Originalveröffentlichung in: Deppermann, Arnulf/Fandrych, Christian/Kupietz, Marc/Schmidt, Thomas (Hrsg.): Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial. – Berlin/Boston: de Gruyter, 2023. S. IX-XI. (Jahrbuch / Leibniz-Institut für Deutsche Sprache 2022)
DOI: *https://doi.org/10.1515/9783111085708*

Arnulf Deppermann (Mannheim)/Christian Fandrych (Leipzig)/
Marc Kupietz (Mannheim)/Thomas Schmidt (Mannheim/Basel)

Zur Einführung: Korpora in der germanistischen Sprachwissenschaft

Im Jahre 2006 fand die Jahrestagung des IDS zum Thema „Sprachkorpora – Datenmengen und Erkenntnisforschritt" statt (Kallmeyer/Zifonun (Hg.) 2007). Die heutige Welt der Korpora ist mit der von vor 16 Jahren nicht mehr zu vergleichen: *big data* und *digital humanities*, die Entwicklung avancierter statistischer Werkzeuge der Korpusanalyse, neuartige Korpora von Video- und *social media*-Daten, automatisierte Workflows und Interoperabilität, Standards und *best practices*, ein grundlegend verändertes Bewusstsein für die rechtliche Absicherung von Korpora, die Verfügbarkeit von Korpusdaten über das Internet und vieles andere mehr haben den Fortschritt und die Diversifizierung der Korpuslandschaft seither geprägt.

Waren zu Beginn des Millenniums viele linguistische Forschungsartikel noch intuitionsbasiert oder philologisch („armchair linguistics", Fillmore 1992), so ist heute eine empirische Datenbasis in den allermeisten Zweigen der Linguistik selbstverständlich geworden. Diese Datenbasis besteht in der überwiegenden Mehrzahl aus Korpora. Die Linguistik ist damit (mit Verspätung z. B. gegenüber der Psychologie und Soziologie) zu einer empirischen Wissenschaft geworden. So haben auch in der Linguistik statistische Methoden Einzug gehalten, doch auch neue Formen der qualitativen Datenanalyse haben sich entwickelt. Zugleich sind aber Korpora im Vergleich zu anderen Disziplinen das methodologische Alleinstellungsmerkmal der Linguistik: Es ist für die Linguistik distinktiv, dass für sie die Rohdaten authentischer gesellschaftlicher Praxis, also Texte und Interaktionen, das Untersuchungsmaterial und der Gegenstand ihrer Beschreibungen und Erklärungen sind. Darin unterscheidet sie sich von den allermeisten Arbeiten in den Sozialwissenschaften, in denen von vornherein codiert, abstrahiert und aggregiert wird und Daten somit nur durch den Filter von Operationalisierungen, Zusammenfassungen und forscherseitigen Interpretationen zum Untersuchungsmaterial werden.

Wie es für Jahrestagungen des IDS üblich ist, hatte die Jahrestagung 2022 das Ziel eine Bestandsaufnahme zu leisten. Dabei geht es einerseits um einen Überblick über die mittlerweile sehr vielfältige und avancierte Korpuslandschaft in der germanistischen Sprachwissenschaft, also um die etablierten maßstabsetzenden Korpora und neue Entwicklungen. Es geht andererseits aber mehr noch um die Frage, wie Korpora für die Untersuchung verschiedenster linguistischer

https://doi.org/10.1515/9783111085708-001

Fragestellungen, z. B. der Lexikografie, der Gesprächsforschung, des Spracherwerbs oder der historischen Sprachwissenschaft, genutzt werden können.

In den Beiträgen dieses Bandes werden daher sämtliche Aspekte der Erstellung und Nutzung von Korpora angesprochen:

- Das Design von Korpora: ihre Zusammenstellung, Fragen der Datenerhebung, Datenqualität, und Datenvollständigkeit,
- die Forschungssoftware zur Annotation, Erschließung, Auswertung und Visualisierung, die für die Arbeit mit Korpora benötigt wird und hilfreich ist,
- die mit den Fragen von Design und Korpusaufbereitung zusammenhängenden linguistischen, aber bei bestimmten Korpora z. B. auch soziologischen oder historischen theoretischen Fragestellungen,
- der Zusammenhang von Korpusaufbereitung und Nutzungsmöglichkeiten bzw. Forschungsfragestellungen,
- ethische und rechtliche Aspekte der Korpussammlung, -aufbereitung, -bereitstellung und -nutzung, also v.a. Datenschutz und Urheberrecht.

Diese Fragen werden im Kontext wissenschaftstheoretischer Überlegungen zur Frage des Nutzens von Korpora für die linguistische Erkenntnisbildung und im Kontext spezifischer fachwissenschaftlicher Fragestellungen, die exemplarisch die Verwendung von Korpusdaten und -funktionalitäten im Kontext konkreter Forschungsvorhaben zeigen, behandelt.

Zu Beginn dieses Bandes stehen synchrone Sprachkorpora im Vordergrund. Das am IDS aufgebaute deutsche Referenzkorpus, DeReKo, wird von Marc Kupietz, Harald Lüngen und Nils Diewald vorgestellt. Sie geben Einblicke in die Korpuskonstruktion und in deren Relevanz für die Entstehung von Untersuchungsergebnissen, die Korpusnutzern oft unbekannt sind und entsprechend bei der Gewinnung von Untersuchungsergebnissen übersehen werden. Das an der Berlin Brandenburgischen Akademie der Wissenschaften (BBAW) beheimatete ZDL-Regionalkorpus wird von Andreas Nolda, Adrien Barbaresi und Alexander Geyken vorgestellt. Dieses Korpus, das Regional- und Lokalseiten von Tageszeitungen beinhaltet, erlaubt den systematischen Vergleich vor allem der regionalen lexikalischen Variation in der standardnahen Schriftlichkeit im deutschen Sprachraum. Alexandra Lenz stellt die österreichische Korpuslandschaft vor. Im Zentrum stehen sowohl ältere Dialektkorpora als auch neuere Medienkorpora und die Sammlungen des Sonderforschungsbereichs „Deutsch in Österreich“.

Anschließend widmet sich dieses Jahrbuch den Gesprächs- und Diskurskorpora. Silke Reineke, Arnulf Deppermann und Thomas Schmidt stellen das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) des IDS vor, das eine große Sammlung von Audio- und Videoaufnahmen authentischer sozialer Interaktionen beinhaltet. Es wird an einer exemplarischen Untersuchung gezeigt, wie

das Korpus speziell für interaktionslinguistische Fragestellungen genutzt werden kann. Ebenfalls ein Gesprächskorpus ist das Parallel European Corpus of Informal Interaction (PECII), das deutsche, französische, italienische, polnische und finnische Daten enthält. Laurenz Kornfeld, Uwe Küttner und Jörg Zinken stellen die Konzeption dieses Korpus vor, das sprach- und situationsvergleichende Untersuchungen zur Sprachverwendung in der sozialen Interaktion ermöglicht. Ein weiteres komparatives Korpus mit natürlichen Gesprächsdaten ist das GeWiss-Korpus zur Hochschulkommunikation, das deutsche, englische, italienische und polnische Aufnahmen unterschiedlicher mündlicher akademischer Genres enthält. Christian Fandrych und Franziska Wallner präsentieren neue Erschließungsinstrumente des Korpus und ihre Nutzung für die Untersuchung von Fragestellungen der Sprachvermittlung. Über diskursanalytische Schriftkorpora berichtet Marcus Müller. Er bespricht insbesondere die Verfahren der Repräsentation von Kontexten und Annotationen, die im *Discourse Lab* der Universitäten Darmstadt und Heidelberg angewendet werden.

Abschließend geht es um spezielle Fragen des Korpusdesigns. Am Beispiel eines Lernerkorpus diskutieren Carolin Odebrecht und Malte Belz Kriterien der Wiederverwendbarkeit von Korpora in Bezug auf die Bereitstellung akustischer Daten, die Mehrebenenannotation und das Aufgabendesign. Volker Emmrich und Mathilde Hennig stellen in ihrem Beitrag die Frage der Standardisierung der Korpusannotation, ihre Probleme und Notwendigkeiten in den Vordergrund. Anhand der Arbeiten zur Wortartenannotation im frühneuhochdeutschen GiesKaNe-Korpus werden Möglichkeiten, wie Standardisierung des Korpus und innovative Korpuserschließung gleichermaßen ermöglicht werden können, diskutiert. Im letzten Beitrag berichtet Alexander Lasch vom Forschungshub #DigitalHerrnhut, das textuelle, kartografische und audiovisuelle Quellen umfasst. Das besondere Augenmerk gilt hier den Möglichkeiten der Citizen Science, Modalitäten internationaler Kooperation sowie der Dissemination in Forschung und Lehre.

Literatur

Fillmore, Charles J. (1992): „Corpus linguistics“ or „Computer-aided armchair linguistics“. In: Svartvik, Jan (Hg.): Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991. (= Trends in Linguistics. Studies and Monographs 65). Berlin/New York: De Gruyter Mouton, S. 35–60.

Kallmeyer, Werner/Zifonun, Gisela (Hg.) (2007): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. (= Jahrbuch des Instituts für Deutsche Sprache 2006). Berlin: De Gruyter.