszentrum Deutscher Sprachatlas (DSA) hat sich dieser Aufgabe verpflichtet und setzt neue Impulse im Bereich des sprachbezogenen Forschungs-

datenmanagements: Das Forschungszentrum verfügt über eine umfangreiche Dokumentation regionalsprachlicher Datensätze aus Ton, Bild und Text, die digital aufgearbeitet (digitalisiert, indiziert, getaggt) und online in einem lebendigen Repositorium publiziert sind – u.a. in Form von digitalen Katalogen für Audiokorpora, einer Online-Bibliografie und digitalen Sprachkarten (Ganswindt/Kehrein/Lameli 2015). Die Audiokorpora umfassen circa 4.000 Übertragungen der „Wenkersätze" in Dialekt und intendierte Standardsprache (vgl. z.B. Göschel 1977). Zum anderen enthält das Repositorium zahlreiche Sprachaufnahmen unterschiedlicher Erhebungssituationen aus internen und externen Projekten (z. B. Aufnahmen zum Mittelrheinischen Sprachatlas (Bellmann 1994) oder Tonkorpora des Preußischen Wörterbuchs (Goltz 2005) sowie flächendeckende Aufnahmen von Vorleseaussprache). Insgesamt stammen die Sprachaufnahmen aus mehr als 70 Jahren Dialektologie, so dass sie auch als Datengrundlage für kurzzeitdiachrone Forschungen geeignet sind.

Die verschiedenartigen Datenbestände sind zum gegenwärtigen Zeitpunkt in dem sprachgeografischen Informationssystem „REDE SprachGIS" zusammengeführt, dessen Basis die geografische Informationsverarbeitung und -vernetzung ist. Es verfügt über spezifische Suchwerkzeuge, die Filterungen nach verschiedenen Korpora oder Dialekträumen erlauben und zudem einen text- und lemmabasierten Zugriff ermöglichen.

Aktuell wird weiterhin ein „Korpus der Regionalsprachen" aufgebaut. Dieses Korpus speist sich vorwiegend aus Aufnahmen freier Gesprächssituationen, die momentan transkribiert und für Annotation und Analyse aufbereitet werden. In einer modularen Vorgehensweise wird das Korpus sukzessive erweitert mit dem Ziel, ein Referenzkorpus der deutschen Regionalsprachen anbieten zu können. In der Korpusgenerierung stehen neben grammatischen Aspekten (Morphologie, Syntax) auch gesprächs- und diskurslinguistische Fragestellungen im Erkenntnisinteresse (Diskursmarker und -organisation, sprachreflexive Handlungen), welche die bisherigen Datenaufbereitungen am DSA ergänzen. Das „Korpus der Regionalsprachen" ist als Teil des frei zugänglichen Repositoriums des DSA mit einem eigenständigen Webinterface für Recherche und Export angelegt. Wert wird dabei auf das in der Korpuslinguistik als momentaner Best-practice-Ansatz angesehene Vorgehen gelegt. Dazu gehört die Darstellung in KWIC (Keyword in Context) und verschiedene Exportformate für eine optimale Les- und Publizierbarkeit sowie multifunktionale Suchoptionen, die Operatoren wie jene der RegEx sowie auch kategoriale Bestimmungen wie jene des POS-Taggings kennen.

Mit der Weiterentwicklung des Repositoriums in den nächsten Jahren werden zentrale Zugänge zu dialektalen und regiolektalen Sprachdaten unterschiedlicher Provenienz aus dem gesamten deutschen Sprachgebiet der wissenschaftlichen Community und interessierten Öffentlichkeit zur Verfügung gestellt. Die Sprachaufnahmen und ihre teils mehrschichtigen Annotationslayer bieten dabei eine einmalige Forschungsressource für die germanistische Variationslinguistik und darüber hinaus.

Auf der Methodenmesse bekommen Sie einen Überblick über die Fülle und die Qualität der verfügbaren Daten und Annotationen sowie den benutzerfreundlichen Zugriff darauf. Des Weiteren werden erste Einblicke in das im Entstehen befindliche „Korpus der Regionalsprachen" ermöglicht.

Literatur:

Bellmann, Günter (1994): Einführung in den Mittelrheinischen Sprachatlas (MRhSA). Tübingen.

Ganswindt, Brigitte / Roland Kehrein / Alfred Lameli (2015): Regionalsprache.de (REDE). In: Kehrein, Roland / Alfred Lameli / Stefan Rabanus (Hrsg.): Regionale Variation des Deutschen – Projekte und Perspektiven. Berlin/Boston: De Gruyter, 421–453.

Goltz, Reinhard (2005): Preussisches Wörterbuch: deutsche Mundarten Ost- und Westpreußens. Akademie der Wissenschaften und der Literatur Mainz. Begr. von Erhard Riemann. Fortgef. von Ulrich Tolksdorf. Hrsg. von Reinhard Goltz. Neumünster: Wachholtz

Göschel, Joachim (Hrsg.) (1977): Die Schallaufnahme deutscher Dialekte im Forschungsinstitut für deutsche Sprache. Marburg.

Forschungsdatenbank Lernertexte (FD-LEX)

Sabine Hachmeister / Sandra Tietjens / Rebekka Wanka / Charlotte Stehr / Michael Becker-Mrotzek (Universität zu Köln)

Die webbasierte Forschungsdatenbank Lernertexte (FD-LEX) ist eine stetig wachsende Datenbank, die Texte von Lernerinnen und Lernern samt ausgewählter Meta- und Testdaten der Probandinnen und Probanden aus verschiedenen Forschungsprojekten zur wissenschaftlichen Nachnutzung bereitstellt. Die Korpora stammen aus unterschiedlichen Forschungsprojekten und umfassen mehrere tausend Texte auf verschiedenen Sprachen, darunter Deutsch, Türkisch oder Russisch. Entsprechend der unterschiedlichen Forschungsdesigns und Fragestellungen der Projekte, liegen Texte verschiedener Textsorten von Schülerinnen und Schülern unterschiedlicher Altersgruppen und Schulformen vor. Alle Texte werden in transkribierter Form im .pdf-Format und zum Teil auch als Scans der handschriftlichen Originale bereitgestellt. Die Daten der einzelnen Korpora können anhand verschiedener Kriterien nach den eigenen Interessen gefiltert werden. Zu jeder Filtervariable sind kurze Erläuterungen hinterlegt. Die Rechercheergebnisse können als Excel- und Datentabelle ausgegeben werden, die transkribierten und – falls vorhanden – die handschriftlichen Lernertexte als PDF-Dateien. Außerdem liegen Informationen zu den jeweiligen Forschungsprojekten vor, beispielsweise die dort verwendeten Schreibimpulse als PDF-Dateien. Hier beispielhaft eine Übersicht über zwei Korpora, die derzeit zur Verfügung stehen:

Das Scriptoria-Textkorpus umfasst 5.628 Texte von 938 Schülerinnen und Schülern auf Deutsch. Es handelt sich hierbei um Berichts- und Argumentationstexte, die in fünften und neunten Klassen an Gesamtschulen und Gymnasien in Hannover und Köln zwischen den Jahren 2013 und 2015 erhoben wurden. Von den Schülerinnen und Schülern wurden zu drei Messzeitpunkten je zwei Texte geschrieben. Dies geschah im Rahmen des Forschungsprojekts *Unterrichtliche Förderung von Teilkomponenten der Schreibkompetenz*, einem Verbundvorhaben der Professur für Pädagogische Psychologie an der Leibniz Universität Hannover (Prof. Grabowski) und der Professur für Deutsche Sprache und ihre Didaktik an der Universität zu Köln (Prof. Becker-Mrotzek). Es

handelt sich um die Weiterführung des Projektes *Diagnose und Förderung von Teilkomponenten der Schreibkompetenz.*

Das TSK-Korpus umfasst 585 Texte von 39 Schülerinnen und Schülern auf Deutsch. Die Texte stammen aus einer Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter, welche an der Universität Siegen durchgeführt wurde. Es wurde anhand von zwei Klassen ermittelt, wie sich die Fähigkeit von Kindern in der 2., 3. und 4. Klasse entwickelt, Texte im Rahmen verschiedener kommunikativer Anforderungen (Textsorten) zu schreiben: Erzählung, Bericht, Instruktion, Beschreibung, Argumentation. Der Erhebungszeitraum lag zwischen den Jahren 2001 und 2003

Die Datenbank ist als ein gemeinsames Projekt des Mercator-Instituts für Sprachförderung und Deutsch als Zweitsprache der Universität zu Köln und der Professur für Pädagogische Psychologie der Leibniz-Universität Hannover entstanden und wurde von der Stiftung Mercator gefördert. Nach der Entwicklungsphase im Jahr 2017, konnte sie 2018 online geschaltet werden, zunächst nur mit den Textdaten des umfangreichen Scriptoria-Korpus. Ende 2019 startete das Projekt zum Ausbau der Datenbank in Kooperation mit dem Verbund Forschungsdaten Bildung und im Dezember 2021 fand der Relaunch der Datenbank-Website statt. Danach werden stetig neue Textkorpora aufgenommen.

Derzeit werden die frequentesten Wörter aus dem Scriptoria-Korpus in einem weiteren Projekt des Mercator-Instituts verwendet. Das Projekt „Entwicklung eines Referenzwortschatz für die Schule“ beschäftigt sich dabei mit der Entwicklung und Bereitstellung eines annotierten, sprachwissenschaftlich und sprachdidaktisch begründeten Wortschatzes für die Klassen 1 bis 10 in Form einer webbasierten Anwendung. Die häufigsten Wörter aus dem Scriptoria-Korpus werden im Teilkorpus „Schreibwortschatz“ des Referenzwortschatzes in die webbasierte Anwendung annotiert eingespeist.

In unserem Beitrag wird der Aufbau der Forschungsdatenbank samt bereits vorhandener Korpora und Recherchemöglichkeiten vorgestellt. Dabei wird beispielhaft eine Recherche im Bestand der Datenbank durchgeführt, um die verschiedenen Filtermöglichkeiten und die Datenausgabe vorzuführen. Zudem wir das Projekt „Entwicklung eines Referenzwortschatzes für die Schule“ vorgestellt und dabei auf die Verknüpfungspunkte zur Datenbank FD-LEX eingegangen.

Links:

www.fd-lex.de

https://www.mercator-institut-sprachfoerderung.de/de/forschung-entwicklung/aktuelle-projekte/forschungsdatenbank-lernertexte-fd-lex/

https://www.mercator-institut-sprachfoerderung.de/de/forschung-entwicklung/aktuelle-projekte/referenzwortschatz-fuer-die-schule/

Entwicklungen in der Erforschung der Jugendsprache im digitalen Raum: Das *NottDeuYTSch*-Korpus

Louis Cotgrove (University of Nottingham / IDS)

In diesem Vortrag wird das *Nottinghamer Korpus Deutscher* YouTube-*Sprache* (Das *NottDeuYTSch*-Korpus) vorgestellt. Das Korpus wurde für meine Dissertation aufgebaut und besteht aus 3,1 Mio. *YouTube*-Kommentaren und 33,7 Mio. Token aus dem Zeitraum zwischen 2008 und 2018. Sie wurden von 112 deutschsprachigen *YouTube*-Kanälen gesammelt, die auf Jugendliche ausgerichtet sind. Das *NottDeuYTSch*-Korpus ist eines der ersten großen Korpora, das deutschsprachige Daten von Jugendlichen aus You Tube-Kommentaren sammelt. Diese bilden einen wichtigen Teil der Jugendkultur und des Jugenddiskurses ab, denn über 85% der 12- bis 19-Jährigen „schauen sich regelmäßig *YouTube*-Videos an" (Bahlo et al. 2019: 80).

Das Korpus wurde entsprechend der Ansprüche an speziellere Web-Korpora aufgebaut (siehe Barbaresi 2019), um korpuslinguistische Arbeiten zur weiteren Untersuchung der Jugendsprache im digitalen Raum zu ermöglichen und die vorhandenen Korpora zu ergänzen, die sich auf andere Bereiche der Sprache im digitalen Raum konzentrieren. Hier wären zum Beispiel das MoCoDa2-Korpus von *WhatsApp*-Nachrichten (Beißwenger et al. 2020), das DiDi-Korpus von *Facebook*-Texten (Glaznieks and Frey 2020) und das IBK-Korpus mit mehreren Online-Quellen, z.B. Emails, IRC-Chats und Blogs (Lüngen and Kupietz 2020) zu nennen. Ab Frühling 2022 wird das *NottDeuYTSch*-Korpus über *KorAP* und *COSMAS II* als Teil des Deutschen Referenzkorpus (DeReKo) verfügbar sein und so den Zugriff für zukünftige Forschung verfügbar machen.

Im Vortrag wird die Methodik vorgestellt, die der Datenauswahl für das *NottDeuYTSch*-Korpus zugrunde liegt, einschließlich der Leitprinzipien für den Aufbau des Korpus und des Identifikationsverfahrens der *YouTube*-Kanäle, von denen die Kommentare gesammelt wurden. Darüber hinaus werden die Prozesse zur Erstellung des *NottDeuYTSch*-Korpus skizziert, methodische Aspekte wie die Ausgewogenheit und der Umfang des Korpus reflektiert und das Stichprobenverfahren eingehend erläutert.

Dieser Vortrag wird die vielen neuen Untersuchungsmöglichkeiten zeigen, die durch die sich aus den zentralen Eigenschaften des *NottDeuYTSch*-Korpus ergeben:

Ausführliche Metadaten: Das Korpus enthält viele Metadaten zu den Videos und Kommentaren. Diese ermöglichen sowohl die Analyse von Videogenres, Zeitrahmen, Nutzer*innen und *YouTuber*innen* als auch Interaktions- und Diskursanalysen von Gesprächen zwischen Kommentator*innen, da die Kommentarstruktur der Threads erhalten bleibt.

Großer Datenbestand: Die hohe Anzahl an Token im *NottDeuYTSch*-Korpus ermöglicht eine umfangreiche quantitative Forschung, z.B.: Längsschnittstudien für lexikalische, orthografische, morphologische und syntaktische Merkmale, die Untersuchung von linguistischen Merkmalen, die nicht häufig vorkommen, wie z.B. einige syntaktische Konstruktionen, und die Erforschung von linguistischen Merkmalen, die für die digital vermittelte Kommunikation spezifisch sind, wie z. B. Emojis und Hashtags.

Multilinguale Daten: Die Kommentare im *NottDeuYTSch*-Korpus sind überwiegend deutschsprachig (einschließlich dialektaler Verwendungen), aber es gibt auch eine nennenswert Präsenz anderer Kommentarsprachen wie Englisch, Türkisch und Russisch sowie sprachliche Elemente aus mehreren Sprachen innerhalb desselben Kommentars (das sogenannte *Translanguaging*, siehe García and Li 2014). Im Vortrag werden auf quantitativen und qualitativen Analysen basierende Beispiele der Mehrsprachigkeit in der Jugendsprache gezeigt.

Der Vortrag illustriert die breiten Anwendungsmöglichkeiten des *NottDeuYTSch*-Korpus zur Analyse der internetbasierten schriftlichen Kommunikation von Jugendlichen auf der Basis von Daten aus bisher noch wenig erforschten Social Media-Quellen. Damit kann er einen wichtigen Beitrag zu zukünftigen korpuslinguistischen Studien sowohl über digital vermittelte Kommunikation als auch über Jugendsprache selbst leisten.

Literatur:

Bahlo, Nils et al. 2019. *Jugendsprache: Eine Einführung* (Berlin: J.B. Metzler)

Barbaresi, Adrien. 2019. 'The Vast and the Focused: On the Need for Thematic Web and Blog Corpora', in *Proceedings of the Workshop on Challenges in the Management of Large Corpora*, hrsg. v. Piotr Bański et al. (Mannheim: Leibniz-Institut für Deutsche Sprache), pp. 29–32

Beißwenger, Michael et al. 2020. 'Die Mobile Communication Database 2 (MoCoDa 2)', in *Deutsch in Sozialen Medien: Interaktiv – multimodal – vielfältig*, hrsg. v. Konstanze Marx, Henning Lobin und Axel Schmidt (Berlin: De Gruyter), pp. 349–52

García, Ofelia & Wei Li. 2014. *Translanguaging: Language, Bilingualism and Education* (London: Palgrave Macmillan)

Glaznieks, Aivars & Jennifer-Carmen Frey. 2020. 'Das DiDi-Korpus: Internetbasierte Kommunikation Aus Südtirol', in *Deutsch in Sozialen Medien: Interaktiv – multimodal – vielfältig*, hrsg. v. Konstanze Marx, Henning Lobin und Axel Schmidt (Berlin: De Gruyter), pp. 353–54

Lüngen, Harald & Marc Kupietz. 2020. 'IBK- und Social-Media-Korpora am Leibniz-Institut für Deutsche Sprache', in *Deutsch in Sozialen Medien: Interaktiv – multimodal – vielfältig*, hrsg. v. Konstanze Marx, Henning Lobin und Axel Schmidt (Berlin: De Gruyter), pp. 319–42 <https://doi.org/10.1515/9783110679885-016>

Die österreichische Presselandschaft digital: Das Austrian Media Corpus (AMC) – Aufbau, Bedienung und Möglichkeiten

Amelie Dorn[1] / Jan Höll[1] / Wolfgang Koppensteiner[1] / Hannes Pirker[2] / Theresa Ziegler[1] ([1]Universität Wien, [2]ACDH-CH ÖAW)

Der Einsatz von Korpora in diversen Gebieten der Sprachforschung verzeichnete in den letzten Jahren einen deutlichen Anstieg, nicht zuletzt bedingt durch die voranschreitenden und umfassenden Digitalisierungsmaßnahmen u. A. in Geistes- und Sozialwissenschaften, aber auch durch die stetig neue Entwicklung unterschiedlicher Methoden und Werkzeuge, um digitalisierte Korpora zu erschließen oder digital erzeugte („born-digital") Korpora systematisch zu analysieren und dadurch Muster, Gebrauch und Verände-

rungen in Sprache und Kommunikation aufzuzeigen und nachzuweisen. Das Austrian Media Corpus (AMC) (siehe Ransmayr 2014; Ransmayr, Mörth & Ďurčo 2013; 2017) stellt ein wertvolles Beispiel eines digital erzeugten Textkorpus dar, das für weitreichende Analysen und auch als Vergleichskorpus herangezogen werden kann. Das AMC wird durch eine Kooperation der Austria Presse Agentur (APA) und der Österreichischen Akademie der Wissenschaften (ÖAW), konkret dem Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) zu Forschungszwecken bereitgestellt. Es ist derzeit mit ca. 45 Mio. Artikeln und 11 Milliarden Token eines der größten deutschen Textkorpora, das inhaltlich für unterschiedlichste sprachwissenschaftliche Fragestellungen zum (österreichischen) Deutschen herangezogen werden kann. Das AMC umfasst neben allen österreichischen Tageszeitungen auch die gängigsten Wochen- und Monatszeitschriften. So macht das AMC nahezu die gesamte Printmedienproduktion der letzten 25 Jahre eines Landes für die wissenschaftliche Forschung zugänglich. Diese Tatsache macht das AMC einzigartig in seinem inhaltlichen Umfang.

Der vorliegende Beitrag leistet zweierlei: Einerseits erfolgt eine extensive Darstellung des AMC, welche neben theoretisch-technischen Aspekten auch eine praktische, interaktive Auseinandersetzung mit dem Korpus bietet (NoSketch Engine (Kilgarriff et al. 2014; Rychlý 2007)). BesucherInnen der Methodenmesse erhalten die Möglichkeit über das NoSketch Engine-Interface selbst Abfragen im Korpus vorzunehmen. Andererseits wird an einer Auswahl an bereits durchgeführten Studien der Nutzen des AMC für die korpuslinguistisch orientierte Forschung demonstriert (vgl. Höll 2020, Koppensteiner 2015, Ziegler 2021). Unter anderem wird der Einsatz des AMC in den Kontext des Spezialforschungsbereichs „Deutsch in Österreich. Variation – Kontakt – Perzeption“ (Lenz 2018) gestellt, der dieses Korpus bereits zur Analyse des Gebrauches von sogenannten „Austriazismen“ aus variationslinguistischer Perspektive einsetzt.

Literatur:

Austrian Media Corpus (amc) https://amc.acdh.oeaw.ac.at/ [letzter Zugriff: 11.11.2021]

Höll, Jan (2020): Wissensstrukturen im Kontext der deutschen Rechtschreibreform. Framebasierte Analyse eines medialen Diskurses in Österreich. Masterarbeit, Universität Wien. Wien.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel (2014): The Sketch Engine: Ten Years on. In: Lexicography 1, 7–36.

Koppensteiner, Wolfgang (2015): Das österreichische Deutsch im plurizentrischen Kontext. Diplomarbeit, Universität Wien. Wien.

Lenz, Alexandra N. (2018): The Special Research Programme „German in Austria. Variation – Contact – Perception“. In: Ammon, Ulrich & Marcella Costa (Hgg.): Sprachwahl im Tourismus – mit Schwerpunkt Europa. (Sociolinguistica. Internationales Jahrbuch für europäische Soziolinguistik 32). Berlin/Boston, 269–277.

NoSketch Engine https://nlp.fi.muni.cz/trac/noske [letzter Zugriff: 24.11.2021].

Ransmayr, Jutta (2014): Neue Forschungsmöglichkeiten zum österreichischen Deutsch mit dem Austrian Media Corpus (AMC). In: Ransmayr, Jutta, Andrea Moser-Pacher & Ilona Elisabeth Fink: Österreichisches Deutsch und Plurizentrik. (Informationen zur Deutschdidaktik. Zeitschrift für den Deutschunterricht in Wissenschaft und Schule 3/2014). Innsbruck: StudienVerlag, 63–68.

Ransmayr, Jutta, Karlheinz Mörth & Matej Ďurčo (2013): Linguistic Variation in the Austrian Media Corpus. Dealing with the challenges of large amounts of data. In: Procedia – Social and Behavioral Sciences 95, 111–115.

Ransmayr, Jutta, Karlheinz Mörth & Matej Ďurčo (2017): AMC (Austrian Media Corpus) – Korpusbasierte Forschungen zum österreichischen Deutsch. In: Resch, Claudia & Wolfgang U. Dressler (Hgg.): Digitale Methoden der Korpusforschung in Österreich (Veröffentlichungen zur Linguistik und Kommunikationsforschung 30). Wien: Verlag der Österreichischen Akademie der Wissenschaften, 27–38.

Rychlý, Pavel (2007): Manatee/Bonito-A Modular Corpus Manager. In: Proceedings of Recent Advances in Slavonic Natural Language Processing 2007 (RASLAN), 65–70.

Ziegler, Theresa (2021): Über Geschenk-s-körbe und Schokolade-Ø-torten (zu runden Geburtstagen). Sneak Peek auf eine Abschlussarbeit über areal-horizontale Verfugungstendenzen bei NN-Komposita mit schwankenden Fugenelementen in der österreichischen Standard(schrift)sprache. In: Kim, Agnes, Katharina Korecky-Kröll, Ludwig M. Breuer, Jan Höll & Wolfgang Koppensteiner: Vom Tun nicht lassen können. Historische und rezente Perspektiven auf sprachliche Variation (in Österreich und darüber hinaus). Festgabe für Alexandra N. Lenz zum runden Geburtstag. (WLG 89) Wien: Institut für Sprachwissenschaft, Universität Wien.

Das deutsch-dänische XPEROHS-Korpus: Hassrede in sozialen Medien

Eckhard Bick / Klaus Geyer (Süddänische Universität)

Das hier vorgestellte Korpus wurde mit dem Ziel erstellt, Sprachgebrauch in sozialen Medien qualitativ und quantitativ auf Hassrede gegen ethnische und religiöse Minderheiten untersuchen zu können. Das zweisprachige, deutsch-dänische Korpus bildet die datenmäßige Grundlage des XPEROHS-Projekts (*Towards Balance and Boundaries in Public Discourse: Expressing and Perceiving Online Hate Speech*, Baumgarten et al. 2019). Es handelt sich um ein Monitor-Korpus von Twitter- und Facebook-Einträgen über den Zeitraum Ende 2017 bis 2020, wobei die Abfrageschnittstelle für Twitter einen fast 100-prozentigen Deckungsgrad ermöglichte, während Facebook eine Vorauswahl bestimmter Webseiten und damit einen thematischen Bias erzwang. Insgesamt enthalten die Twitter-Korpora ca. 2 Milliarden Wörter für Deutsch und 270 Millionen Wörter für Dänisch, die kleineren Facebook-Korpora ca. 200 Millionen Wörter für Deutsch und 60 Millionen Wörter für Dänisch.

Um gezielte linguistische Statistiken und die qualitative Identifizierung linguistischer Muster von Hassrede zu erlauben, wurde das Korpus mehrfach grammatisch und lexikalisch-semantisch mit Hilfe von Constraint-Grammar-Parsern annotiert (Bick 2020a), wobei Grammatiken und Lexika laufend an spezielle sprachliche Merkmale des Genres (z. B. unvollständige Sätze, orthographische Variationen, kreative Wortbildungen und nicht-textliche Kürzel) sowie an den inhaltlichen Distinktionsbedarf des Projektes angepasst wurden. Neben Lemmatisierung und standardmäßiger morphosyntaktischer Annotation wurde auch auf semantische Aspekte und eine genaue morpholo-

gische Analyse wert gelegt. So werden die Bestandteile von Komposita erkannt und Inhaltswörter semantisch klassifiziert und disambiguiert, z. B. Substantive als ‚Berufsbezeichnung', ‚Tier', ‚Werkzeug', ‚Fahrzeug' oder ‚Gebäude'. Verben (und prädizierende Substantive) erhalten eine Framenet-Kategorie und einen Satz von semantischen Rollen für ihre Argumente (Bick 2011 und Bick 2017). Besonderes Augenmerk lag auch auf der Unterstützung von Sentiment Analysis, z.B. über die Polarität von Adjektiven und indem Text-Emoticons und bildliche Emojis morphosyntaktisch in die Satzanalyse eingebunden und als zehn verschiedene Emotionen „lemmatisiert" wurden (Bick 2020b).

Als grafische Benutzerschnittstelle wurde eine erweiterte Version von Corpuseye (corp.hum.sdu.dk) verwendet, die intern auf einer CQP-Datenstruktur (Hardie 2012) basiert. Die Projektversion enthält neben Feld- und Tag-basierten Suchoptionen auch Neuerungen wie gestaffelte Suchen, N-gram-Auswertung, die Unterstützung von Emojis und den Zugang zu Dependenzrelationen.

In unserer Präsentation besprechen wir eine Reihe von korpusspezifischen Herausforderungen datenrechtlicher, technischer und korpuslinguistischer Art und zeigen entsprechende Lösungsansätze auf. Schließlich erläutern wir an Beispielen, wie das Korpus in der Hassredeforschung eingesetzt werden kann, z.B. bei der Identifikation und Exemplifizierung von indirekter Hassrede durch konstruktionsgrammatische Muster (Bick / Geyer / Kleene 2021 und Geyer / Bick / Kleene i. Dr.), bei der Erkennung bestimmter metaphorischer Prädikationen oder bei der Unterscheidung von Erwähnung vs. Verwendung (*mention* vs. *use*) von Ethnopaulismen (Geyer 2021).

Literatur:

Baumgarten, Nicole / Bick, Eckhard / Geyer, Klaus / Iversen, Ditte Aakær / Kleene, Andrea / Lindø, Anna Vibeke / Neitsch, Jana / Niebuhr, Oliver / Nielsen, Rasmus / Petersen, Nedenskov, Esben (2019). Towards Balance and Boundaries in Public Discourse: Expressing and Perceiving Online Hate Speech (XPEROHS). In: *RASK – International Journal of Language and Communication* 50, 87-108.

Bick, Eckhard (2011). A FrameNet for Danish. In: *Proceedings of NODALIDA 2011*, May 11-13, Riga, Latvia. NEALT Proceedings Series, Vol 11, 34-41. Tartu: Tartu University Library.

Bick, Eckhard (2017). Propbank Annotation of Danish Noun Frames. In: *Proceedings of IWCS2017 – 12th International Conference on Computational Semantics* (Montpellier, September 2017). ACL Anthology W17-69. URL: http://aclweb.org/anthology/W17-69.

Bick, Eckhard (2020a). An Annotated Social Media Corpus for German. In: Calzolari, Nicoletta et al. (eds.), *Proceedings of the 12th International Conference on Language Resources and Evaluation, LREC2020* (Marseille, May 2020), 6129-6137. ACL / ELRA.

Bick, Eckhard (2020b). Annotating Emoticons and Emojis in a German-Danish Social Media Corpus for Hate Speech Research. In: *RASK – International Journal of Language and Communication* 52, 1-20.

Bick, Eckhard / Geyer, Klaus / Kleene, Andrea. (2020). „Die ách so friedlichen Muslime": eine korpusbasierte Untersuchung von Formulierungsmustern fremdenfeindlicher Aussagen in Sozialen Medien. In: Wachs, Sebastian / Koch-Priewe, Barbara / Zick, Andreas (Hrsg.), *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen: Theoretische und empirische Annäherungen an ein interdisziplinäres Phänomen.* Heidelberg: Springer VS. 81-103

Geyer, Klaus (2021). Verwendung (*use*) vs. Erwähnung (*mention*) von Ethnophaulismen: Eine Untersuchung zu Hassrede in Facebook-Beiträgen. In: Lüger, Heinz-Helmut / Giessen, Hans Werner (Hrsg.), *Text-, Diskurs- und Kommunikationsforschung: Festschrift für Hartmut Lenk*. Landau: Verlag Empirische Pädagogik, 165-184.

Geyer, Klaus / Bick, Eckhard / Kleene, Andrea (i. Dr.). "I am not a racist, but ...". A Corpus-Based Analysis of Xenophobic Hate Speech Constructions in Danish and German Social Media Discourse. In: Natalie Knoblock (ed.), *Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanizing Discourse*. Cambridge: Cambridge University Press.

Hardie, Andrew (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. In: *International Journal of Corpus Linguistics* 17 (3), 380-409.

KompAS-Korpus: 180 argumentative Gespräche Deutschschweizer Schulkinder. Korpus und interaktive Visualisierung.

Daniel Müller-Feldmeth / Tamara Koch / Chantal Wanderon / Martin Luginbühl (Universität Basel)

Im Rahmen des SNF Projekts «Argumentative Gesprächskompetenz in der Schule: Kontexte, Anforderungen, Erwerbsverläufe» (Hauser & Luginbühl, 2017) sowie des derzeit laufenden SNF Projekts «Kompetenzniveaus mündlichen Argumentierens unter Schulkindern (KompAS)»[1] wurde ein Korpus argumentativer Gespräche Deutschschweizer Schulkinder (7-13 Jahre) erstellt. Das Korpus umfasst 180 videographierte, transkribierte und annotierte argumentative Gespräche. Jeweils vier Kinder aus den Klassenstufen 2, 4, oder 6 (Primarschule) bearbeiteten mündlich eine von zwei Entscheidungsaufgaben (ohne Anwesenheit von Erwachsenen): In der *Robinsonaufgabe* sollten sich die Kinder auf drei von zwölf Gegenständen einigen, die sie auf eine einsame Insel mitnehmen würden (60 Gespräche, je 20 pro Klassenstufe). In der *Spendenaufgabe* sollten die Kinder eine nach Priorität sortierte Rangfolge von vier Projekten erstellen, welchem sie 50 CHF spenden würden. In 60 Gesprächen (je 20 pro Klassenstufe) sollten sich die Kinder vorstellen, sie hätten 50 CHF, in weiteren 60 Gesprächen wurde das Geld wirklich gespendet (fiktives vs. reales Setting).

Die Daten umfassen die Transkription der (meist in Schweizerdeutsch geführten) Gespräche nach GAT2 (Selting et al., 2009), eine Übersetzung ins Standarddeutsche, sowie zentrale nonverbale Aktivitäten. Zudem wurden unterschiedliche argumentativ relevante Aspekte kodiert, unter anderem das Thema (Objekt/Spendenprojekt) jeder Äusserung, Auslösehandlungen (Thematisierungen, Positionierungen für/gegen Objekt/Projekt), der interaktionale Status von Begründungen (zustimmend, ablehnend, …), sowie zusätzliche Informationen (Metakommunikation etc.). Derzeit arbeiten wir an der Kodierung weiterer Aspekte, z.B. konversationeller «Jobs» und pragmatischer

1) https://germanistik.philhist.unibas.ch/de/forschung/forschungsprojekte/sprachwissenschaft/kompetenzniveaus-muendlichen-argumentierens-unter-schulkindern-kompas

Mittel (Heller, 2012; Domenech et al., 2018; Quasthoff et al., 2021), Topoi (Arendt, 2019; Schwarze, 2021; Kienpointner, 2017; Ottmers, 2007) argumentativer Episoden sowie der argumentativen Struktur.

Die Daten umfassen 873 Minuten Gesprächszeit, ca. 164.800 Wörter sowie 49.830 vergebene Codes.

Verfügbar ist das Korpus über eine interaktive *R-shiny* Anwendung (Chang et al., 2021), die kontinuierlich erweitert wird (https://argcomp.shinyapps.io/baselargVis/). In der Anwendung lässt sich die sequentielle Struktur der einzelnen Gespräche in einer Partitur-Ansicht darstellen. Dabei können unterschiedliche Kodierungen ausgewählt und kombiniert werden und so im Zeitverlauf exploriert werden. Eine integrierte Suchfunktion erlaubt es zudem, mithilfe regulärer Ausrücke in den unterschiedlichen Annotationsspuren zu suchen, wobei die Äusserungen hervorgehoben werden, die den Suchausdruck enthalten. In Kooperation mit RISE (Research & Infrastructure Support, https://rise.unibas.ch) werden die (anonymisierten) Daten, einschliesslich der Videos und Transkripte, zukünftig in der «ZuMult» Umgebung zur Verfügung stehen (https://zumult.org).

Das Korpus bietet aufgrund der relativ «natürlichen» Peer-Interaktionen (dazu Quasthoff, 2021) reichhaltige Möglichkeiten für gesprächsanalytisch-qualitative Analysen argumentativer Kompetenz im frühen Schulalter, lässt aber aufgrund der grossen Datenmenge und der reichhaltigen Kodierung auch vielfältige Möglichkeiten quantitativer Analysen zu (Mundwiler et al., 2019). Die interaktive Visualisierung ermöglicht dabei die browserbasierte Exploration der Daten unter unterschiedlichsten Gesichtspunkten.

Literatur:

Chang, Winston / Cheng, Joe / Allaire, JJ / Sievert, Carson / Schloerke, Barret / Xie, Yihuo / Allen, Jeff / McPherson, Jonathan / Dipert, Alan / Borges, Barbara (2021): *shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny

Domenech, Madeleine / Heller, Vivien / Petersen, Inger (2018): Argumentieren mündlich, schriftlich, zweitsprachlich. Verfahren und Anforderungen. – In: Massud, Abdel-Hafiez (Hrsg.): *Argumentieren im Sprachunterricht.* Landau: Verlag Empirische Pädagogik (= Beiträge zur Fremdsprachenvermittlung), S. 15-35.

Hauser, Stefan / Luginbühl, Martin (2017): Wenn Kinder argumentieren – Grundlagen und erste Befunde einer Studie zur mündlichen Argumentationskompetenz von Schulkindern. – In: Meißner, Iris; Wyss, Eva Lia (Hrsg.): *Begründen – Erklären – Argumentieren. Konzepte und Modellierung in der angewandten Linguistik*. Tübingen: Stauffenburg (= Stauffenburg Linguistik 93), S. 89-105.

Heller, Vivien (2012): *Kommunikative Erfahrungen von Kindern in Familie und Unterricht. Passungen und Divergenzen*. Tübingen: Stauffenburg (= Stauffenburg Linguistik 67).

Kienpointner, Manfred (2017): Topoi. – In: Roth, Kersten S. / Wengeler, Martin / Ziem, Alexander (Hrsg.): *Handbuch Sprache in Politik und Gesellschaft*. Berlin: de Gruyter (= Handbücher Sprachwissen HSW 19), S. 187-211.

Mundwiler, Vera / Kreuz, Judith / Müller-Feldmeth, Daniel / Luginbühl, Martin / Hauser, Stefan (2019). Quantitative und qualitative Zugänge in der Gesprächsforschung. Methodologische

Betrachtungen am Beispiel einer Studie zu argumentativen Gruppendiskussionen. *Gesprächsforschung*, 20, S. 323-383.

Ottmers, Clemens (2007): *Rhetorik*. J.B. Metzler.

Quasthoff, Uta / Heller, Vivien / Morek, Miriam (2021): Diskurskompetenz und diskursive Partizipation als Schlüssel zur Teilhabe an Bildungsprozessen: Grundlegende Konzepte und Untersuchungslinien. – In: Quasthoff, Uta / Heller, Vivien / Morek, Miriam (Hrsg.): *Diskurserwerb in Familie, Peergroup und Unterricht: Passungen und Teilhabechancen.* Berlin, Boston: De Gruyter, S. 13-34.

Quasthoff, Uta (2021): Methodische Überlegungen zur Datenbasis in der Interaktionalen Diskursanalyse: Grundlegendes zur Erhebung von Stichproben und Korpora. – In: Quasthoff, Uta / Heller, Vivien / Morek, Miriam (Hrsg.): *Diskurserwerb in Familie, Peergroup und Unterricht: Passungen und Teilhabechancen.* Berlin, Boston: De Gruyter, S. 43-76.

Schwarze, Cordula (2021): Gesprächsrhetorik und Topik. – In: Hess-Lüttich, Ernest W. B. (Hrsg.): *Handbuch Gesprächsrhetorik*. Berlin, Boston: de Gruyter (= Handbücher Rhetorik 3), S. 143-153.

Selting, Margret / Auer, Peter / Barth-Weingarten, Dagmar / Bergmann, Jörg / Bergmann, Pia / Birkner, Karin / Couper-Kuhlen, Elizabeth / Deppermann, Arnulf / Gilles, Peter / Günthner, Susanne / Hartung, Martin / Kern, Friederike / Mertzlufft, Christine / Meyer, Christian / Morek, Miriam / Oberzaucher, Frank / Peters, Jörg / Quasthoff, Uta / Schütte, Georg / Stukenbrock, Anja / Uhmann, Susanne (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10, 353-402.

Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650) (ReN) als Impuls für die historische Grammatikforschung

Sarah Ihden / Ingrid Schröder (Universität Hamburg)

Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650) (ReN) erschließt eine Schriftkultur, die durch ihre prominenten Funktionen als internationale Verkehrssprache der Hanse im Nord- und Ostseeraum, als Rechts- und Verwaltungssprache in den Städten, als Fachsprache v.a. der Medizin, aber auch als Sprache religiöser Alltagspraxis im Mittelalter und in der frühen Neuzeit geprägt wird.

Das nach raum-zeitlichen Kriterien sowie Textgruppen (‚Feldern der Schriftlichkeit') gegliederte Korpus umfasst in der Version 1.1 vom 6.1.2021 (ReN-Team 2021) insgesamt 2,3 Millionen Token, von denen ca. 1,5 Millionen Token grammatisch annotiert sind, und ist über das Such- und Visualisierungstool ANNIS (Krause/Zeldes 2016) im Internet frei zugänglich. Es bietet für jeden Text Metadaten zum Entstehungskontext, zu weiteren quellenrelevanten Aspekten (z.B. Aufbewahrungsort des Originals, Editionen und weiterführende Forschungsliteratur) sowie zur Bearbeitung (z.B. Umfang des aufgenommenen Textes/Ausschnitts, Name der Bearbeiter:innen, Hinweise zum Vorgehen der Transkription und Annotation). Das ReN liefert auf der Textebene eine originalgetreue sowie eine

hinsichtlich der grammatischen Worteinheiten normalisierte Transkription und enthält Annotationen zum Lemma, zur Wortart (PoS), zur Flexionsmorphologie sowie zu Satzeinheiten. Das für die grammatische Annotation genutzte Historische Niederdeutsch-Tagset HiNTS (Barteld et al. 2018) orientiert sich grundlegend am Historischen Tagset HiTS (Dipper et al. 2013), das in den anderen Referenzkorpora zur deutschen Sprachgeschichte im Verbund Deutsch Diachron Digital (Altdeutsch, Mittelhochdeutsch, Frühneuhochdeutsch und Deutsche Inschriften) verwendet wird. Die ähnliche Anlage der Referenzkorpora hinsichtlich der Struktur und der Annotationen stellt die Basis für sprachstufenübergreifende korpusbasierte Analysen zum älteren Deutschen dar.

Zu ReN-spezifischen Besonderheiten zählen drei nach Grad der Diplomatizität unterschiedene Textansichten in ANNIS und eine von ANNIS unabhängige Leseversion aller Texte im PDF-Format sowie die verschiedenen Darstellungsformen des Lemmas von der eindeutigen Form wie im Wörterbuch (MNWB 1956 ff.) bis zu einer um sämtliche Sonderzeichen reduzierten Form zum einfachen Einstieg in die lemmabasierte Suche. Ein besonderes Merkmal der flexionsmorphologischen Annotation stellt der Einsatz von Portmanteau-Tags (z.B. Dat-Akk für Dativ oder Akkusativ) dar, welche die genaue Abbildung formaler Ambiguitäten erlauben. Auf diese Weise sollen Interpretationen der Annotator:innen, welche die Analysen vorwegnehmen und so zu Zirkelschlüssen führen, vermieden werden.

Mit seiner hinsichtlich formaler Ambiguitäten umsichtigen Annotation stellt das ReN eine geeignete Datengrundlage für Analysen im Rahmen der Erarbeitung einer neuen mittelniederdeutschen Grammatik (DFG: SCHR 999/9-1) bereit. Gegenwärtig wird auf der Basis des ReN die Flexion der Appellativa analysiert. Erste Auswertungen zeigen, dass die Struktur und die Annotationen im Korpus variationssensitive Untersuchungen ermöglichen, die neue Einblicke in die Substantivflexion im Mittelniederdeutschen, beispielsweise bezüglich diachroner Entwicklungen, liefern (vgl. Ihden/Schröder 2021).

Im Bereich der Lexembildung bietet das ReN gemeinsam mit den anderen historischen Referenzkorpora eine geeignete Basis für das geplante Projekt Wortfamilien diachron (WoDia) – Eine Forschungsumgebung zur historischen Wortbildung des Deutschen, das den alt- und mittelhochdeutschen, altsächsischen und mittelniederdeutschen Wortschatz anhand der Wortfamilienstruktur abbilden und miteinander verknüpfen soll. Über Strukturformeln werden die Lexembildungseinheiten sowie ihre hierarchischen Verbindungen dargestellt. In WoDia sind außerdem die Lemmata der Referenzkorpora mit denen der Zeitstufenwörterbücher verlinkt, sodass für Analysen die unterschiedlichen Vorteile dieser Ressourcen sinnvoll kombiniert werden können. Die Korpora belegen das Vorkommen bestimmter Lexembildungseinheiten und -arten in verschiedenen Schreibsprachlandschaften, Zeiträumen, Textsorten und Texten.

In der Posterpräsentation soll das ReN mit seinen Spezifika erläutert und die Anwendungsmöglichkeiten der historischen Korpora, insbes. des ReN, sollen anhand der zwei genannten korpusbasierten Projekte demonstriert werden. Zur Veranschaulichung sollen exemplarische Suchen im Korpus in ANNIS am Rechner präsentiert werden.

Literatur:

Barteld, Fabian/Ihden, Sarah/Dreessen, Katharina/Schröder, Ingrid (2018): HiNTS: A Tagset for Middle Low German. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, 3940–3945. Online unter: http://www.lrec-conf.org/proceedings/lrec2018/pdf/870.pdf (Stand: 15.11.2021).

Barteld, Fabian/Dreessen, Katharina/Ihden, Sarah/Schröder, Ingrid (2017): Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650) – Korpusdesign, Korpuserstellung und Korpusnutzung. In: Becker, Anja/Hausmann, Albrecht (Hrsg.): Mittelniederdeutsche Literatur. Mitteilungen des deutschen Germanistenverbandes 64/3, 226–241.

Dipper, Stefanie/Donhauser, Karin/Klein, Thomas/Linde, Sonja/Müller, Stefan/Wegera, Klaus-Peter (2013): HiTS: ein Tagset für historische Sprachstufen des Deutschen. In: Journal of Language Technology and Computational Linguistics 28/1, 85–137. Online unter: https://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/jlcl13_webVersion.pdf (Stand: 15.11.2021).

Ihden, Sarah/Schröder, Ingrid (2021): Mittelniederdeutsche Grammatik: Konzeption und erste Analysen. In: Niederdeutsches Jahrbuch 144, 79–104.

Krause, Thomas/Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: Digital Scholarship in the Humanities 31, 118–139. Online unter: http://dsh.oxfordjournals.org/content/31/1/118 (Stand: 15.11.2021).

MNWB (1956 ff.) = Lasch, Agathe/Borchling, Conrad/Cordes, Gerhard/Möhn, Dieter/Schröder, Ingrid (Hrsg.) (1956 ff.): Mittelniederdeutsches Handwörterbuch. Bd. 1 ff. Neumünster.

Korpus:

ReN-Team 2021: Reference Corpus Middle Low German/Low Rhenish (1200–1650); Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650). Version 1.1 vom 06.01.2021. http://doi.org/10.25592/uhhfdm.9195 (Stand: 15.11.2021).

Ein Briefwechsel-Korpus des 16. Jahrhunderts in Frühneuhochdeutsch

Patricia Scheurer (Universität Zürich)

Der Zürcher Reformator Heinrich Bullinger (1504-1575) war Mitarbeiter und Nachfolger von Huldrych Zwingli und ein wichtiger Multiplikator für die Ideen der Reformation in der Schweiz und in Europa. Er stand mit wichtigen Persönlichkeiten seiner Zeit (darunter Calvin, Luther und Könige von England) in Kontakt und entsprechend groß war sein Korrespondenz-Netzwerk. Die Briefe behandeln ein breites Spektrum an Themen von theologischen Fragen über politische Ereignisse bis zu alltäglichen Begebenheiten und geben Aufschluss über Beziehungen und Bildung, aber auch über Lebensmittel des 16. Jahrhunderts.

Von Bullingers Briefwechsel sind rund 10.000 Briefe an und 2.000 Briefe von Bullinger erhalten, etwa 20% davon sind in Frühneuhochdeutsch verfasst, der Rest hauptsächlich in Latein. Zu rund 9.000 Briefen liegen bereits Transkriptionen des Schweizerischen

Instituts für Reformationsgeschichte vor, die das Institut für Computerlinguistik der Universität Zürich im Rahmen des Digitalisierungsprojekts «Bullinger digital» in eine Datenbank eingepflegt und mit Metainformationen versehen hat.

Daraus ist ein Korpus mit aktuell 2.100 Briefen in Frühneuhochdeutsch entstanden (davon über 1.000 Briefe aus der in Buchform publizierten HBBW-Edition), das über eine Million laufende Wortformen umfasst. Ziel des Projektes «Bullinger digital» mit Laufzeit bis Ende 2022 ist es, die fehlenden 3.000 Transkriptionen mit Hilfe eines eigens entwickelten Handwritten Text Recognition (HTR)-Systems zu erstellen und in die Datenbank einzuspielen. Auch ist die automatische Identifizierung der frühneuhochdeutschen Passagen, die in den Latein-Briefen vorkommen, in Arbeit. Die Körpergröße wird also noch deutlich zunehmen.

Aktuell liegen von 327 verschiedenen Korrespondenten Briefe auf Frühneuhochdeutsch vor, was beispielsweise die Untersuchung regionaler Unterschiede in der Schreibweise ermöglicht. Da sich Bullingers Briefwechsel über ein halbes Jahrhundert erstreckt (1523-1575), sind auch diachrone Untersuchung zum Sprachwandel oder zur allfälligen Standardisierung von Schreibvarianten denkbar. Die gemischt-sprachlichen Briefe bieten sich für Untersuchungen zum Code-Switching an.

Zu über 90% der Briefe wird bis Ende 2022 ein Faksimile der Handschrift vorliegen, inklusive Alignierung zwischen Faksimile und Transkription. Ebenfalls Bestandteil des Projektes ist die Normalisierung der frühneuhochdeutschen Texte, sodass vorkommende Schreibvarianten, z.B. *zyten*, *zytten*, *ziten*, *zitten*, *zeyten*, *zeiten* mit der Abfrage des standarddeutschen Begriffs z.B. *Zeiten* aufgefunden werden können. Die Normalisierung erfolgt mit einem statistischen Verfahren, das die historischen Wortformen auf moderne Entsprechungen abbildet.

Gegenwärtig kann das Frühneuhochdeutsche Korpus nach Absprache zur Verfügung gestellt werden. Im weiteren Projektverlauf ist vorgesehen, dass das Korpus über die Suche zusammengestellt und im TEI-XML-Format heruntergeladen werden kann. Die Bullinger-Korrespondenz ist über www.bullinger-digital.ch durchsuchbar.

Folgende zwei Beispiele sollen einen Eindruck von dem zur Verfügung stehenden Korpus vermitteln:

Beispiel für Transkription mit historischen Sonderzeichen

Wyter ist min gar ernstlich begår, das ir nunmee die råchnung stellen und mir anzeigen wȯllind, was ich üch schuldig und ze thůn sye, es sye des tischs halb, über das er långer by üch dann Josias by mir xin, ouch alles des, das ir imm gelihen und für inn uußgåben habend. (https://www.bullinger-digital.ch/letter/12651)

Beispiel für Normalisierung mit statistischem Verfahren

Frühneuhochdeutsch	*Normalisiert*
Der mann ist das houpt, das wie imm houpt	Der Mann ist das Haupt, das wie ihm Haupt
alle vernunfft und fürsorg ist	alle Vernunft und Fürsorge ist
für den gantzen lib,	für den ganzen Leib,

also sol der man mitt vernunfft
by dem wyb whonen und sy mitt wyßheyt
in früntligheit leyten.
(https://www.bullinger-digital.ch/letter/11252)

also soll der Mann mit Vernunft
bei dem Weibe wohnen und sie mit Weisheit
in Freundlichkeit leiten.

Literatur:

HBBW-Edition (1974-2019): Heinrich Bullinger, Werke, herausgegeben vom Institut für Schweizerische Reformationsgeschichte, Zweite Abteilung: Briefwechsel, Bände 1-19, Zürich.

Das VinKo-Korpus

Anne Kruijt / Stefan Rabanus / Marta Tagliani
(Universität Verona)

Das VinKo-Korpus ist ein Parallelkorpus, das aus Tonaufnahmen deutscher und italienischer Dialekte und Minderheitensprachen besteht, die in den italienischen Regionen Trentino-Südtirol, Venetien und Friaul-Julisch Venetien gesprochen werden. Die Daten werden über die Crowdsourcing-Plattform des VinKo-Projekts („Varietäten in/m Kontakt"; <https://www.vinko.it/index.php?lang=de>) gesammelt. Für die drei linguistischen Systemebenen Phonetik/Phonologie, Morphologie und Syntax sind Variablen so definiert, dass die Varianten sprachgruppenübergreifend (deutsch-italienisch) verglichen werden können. Bei der Erhebung wird eine Mischung aus traditionellen Übersetzungs- und Ausspracheaufgaben und innovativen Erhebungsmethoden in Form von Storyboard-Übersetzungsaufgaben und Aufgaben zur gesteuerten freien Sprachproduktion (Vervollständigung von konversationellen Paarsequenzen) verwendet. Der ursprüngliche, von der VinKo-Arbeitsgruppe verfolge Zweck der Datensammlung ist die Untersuchung von alter Mehrsprachigkeit, Sprachkontakt und sprachkontaktinduziertem oder -befördertem Sprachwandel im Untersuchungsgebiet. Im Mittelpunkt stehen die Tiroler Dialekte und germanische Minderheitensprachen wie Zimbrisch, Fersentalerisch, Sauranisch und Plodarisch einerseits, und romanische Varietäten wie das Ladinische und die Dialekte des Trentino und Venetiens andererseits.

Die VinKo-Arbeitsgruppe bekennt sich zum „open science"-Ansatz und zu den FAIR-Prinzipien für die Daten (*findable*, *accessible*, *interoperable*, *re-usable*). Aus diesem Grund werden die Daten der Forschung und der interessierten nicht-wissenschaftlichen Öffentlichkeit auf drei Arten zur Verfügung gestellt. (1) Eine Auswahl der Tonaufnahmen ist über eine interaktive Karte (GIS-System) und verschiedene Such- und Darstellungsfunktionen auf der VinKo-Webseite offen konsultierbar: <https://vinko.it/listen-explore.php?lang=de>. (2) Die Gesamtdaten sind erstens in einem passwortgeschützen Bereich der VinKo-Webseite zugänglich (Zugangsdaten können bei der VinKo-Arbeitsgruppe angefordert werden). (3) Zweitens kann das gesamte VinKo-Korpus (Rabanus et al. 2021) aus dem Repository des Eurac Research Clarin Centre (ERCC) in Bozen heruntergeladen werden (ERCC ist Teil der größeren europäischen CLARIN-Initiative [Common

Language Resources and Technology Infrastructure]). Zum gegenwärtigen Zeitpunkt enthält der Datensatz im Repository alle VinKo-Daten aus dem Zeitraum von Juni 2017 bis Mai 2021 und umfasst damit insgesamt 37.806 Tonaufnahmen. Der Datensatz wird regelmäßig aktualisiert, um neue Daten aufzunehmen (im passwortgeschützten Bereich der VinKo-Webseite sind heute [08.02.2022] 90.000 Tonaufnahmen gespeichert). Das VinKo-Korpus besteht aus drei Hauptordnern: dem Audio-Ordner, der die Tonaufnahmen enthält; dem Metadaten-Ordner, der Tabellen mit relevanten Informationen zu den Stimuli (v.a. Explikation der linguistischen Variable und deutsche, italienische bzw. dialektale Ausgangsform für die Aufgabe) sowie soziolinguistische Informationen über die Sprecher und damit auch die für die Analyse notwendigen Bezüge enthält; dem Bilder-Ordner, der die Bilder enthält, die als visueller Kontext für die Storyboard-Aufgaben verwendet werden. Alle Daten sind unter einer Creative-Commons-Attribution-Non-Commercial-Share-Alike-Lizenz lizenziert und können für nicht kommerzielle Zwecke frei verwendet werden.

Literatur:

Rabanus, Stefan, Alessandra Tomaselli, Andrea Padovan, Anne Kruijt, Birgit Alber, Patrizia Cordin, Roberto Zamparelli, and Barbara Maria Vogt. 'VinKo (Varieties in Contact) Corpus'. https://www.vinko.it, 2021. https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/32

Chinesisches Deutschlerner-Korpus (CDLK)

Yuan Li / Zekun Wu (Zhejiang Universität, Hangzhou)

Die deutsche Sprache erlebt in China seit der Jahrtausendwende einen Boom. Der neuesten Erhebung 2020 zufolge gibt es in allen Bereichen, also Schule, Hochschule und Erwachsenenbildung, 20% mehr Deutschlerner als noch vor 5 Jahren. Die bestehenden Forschungsprojekte zur Lernersprache chinesischer Deutschlerner verfolgten meist einen qualitativen und induktiven Ansatz und stützten sich selten auf umfangreiche Korpora. Weltweit gesehen enthalten nur einige Lernerkorpora (Falko, AleSKo, Kobalt und BFSU-Korpus) schriftliche Texte der chinesischen Deutschlerner. Die ersten drei wurden in Deutschland erstellt, enthalten nur wenige Texte und nur von fortgeschrittenen Studierenden (39, 43 bzw. 20 Texte auf dem Niveau B2). Das BFSU-Korpus aus China besteht aus Prüfungsaufsätzen chinesischer DeutschstudentInnen, ist aber nicht öffentlich zugänglich. Es ist zu erkennen, dass es bei den aktuellen Korpora nur begrenzte Textsorten und nicht umfassende Deutschniveaus sowie Lernstufen gibt.

Angesichts dieses Desiderats wird ein umfangreiches „Chinesisches Deutschlerner-Korpus" (CDLK) unter der Leitung von Prof. Dr. Yuan Li an der Zhejiang Universität aufgebaut. Das ist ein multidimensional annotiertes Lernerkorpus, das alle Niveaus bzw. Lernstufen abdeckt und Texte sowohl von SchülerInnen, als auch von StudentInnen enthält. Darüber hinaus werden die Texte von denselben Lernern innerhalb von drei Jahren regelmäßig gesammelt.

Im CDLK werden Texte mit unterschiedlichen Themen und aus verschiedenen Genres (argumentative, bildbeschreibende, deskriptive und erzählende Texte) im Unterricht von Deutschlernern aus unterschiedlichen Jahrgängen von Schulen und Universitäten in verschiedenen Regionen handschriftlich innerhalb von 30 Minuten ohne Hilfsmittel produziert. Ein wichtiger Teil des CDLK ist die Mehrebenenannotation, inklusive Fehlerannotation und Dependenz-Annotation. Die Fehlerkategorien wurden auf Basis Hirschmann und Noldas differenziert und bereichert, um Äußerungen chinesischer Deutschlerner kultursensibel und -spezifisch anzupassen. Die Abweichungen von den Originaläußerungen werden nicht nur auf der Normalisierungsebene (Orthografie, Morphosyntax, Morphologie, Lexik, Syntax, Semantik), sondern auch auf der höheren Ebene (Logik, Stil, Transfer) annotiert. Die Annotation wurde mit EXMARaLDA vorgenommen. Damit wurden die Zielhypothesen automatisch mit POS-, lemma annotiert. Dann wurden alle Abweichungen (ZHDiff) zwischen den Zielhypothesen und den Originaläußerungen automatisch annotiert, und manuell auf die jeweiligen Fehlerkategorien positioniert. Jeder Fehler wurde von drei Forschern gesondert identifiziert und kategorisiert. Bei Dependenz-Annotation wird Dependenzrelation zuerst automatisch mit SpaCy annotiert, dann manuell überprüft.

Auf der Grundlage des CDLK wurden bereits einige Studien durchgeführt. In Anlehnung an der Theorie der dynamischen Systeme und mit dem CAF-Rahmen wurde die Lernersprache querschnittlich oder längsschnittlich aus einer bestimmten Perspektive vergleichend durchgeführt: Auf der lexikalischen Ebene wurden die Verwendung von Konnektoren, Verben auf Basis der Valenzgrammatik, Präpositionen, Chunk sowie lexikalische Fehler untersucht; auf der syntaktischen Ebene konzentrieren wir uns derzeit auf die Entwicklung der syntaktischen Komplexität, die Wortstellung der Verben und die syntaktische Vielfalt; auf der textuellen Ebene wird die Diskurskohärenz thematisiert; darüber hinaus werden Faktorenkorrelationen erforscht, wie z.B. die Beziehung zwischen dem Einstiegsalter der Lerner und der Lerngeschwindigkeit sowie die Korrelation zwischen dem Einstiegsalter und der Korrektheit der schriftlichen Produktion.

Literatur:

Lüdeling, Anke & Merja Kytö (Hrsg.) (2009): Corpus Linguistics. An International Handbook. Vol 2. (Reihe Handbücher zur Sprach- und Kommunikationswissenschaft). Mouton de Gruyter, Berlin.

Nolda, Andreas (2019): Annotation von Lernerdaten mit EXMARaLDA (Dulko). Manuscript. Berlin-Brandenburgische Akademie der Wissenschaften, Berlin.

Reznicek, Marc; Anke Lüdeling; Cedric Krummes; Franziska Schwantuschke; Maik Walter; Karin Schmidt; Hagen Hirschmann & Torsten Andreas (2012): Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01.

Reznicek, Marc; Anke Lüdeling & Hagen Hirschmann (2013): Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In: Ana Díaz-Negrillo, Nicolas Ballier & Paul Thompson (Hrsg.) Automatic Treatment and Analysis of Learner Corpus Data. John Benjiamins, 101-124, Amsterdam / Philadelphia.

Yuan Li u.a. (2021): Das Handbuch CDLK. Version 2.0. (unveröffentliches Manuskript)

文秋芳，梁茂成 & 晏小琴 编著(2008)：中国学生英语口笔语语料库(2.0版)．外语教学与研究出版社，北京．(Wen, Qiufang; Maocheng Liang & Xiaoqin Yan (Hrsg.) (2008): Spoken and Written English Corpus of Chinese Learners. Version 2.0. Foreign Language Teaching and Research Press, Beijing.)

Swiss-AL: Korpus und Workbench für mehrsprachige digitale Diskurse

Julia Krasselt / Philipp Dreesen / Matthias Fluor / Klaus Rothenhäusler / Maren Runte / Cerstin Mahlow / Anna Ulasik (Zürcher Hochschule für Angewandte Wissenschaften)

Swiss-AL ist ein mehrsprachiges Korpus zur Analyse digitaler Diskurse in der Schweiz (Krasselt et al. 2020). Es wird seit 2016 am Departement für Angewandte Linguistik der Zürcher Hochschule für Angewandte Wissenschaften (ZHAW) entwickelt. Das Korpus enthält wichtige Akteure der Schweizer öffentlichen Kommunikationen aus den Bereichen Politik (z.B. Bundesämter und Parteien), Wirtschaft (z.B. Verbände und Spitäler), Wissenschaft (z.B. Hochschulen), Medien (Print und Online) und Zivilgesellschaft (z.B. NGOs). Die Medientexte stammen aus Schweizer Tages- und Wochenzeitungen sowie von Online-Newsportalen und werden systematisch über die Schweizer Mediendatenbank bezogen. Die Texte der anderen vier Bereiche werden mit Hilfe von Webcrawling kontinuierlich gesammelt. Dafür wird systematisch auf diejenigen Teile der Webseite eines Akteurs zugegriffen, die News, Medienmitteilungen und Blogeinträge enthalten. Die Erstellung des Korpus wird mit einer eigenen Verarbeitungspipeline prozessiert, mit je eigenständigen Komponenten für Datenverwaltung (mit Elastic Search), Preprocessing (insbesondere mit Komponenten zu Deduplikation und Scraping) und linguistische Verarbeitung (Wortarten, Dependenz-Parsing, Named Entity Recognition, semantische Annotationen). Das Korpus wird außerdem ergänzt durch Daten aus sozialen Medien (Twitter), Gesundheitsforen sowie Blogs (Fokus Rechtspopulismus).

Swiss-AL steht der breiten Öffentlichkeit zur Verfügung und wird in anwendungsorientierten Forschungsprojekten sowie in der linguistischen und kommunikationswissenschaftlichen Lehre genutzt. Dafür werden basierend auf spezifischen Modellierungskriterien Teilkorpora erstellt (Dreesen & Stücheli-Herlach 2019). Das Korpus wurde bisher für mehrsprachige Diskursanalysen eingesetzt, beispielsweise Diskurse zu erneuerbaren Energien, Covid-19, Antibiotikaresistenzen und Sozialhilfe. Die in Swiss-AL enthaltenen Medientexte bilden außerdem die Grundlage für die Wahl des Schweizer Wort des Jahres in den Sprachen Deutsch, Französisch, Italienisch und Rätoromanisch (Liste Lamas 2018).

Das Korpus kann über die an der ZHAW entwickelte Swiss-AL-Workbench genutzt werden (Krasselt et al. 2021). Die Workbench zeichnet sich durch eine leicht zugängliche und einfach gestaltete Benutzeroberfläche aus, bietet eine breite Palette von Analysemöglichkeiten und visualisiert Ergebnisse in unterschiedlicher Form. Die Workbench ermöglicht klassische korpuslinguistische Analysen, die für Diskursanalysen inzwi-

schen etabliert sind (Konkordanzen, Kollokationen, Keywords). Darüber hinaus können computerlinguistische Methoden genutzt werden, die sich insbesondere für vertiefte Analysen an der Schnittstelle von Semantik und Pragmatik eignen (Topic Modeling, Word Embeddings). Die Swiss-AL-Workbench basiert auf RShiny und erstellt Abfragen in Echtzeit mit Hilfe der IMS Corpus Workbench.

Auf der Methodenmesse der IDS-Jahrestagung 2022 sollen sowohl das Korpus als auch die verfügbare Workbench vorgestellt werden.

Link:

https://swiss-al.linguistik.zhaw.ch/rsconnect/workbench/

Literatur:

Dreesen, P., & Stücheli-Herlach, P. (2019). Diskurslinguistik in Anwendung. Ein transdisziplinäres Forschungsdesign für korpuszentrierte Analysen zu öffentlicher Kommunikation. Zeitschrift für Diskursforschung, 7(2), 123–162.

Krasselt, J., Fluor, M., Rothenhäusler, K., & Dreesen, P. (2021). A workbench for corpus linguistic discourse analysis. In D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, & B. Heinisch (Hrsg.), *3rd conference on language, data and knowledge* (LDK 2021) (Bd. 93, S. 26:1-26:9). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/OASIcs.LDK.2021.26

Krasselt, J., Dreesen, P., Fluor, M., Mahlow, C., Rothenhäusler, K., & Runte, M. (2020). Swiss-AL: A Multilingual Swiss Web Corpus for Applied Linguistics. Proceedings of the 12th Language Resources and Evaluation Conference, 4138-4144. https://www.aclweb.org/anthology/2020.lrec-1.510/

Liste Lamas, E. (2018). „Wort des Jahres Schweiz“ als Ergebnis systematischer Diskursmessung im mehrsprachigen Kontext. *Text und Diskurs.* 2018(11), S. 503-516. https://doi.org/10.7311/tid.11.2018.24

DONNERSTAG, 17. MÄRZ 2022, 9.30 UHR

Das GeWiss-Korpus: Neue Forschungs- und Vermittlungsperspektiven zur mündlichen Hochschulkommunikation

Christian Fandrych / Franziska Wallner (Herder-Institut / Universität Leipzig)

Im Vortrag sollen nach einer kurzen Vorstellung des gesprochensprachlichen Korpus GeWiss ausgewählte Forschungs- und Vermittlungsperspektiven, aber auch Begrenzungen präsentiert werden, die durch das Korpus systematisch angelegt sind. Dabei wird u.a. auch auf die neuen Zugangs- und Nutzungsmöglichkeiten des Korpus eingegangen, die im Rahmen des Projekts ZuMult (IDS Mannheim, Herder-Institut/Universität Leipzig, Universität Hamburg) entwickelt wurden.

Mit GeWiss (Gesprochene Wissenschaftssprache kontrastiv) liegt seit 2013 eine Ressource vor, die es ermöglicht, mündliche Hochschulkommunikation empirisch

genauer zu untersuchen. Es handelt sich dabei um ein Vergleichskorpus, das verschiedene mündliche Genres in philologischen Fächern aus unterschiedlichen akademischen Kontexten in deutscher, englischer, italienischer und polnischer Sprache bereitstellt. Das Korpus enthält „natürliche" (d.h. nicht zum Zweck der Korpuserstellung elizitierte) Sprachdaten, insbesondere dialogische Genres wie das Prüfungsgespräch sowie stärker monologisch geprägte kommunikative Ereignisse wie studentische Referate und Vorträge von Expert/-innen. Neben L1-Daten umfasst das Korpus auch L2-Produktionen für die Sprachen Deutsch und Englisch. Mit GeWiss eröffnen sich damit vielfältige Forschungs- und Anwendungsperspektiven. Aus sprachdidaktischer Perspektive – die hier im Mittelpunkt stehen soll – ist GeWiss in verschiedener Hinsicht von großer Relevanz: Zum einen können relevante sprachlich-kommunikative Vermittlungsgegenstände untersucht und beschrieben werden, für die es bisher noch keine empirische Grundlage gab; diese können zudem auch sprach- und kontextvergleichend betrachtet werden. Daneben ermöglicht es das Korpus, den Erwerb von gesprochensprachlichen, genrespezifischen Kompetenzen bei verschiedenen Studierendengruppen zu untersuchen. Nicht zuletzt eignen sich die Daten aber auch für Vermittlungszwecke – etwa zur Veranschaulichung von typischen Vortragseröffnungen, von ausgewählten sprachlichen Handlungen (z.B. mündliche Literaturverweise in Vorträgen) oder von kommunikativen Verfahren wie der Bearbeitung von Problemen in Prüfungsgesprächen. Zugänglich ist das Korpus über zwei Schnittstellen (das GeWiss-Portal und die DGD); diese Zugänge sind teils mit unterschiedlichen Nutzungsmöglichkeiten verbunden.

Da Korpora wie GeWiss (auch) mit dem Ziel des Transfers etwa in die Sprachdidaktik und die propädeutische Vorbereitung von Studierenden geschaffen wurden, wurde mittels einer Nutzungsstudie (Fandrych et al. 2016) erhoben, welche Voraussetzungen, Interessen und Nutzungsbedürfnisse gegenwärtige Nutzer/-innen haben. Es zeigte sich, dass viele interessierte Nutzer/-innen kaum über die relevanten korpuslinguistischen (Recherche-)Kenntnisse verfügen, die für eine zielgerichtete Abfrage des Korpus notwendig sind. Die Studie ergab zudem Nutzungsbedürfnisse, die so nicht direkt abfragbar waren, etwa Zugriffswege nach sprachdidaktisch relevanten Kriterien (wie z.B. Schwierigkeitsparametern), verbesserte Möglichkeiten für die didaktische Aufbereitung der Transkripte sowie die Kombinierbarkeit der unterschiedlichen Suchoptionen aus dem GeWiss-Portal und der DGD.

Als Folge wurde mit dem Projekt ZuMult eine Ressource geschaffen, die insbesondere neue Zugänge, Recherche- und Anwendungsmöglichkeiten bereitstellen soll, die in der (Fremd-)Sprachendidaktik, aber auch in verschiedenen weiteren Anwendungs- und Forschungsfeldern von besonderem Interesse sind. Die neu entwickelten Tools ermöglichen unter anderem

- eine gezielte Suche nach Sprechereignissen mit dem Werkzeug ZuMal, wobei verschiedene Metadaten- und schwierigkeitsbezogene Auswahlfilter genutzt werden können,
- eine vielfältige Analyse und didaktische Aufbereitung von Sprechereignissen mit dem Transkriptbrowser ZuViel
- sowie komplexe CQP-basierte Suchanfragen mit der Oberfläche ZuRecht.

Im Vortrag werden exemplarisch einige im Korpus angelegte vermittlungsrelevante Forschungsfragen vorgestellt und anhand von ersten Analysen verdeutlicht. Daneben wird eine Auswahl der neuen, im Kontext von ZuMult geschaffenen Werkzeuge zur Erforschung und Vermittlung der mündlichen Hochschulkommunikation vorgestellt. Abschließend werden einige wichtige Desiderate in Bezug auf die korpusbasierte Erforschung und Nutzbarmachung der Wissenschaftskommunikation aus sprachdidaktischer Perspektive diskutiert.

DONNERSTAG, 17. MÄRZ 2022, 10.10 UHR

Akustisches Signal, Mehrebenenannotation und Aufgabendesign. Flexible Korpusarchitektur als Voraussetzung für die Wiederverwendung gesprochener Korpora

Carolin Odebrecht / Malte Belz (Humboldt-Universität zu Berlin)

Eine Korpusarchitektur muss – insbesondere für Korpora nicht-standardisierter Sprache – verschiedene epistemische Anknüpfungspunkte für unterschiedliche Forschungskontexte bieten, wenn sie die Weiterentwicklung und Wiederverwendung von Korpora ermöglichen soll. In unserem Vortrag diskutieren wir drei besonders relevante Korpusarchitekturkomponenten für gesprochene Korpora: die Bereitstellung eines akustischen Signals, die konzeptionell und technisch damit verbundene Annotation in einem Mehrebenenmodell und das Aufgabendesign. Zudem stellen wir methodische Überlegungen zur Normalisierung und zum Forschungsdatenmanagement vor, die für iterative Verbesserungen und für die Wiederverwendung relevant sind. Exemplarisch hierfür ziehen wir das gesprochene Lernerkorpus WroDiaCo heran (Wesolek et al. 2021), welches für eine studentische Abschlussarbeit erhoben (Version 1) und schon für eine andere linguistische Fragestellung zu Normalisierungsmechanismen wiederverwendet wurde (Version 2, vgl. Belz & Odebrecht [angenommen]). Die methodischen Überlegungen werden anhand einer Forschungsfrage zum Fremdsprachnakzent (Vokalrealisierung polnischer Deutschlerner*innen) illustriert.

Literatur:

Belz, Malte & Carolin Odebrecht. [angenommen]. Abschnittsweise Analyse sprachlicher Flüssigkeit in der Lernersprache: Das Ganze ist weniger informativ als seine Teile. Zeitschrift für germanistische Linguistik.

Wesolek, Sarah, Malte Belz & Christine Mooshammer. 2021. Wroclaw Dialogue Corpus (WroDiaCo): Version 2. Medien-Repositorium. <https://rs.cms. hu-berlin.de/phon>.

Korpusaufbau zwischen Standard und Innovation (am Beispiel von GiesKaNe)

Volker Emmrich / Mathilde Hennig (Universität Gießen)

Die Auseinandersetzung mit Standards ist zentral für jedes Korpusprojekt und muss unter Aspekten wie Forschungsinteresse, wissenschaftliche Originalität, eristischer Diskurs und Usability betrachtet werden. In einer solchen kritischen Auseinandersetzung wird deutlich, dass Standards einerseits im Konflikt mit Grundprinzipien wissenschaftlicher Forschung wie Innovation und Originalität stehen, während sie andererseits unter Gesichtspunkten wie Usability und Vergleichbarkeit von Daten unverzichtbar für die Einbettung eines neuen Korpus in die Korpuslandschaft sind. Beide Aspekte müssen aber keineswegs als einander ausschließende Alternativen betrachtet werden, da über die Möglichkeit der Mehrebenenannotationen die datenstrukturellen Voraussetzungen für einander ergänzende innovative und standardkonforme Annotationen gegeben sind und durch die Methoden des maschinellen Lernens eine effiziente Ergänzung möglich sein sollte: Wenn qualitative hochwertige komplexe Annotationen nach einem Annotationsschema vorliegen, können diese als unabhängige Variablen bzw. als Eingabe für die Vorhersage der Analyse nach einem anderen Annotationsschema dienen.

Überlegungen dieser Art werden im Vortrag am Beispiel von GiesKaNe (= GießenKassel-Neuhochdeutsch) präsentiert. Das aktuell im Rahmen eines DFG-Langfristvorhabens an den Standorten Kassel (Leitung: Vilmos Ágel) und Gießen (Leitung: Mathilde Hennig) im Aufbau befindliche Korpus arbeitet im Rahmen der Zielsetzung der Erschließung syntaktischer Grundstrukturen des Neuhochdeutschen mit innovativen Ansätzen, die der Komplexität der Annotationsziele (syntaktische Grundstrukturen von Text über Satz bis hin zu Wortgruppe und Wort, systematische Form-Funktionszusammenhänge, Syntax-Semantik-Beziehungen) gerecht werden sollen. Herzstück ist ein eigener Baumbankansatz, aber auch das Wortarttagging folgt einem neuen Modell. Das Ziel des geplanten Vortrags besteht darin, am Beispiel der Wortartannotation (GiesKaNe vs. STTS und HiTS) zu zeigen, dass mit Innovationen dieser Art einerseits neue Akzente in der Korpuslinguistik gesetzt werden können, dass sie aber nicht in eine idiosynkratische Einbahnstraße führen müssen, wenn man die im vorliegenden Modell vorhandenen Annotationen systematisch für die Rekonstruktion anderer Modelle nutzt und diese in der Mehrebenenannotation ergänzend für die Korpusnutzung anbietet.

Multimodale und agile Korpora im Forschungshub #DigitalHerrnhut

Alexander Lasch (TU Dresden)

In Dresden entsteht für den Forschungshub #DigitalHerrnhut der Pilot für ein agiles und multimodales Referenzkorpus der nächsten Generation (NexGen Agile Reference Corpus (NARC)) in Zusammenarbeit mit der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB). Dieses Korpus (N-ARC1) wird textliche, kartografische und audiovisuelle Quellen sowie weitere Artefakte fassen, die, miteinander vernetzt, als offene Forschungsdaten (teil-)maschinell angereichert werden können und in einer virtuellen Forschungsumgebung öffentlich und nachnutzbar zur Verfügung stehen sollen. Die Dokumente und Spuren der Herrnhutischen Brüdergemeine – eine am Beginn des 18. Jh. gegründete und in nur wenigen Jahrzehnten weltumspannende Glaubensgemeinschaft – bieten dafür den idealen Ausgangspunkt. Sie eröffnen der interdisziplinären Forschung einen unvergleichbar reichen Wissensschatz, um globale Auswirkungen europäischer Sendungskultur zu verstehen. N-NARC1 wird die systematische Er- und Bearbeitung von Forschungsgegenständen im weltweit vernetzten Forschungshub #DigitalHerrnhut aus den verschiedensten wissenschaftlichen Perspektiven – linguistisch, geo-informatisch, kulturhistorisch, landesgeschichtlich, landschaftarchitektonisch, botanisch, theologisch etc. – ermöglichen. Insbesondere verspricht die herrnhutische, weltweit archivierte und in großen Teilen noch nicht erschlossene Datengrundlage, die Methoden der Digital Humanities synergetisch anzuwenden und vorbildlich in Forschung, Lehre wie CitizenSciences auszustrahlen. N-ARC1 steht in den Digitalen Sammlungen der SLUB nachhaltig und offen zur Verfügung und es befördert (z.B. postkoloniale) Anschlussforschung zum einen durch stetige Erweiterung und zum anderen dank entwickelter Infrastrukturen.

Im Beitrag werde ich exemplarisch an drei Phänomenen die Herausforderungen illustrieren, vor denen wir derzeit in der Umsetzung in Dresden stehen, die (1) auf CitizenScience, (2) internationale Kooperation und (3) einzelprojektübergreifende Kollaboration fokussieren. (Ad 1) In Forschungsvorhaben wird der Datenerschließung leider nur noch eine untergeordnete Rolle zugestanden, diese ist im Kontext von #DigitalHerrnhut aber mitzudenken. Ohne Archivarbeit, Digitalisierung und großflächige Erschließung handschriftlicher Quellen in erheblichem Umfang ist fast keine wissenschaftliche Arbeit im Kontext der Herrnhuter Brüdergemeine vorstellbar. Dafür wird automatische Optical Character Recognition (OCR) immer notwendiger, die aber trainiert werden muss, wofür zunächst Ground Truth herzustellen ist. Diese erreichen wir über einen unkonventionellen Weg: Wir lassen einen Teil unserer Texte von kompetenten Leser:innen einlesen und transkribieren die eingelesenen Texte automatisch, um sie anschließend mit den Digitalisaten zu alignieren. So erreichen wir eine Basis für die automatische Erkennung. Die Leser:innen sind eingeladen, aus der Lektüre heraus eigene Fragen an die Texte zu entwickeln – wir bieten hierfür eigene Veranstaltungsformate mit verschiedenen

Trägern im Verbund Dresden CONCEPT an. Die zweite Herausforderung ist (ad 2) die internationale Kooperation, die bei einem Projekt wie #DigitalHerrnhut erforderlich ist: Auf welche technischen Standards einigt man sich, welche virtuellen Forschungsumgebungen sind aufzubauen? Setzen wir, wie in Europa, auf OpenSource-Lösungen? Wenn ja, wie arbeiten wir mit unseren Kolleg:innen in den Vereinigten Staaten oder Großbritannien und Skandinavien zusammen, die meist kommerzielle Anwendungen nutzen? Im Moment suchen wir z.B. im IVAC-DAAD-Projekt *Conceptualizing Languages in Colonization* hierfür brauchbare Lösungen. Eine dritte Herausforderung (3) betrifft die digital gestützte Dissemination in Forschung und Lehre (gefördert z.B. im Projekt *virTU-os* durch die Stiftung „Innovation in der Hochschullehre"). Es geht darum, ein Verständnis vom Gesamtbild der Überlieferung zu erzeugen, das sich nicht einstellt, wenn aus Qualifikationsphasen dem Gesamtbild immer nur kleine Puzzlestücke an verstreuten Orten beigefügt werden. Wir tragen diese deshalb zusammen und laden zur Mitarbeit ein. So ist es uns gelungen, in unserem aktuellen, noch überschaubaren Korpus bspw. 220 Lebensbeschreibungen aus zwei Dissertationsprojekten zu integrieren. Des Weiteren platzieren wir ausgewählte Quellen in „Virtuellen Exkursionen", die in der Lehre entstehen, als Showcases zum Einsatz kommen und es Studierenden erlauben, nicht nur „Kleinwelke" in der Nähe von Dresden, sondern mittelfristig auch Bethlehem in den Vereinigten Staaten, oder langfristig Gnadental in Südafrika zu besuchen. Sie sind uns Ansatzpunkte, um die Verzahnung von Mission und Bildung in Selbstzeugnissen der Gemeinschaft kennen- und verstehen zu lernen.

Literatur:

Alexander Lasch. 2021. (Wissenschafts-)Kultur der Digitalität. In: Lessons Learned I. DOI: https://doi.org/10.25369/ll.v1i1/2.27.

Alle für den Beitrag relevanten Onlineressourcen sind hier zusammengestellt und werden fortwährend aktualisiert: https://dhh.hypotheses.org/ids2022.

IDS 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
www.facebook.com/ids.mannheim
@IDS_Mannheim

Leibniz Gemeinschaft