
Claudia Posch, Karoline Irschara, Gerhard Rampl (Hg.)
Wort – Satz – Korpus:
Multimethodische digitale
Forschung in der Linguistik
innsbruck university press

EDITED VOLUME SERIES

innsbruck university press

universität innsbruck

Claudia Posch, Karoline Irschara, Gerhard Rampl (Hg.)

Wort – Satz – Korpus: Multimethodische digitale Forschung in der Linguistik

Claudia Posch, Karoline Irschara, Gerhard Rampl
Institut für Sprachwissenschaft, Universität Innsbruck

Diese Publikation wurde mit finanzieller Unterstützung der Österreichischen Akademie der Wissenschaften und des Vizerektorats für Forschung der Universität Innsbruck gedruckt.

Universität Innsbruck
1. Auflage
Coverbild: Adelheid Schmid-Nuss
Layout: Romana Fiechtner

www.uibk.ac.at/iup
ISBN 978-3-99106-061-1

Inhaltsverzeichnis

Vorwort und Danksagung

Der Sammelband *Wort – Satz – Korpus: Multimethodische digitale Forschung in der Linguistik* dokumentiert einige Vorträge, die im Rahmen eines Workshops zur Korpuslinguistik bei der 42. Österreichischen Linguistiktagung 2018 an der Universität Innsbruck präsentiert und diskutiert wurden[1]. Versammelt waren Vertreter*innen verschiedener Ansätze, die innerhalb der Linguistik mit Textkorpora (in unterschiedlichen Definitionen) arbeiten. Der vorliegende Sammelband enthält eine Auswahl dieser im weitesten Sinn korpuslinguistischen Beiträge.

Wie viele Arbeiten in dieser Zeit hat sich die Veröffentlichung dieses Bandes unter der weltweiten Krise durch den Ausbruch von COVID19 verzögert. Dies ist einerseits bedauernswert, andererseits hat sich gerade im letzten Jahr die Wichtigkeit der Arbeit mit dem Digitalen sehr klar herauskristallisiert. Wir danken unseren Beiträger*innen deshalb für die Geduld und freuen uns, dass die gemeinsame Arbeit nun in Form dieses Bandes Früchte trägt.

Wir danken dem iup Verlag für die Aufnahme in die Reihe *EDITED VOLUME* und den Mitarbeiter*innen für die verlagsseitige Unterstützung.

Innsbruck, November 2021

1 Nicht alle Vorträge der Tagung wurden in den Sammelband aufgenommen, es wurden einige aus Termingründen abgesagt.

Claudia Posch, Karoline Irschara, Gerhard Rampl

Multimethodische digitale Forschung in der Linguistik: Ansätze, Herausforderungen, Perspektiven

In den vergangenen Jahrzehnten haben eine Reihe von *turns* die Forschung in den Geisteswissenschaften geprägt. Derzeit befinden wir uns mit der Herausbildung der Digital Humanities zweifellos innerhalb eines *digital turn* (Flanders/Jannidis 2019: 3) und es ist noch nicht absehbar, wie dieser die einzelnen Fachgebiete verändern wird. Dies betrifft insbesondere auch die Linguistik, welche sich schon früh mit digitalen Methoden beschäftigt hat (Jensen 2014: 115). Die Korpuslinguistik war grundlegend für die anfänglichen Entwicklungen in den Digital Humanities. Inzwischen haben digitale Arbeitsweisen auf nahezu alle Beschäftigungsfelder der Linguistik übergegriffen und geben dieser Entwicklung nunmehr eine neue Dimension. Sowohl die Digital Humanities als auch die von der Informatik geprägte maschinelle Sprachverarbeitung haben sich etabliert. Wie kann sich also die Korpuslinguistik in diesem neuen Kontext positionieren?

In den anglophonen digitalen Geisteswissenschaften ist, nach kritischen Einwänden aus dazu gehörenden Einzeldisziplinen, eine Debatte um die Digital Humanities entbrannt beziehungsweise hat eine teilweise Neuorientierung stattgefunden. Jensen sieht diese Redefinition und Neuorientierung der Digital Humanities als eher qualitativ an und befürchtet, dass diese Entwicklung zu einem Ausschluss der quantitativ orientierten Korpuslinguistik aus den Digital Humanities führen könnte (Jensen 2014: 131). Ein Beispiel hierfür ist der Hashtag #transformDH, der im Rahmen der Tagung *American Studies Association's annual conference* entstand und der zur Gründung eines sogenannten „academic guerrilla movement" führte, welches die „capital letter" Digital Humanities neu definieren möchte (https://transformdh.org/about-transformdh/). Auch im Sammelband *Debates in the Digital Humanities* wird

für eine Hinwendung zum *Putting the Human Back into the Digital Humanities* (Losh et al. 2016) plädiert.

Andererseits hinterfragen Bubenhofer (2018) bzw. Bubenhofer/Dreesen (2018) in zwei forschungstheoretischen Beiträgen zu Recht, ob nicht eher zu befürchten sei, dass gerade die Explosion quantitativer Ansätze (Stichwort NLP) die Linguistik entbehrlich und linguistische Fragestellungen obsolet mache. Bubenhofer verweist auf die Fülle aktueller Publikationen in den hochrangigen Fachzeitschriften der Disziplin, die teilweise überhaupt nicht mehr mit konkreten Instanzen und Beispielen aus Texten arbeiten, sondern ausschließlich deskriptive, analytische und multifaktorielle Statistiken verwenden, um z. B. Textsorten zu klassifizieren oder maschinelle Übersetzungen zu verbessern etc. (Bubenhofer 2018: 18). Geht der Trend also weg von linguistischen Modellen und Kategorien hin zu Black-Box-Systemen? Oder kann die Linguistik im Sinne einer „antifragilen" (Bubenhofer/Dreesen 2018) Disziplin gegenwärtige Umbrüche nutzen, um gestärkt daraus hervorzugehen? Reine sogenannte Black-Box Systeme können inzwischen sprachliche Informationen sehr verlässlich extrahieren, ermöglichen aber keine Erklärung der dahinter liegenden Phänomene (Bubenhofer 2018: 21f). Bubenhofer plädiert letztendlich dafür, korpuslinguistische Analysen als „neue Daten" zu fassen, die ausgewertet und interpretiert werden müssen und es scheint gerade so, dass viele linguistische Fragen dort beginnen, wo textmining-basierte Fragen enden (Bubenhofer 2018: 25).

Was sind also die aktuellen Interessensgebiete der Korpuslinguist*innen? Schon die Uneinigkeit bei der Definition des Fachgebiets *Korpuslinguistik* (vgl. Lemnitzer/Zinsmeister 2015; Scherer 2014; Perkuhn/Keibel & Kupietz 2012; McEnery/Hardie 2012) zeigt, dass diese weder klar als Methode, noch als reines Hilfsmittel der Linguistik gesehen wird. Sie hat sich inzwischen vielmehr zu einem breitgefächerten Gebiet entwickelt, wie beispielsweise Beiträge zur raumorientierten Korpuslinguistik (Gregory/Hardie 2011) oder die zahlreichen diskursorientierten Arbeiten zeigen. So beobachten Nartey et al. in ihrer Metastudie seit 2008 „a massive surge" (Nartey/Mwinlaaru 2019: 214) in Bezug auf die Anzahl der publizierten Arbeiten, die Diskursanalyse mit Korpuslinguistik verknüpfen.

Der vorliegende Sammelband möchte in die unterschiedlichen Handlungsfelder der Korpuslinguistik eintauchen. Er umfasst Beiträge zur Genese und Zusammenstellung von Korpora (corpus building), zur datengeleiteten (corpus-driven) Auseinandersetzung mit Korpora, zur korpusgestützten historischen Linguistik sowie zur korpusgestützten Diskursforschung. Sowohl quantitative Techniken der Korpuslinguistik als auch qualitative Methoden der Diskursanalyse, wie z. B. das close-reading von zusammenhängenden Textsegmenten, sind in dem Band vertreten.

Eine Erweiterung des traditionellen Verständnisses von Korpuslinguistik in Richtung eines offenen Feldes maschinell arbeitender Ansätze, die von linguistischen Fragestellungen geleitet sind bis möglicherweise hin zu einer Digitalen Linguistik könnte eine Bereicherung in der oben genannten antifragilen Disziplin sein. Wir hoffen, dass der vorliegende Band zu diesem Verständnis beitragen kann.

Die Beiträge in diesem Band

Iris Jammernegg, Sonja Kuri und Federico Collaoni verwenden korpus- und diskurslinguistische Methoden, um diskursiven Aushandlungen in Bezug auf erneuerbare Energien nachzuspüren. Ihr Beitrag *Diskursive Implementation der Energiewende in Deutschland: Eine linguistische Analyse zu Diskursorganisation und Wissensaufbau anhand von drei Akteuren des deutschen Energiesektors* untersucht, wie der Diskurs zur Energiewende im gegenwärtigen Deutschland maßgeblich durch drei unterschiedliche Energieversorgungsakteure geprägt wird. Dafür werden multimodal die Webauftritte und Schlüsseltexte von *RWE*, dem *Ministerium für Energiewende, Landwirtschaft, Umwelt, Natur und Digitalisierung* von Schleswig-Holstein und den *Bürgerwerken* analysiert. Die im Beitrag dargestellten Daten beziehen sich auf den Zeitraum zwischen August und Oktober 2018 und wurden aus einem größeren Gesamtkorpus extrahiert, welches Online-Auftritte von jeweils 14 deutsch- und italienischsprachigen Akteuren öffentlicher oder wirtschaftlicher Sektoren bzw. von Non-Profit-Organisationen der Energieversorgung umfasst (insgesamt 1.000 Textdokumente, 685.000 Tokens). Die Analyse reiht sich methodisch in den

Bereich der Corpus-Assisted Discourse Studies (CADS) ein: Durch die Ermittlung von Keywords, Kookkurrenzen und Kollokationen, aber auch durch den Einbezug textueller Makrostrukturen, audiovisueller Elemente und intertextueller Bezüge werden kognitive Frames der jeweiligen Akteure eruiert.

Rosemarie Lühr behandelt in ihrem Beitrag *Diskursforschung in alten Sprachen* die Frage, ob sich das für gegenwärtige Sprachen angenommene Konzept von Nähe- vs. Distanzdiskursen auch in einem Korpus alter Sprachen aufspüren lässt. Als Untersuchungsgrundlage dient das diachrone, geparste Kontrastkorpus ISAIS (*Informationsstruktur in älteren indogermanischen Sprachen,* ca. 125.970 Tokens, verfügbar in der ANNIS-Datenbank https://korpling.org/annis3/#_c=SVNBSVNfMS4w). Es enthält Texte u.a. für Latein, Indisch, Griechisch, Avestisch, Hethitisch, Luwisch und Lykisch und ist hinsichtlich unterschiedlicher informationsstruktureller Parameter wie *saliency*, *topic*, *focus*, *information-particle*, *discourse*, *style* usw. getaggt. Unter der Berücksichtigung von Diskurs- und Syntaxkonfigurationalität werden emphatische Strukturen der Wortstellung für die einzelnen Sprachen abgefragt. Im Zentrum der Analyse stehen jeweils initiale Topiks und Hyperbata, die durch differenzierte korpuslinguistische Analysen jeweils Nähe- und/oder Distanzdiskursen zugeordnet werden.

Tony McEnery und Gavin Brookes beschäftigen sich in ihrem Beitrag *Register, belief and violence: A multi-dimensional approach* mit Texten, die zu extremistischer Gewalt aufrufen. Die Untersuchungsgrundlage bildet ein Korpus aus Texten, die im Besitz von verurteilten dschihadistischen Terroristen gefunden wurden (275 Texte, 3.983.432 Tokens). Die Texte – dabei handelt es sich u.a. um Zeitungsartikel, Interviewtranskripte, Vorlesungen, Biographien usw. – wurden von Expert*innen für Terrorismusbekämpfung hinsichtlich ideologischer Gesichtspunkte kategorisiert und jeweils entweder als *moderat*, *randständig* oder *extrem* eingestuft. Durch eine multidimensionale Analyse gehen die Autoren datengeleitet der Frage nach, inwiefern die dschihadistischen Texte ein individuelles, homogenes Register abbilden oder aber ein bestimmtes Spektrum an unterschiedlichen Registern bespielen: Je nach Grad des Extremismus werden Ähnlichkeiten und Unterschiede der Registermerkmale zwischen den einzelnen Extremismus-Klassifikationen herausgearbeitet und analysiert. McEnery/Brookes berücksichtigen den Einfluss

der jeweiligen Autoren und der Textproduktion (schriftlich vs. mündlich) auf die Texte und untersuchen auch, inwiefern Textsorten und Registervariation zusammenspielen.

Gerhard Rampl und Claudia Posch gehen in ihrem Beitrag *Methodisch geht es korpuswärts! Zur Produktivität des Suffixes -wärts* der Frage nach, ob bzw. inwiefern das Suffix *-wärts* als produktiv erachtet werden kann: Ausgangspunkt ist das thematisch fokussierte Korpus Alpenwort – Korpus der Zeitschrift des österreichischen Alpenvereins, welches alle Bände der als Jahrbuch erscheinenden Zeitschrift von 1870 bis 2010 (19,9 Millionen Wörter) enthält. Keyword-Analysen förderten verschiedene Raum- und Direktionaladverbien zu Tage, darunter mehrere Bildungen mit dem Suffix *-wärts*, welche im Korpus unterschiedlich verteilt sind und diachron betrachtet eine signifikante Keyness aufweisen. Mittels der als Produktivitätsindizes geläufigen Type-Token-Ratio und Hapax-Token-Ratio Berechnungen untersuchen Rampl/Posch, wie sich das Suffix *-wärts* diachron entwickelt und ob sich diese Entwicklung auch in anderen Korpora (Text+Berg, DWDS, DeReKo und Schweizer Textkorpus) bzw. Textsorten zeigen lässt. Eine Analyse der relativen Tokenfrequenz zeigte, dass in Texten mit Alpinbezug direktionale Adverbien häufiger vorkommen als in anderen Textsorten; außerdem entpuppte sich das Suffix *-wärts* in den untersuchten Texten als sinkend produktiv. Rampl/Posch illustrieren ferner auf einer allgemeineren Ebene, welche methodischen Fragestellungen bei quantitativen Analysen im Vergleich zwischen unterschiedlichen Korpora berücksichtigt werden müssen.

Im Beitrag *Digital echo of mountains: Content indexing of alpine texts* liefern Elisabeth Gruber-Tokić, Gerald Hiebel, Gerhard Rampl und Claudia Posch einen Einblick in die Semantik alpiner Diskurse: Im Fokus steht das Alpenwort Korpus (Posch/Rampl 2017), welches die digitalisierten Zeitschriften des Deutschen und Österreichischen Alpenvereins (ZAV) von 1870 bis 2010 umfasst. Der Beitrag bespricht, wie dieses Korpus im Rahmen des Projekts *Semantics for Mountaineering History* semantisch annotiert wird und welche Herausforderungen dabei berücksichtigt werden müssen. Die semantische Annotation bezieht sich dabei auf die in den Daten auftretenden Orte, Personen und alpinen Aktivitäten (z.B. Erstbesteigungen), welche aus den Daten extrahiert und mit Hilfe von TEI und CIDOC CRM Standards se-

mantisch repräsentiert werden. Darüber hinaus illustrieren die Autor*innen, welche Probleme auf der automatisierten Ebene von Named Entity Recognition (NER) und Named Entity Linking (NEL) auftauchen und wie diese methodisch gelöst werden können.

Ein weiteres spezialisiertes Korpus wird im Beitrag *Building the MedCorpInn Corpus: Issues and goals* behandelt: Karoline Irschara, Claudia Posch, Birgit Waldner, Anna-Lena Huber, Bernhard Glodny, Leonhard Gruber und Stephanie Mangesius dokumentieren zentrale methodische und technische Schritte der Erstellung des MedCorpInn Korpus, welches über 5 Mio. Befunde der Innsbrucker Universitätskliniken für Radiologie und Neuroradiologie aus den Jahren 2007–2019 enthält. Neben Ausführungen zu den sprachlichen Besonderheiten dieser medizinischen Daten und daraus resultierenden Herausforderungen für die Korpuserstellung, heben die Autor*innen Fragestellungen unterschiedlicher Forschungsbereiche hervor, welche mit Hilfe des Korpus behandelt werden können. Hier wird insbesondere der Einfluss sozialer Faktoren thematisiert: Literaturüberblicke aus den Bereichen der medizinischen Kommunikation und der Gendermedizin belegen unterschiedliche Formen von Biases, welche auch auf sprachlicher, diskursiver Ebene (re)produziert werden. Um solche Biases ausfindig zu machen, werden Ansätze aus der Korpuslinguistik und den Corpus-Assisted Discourse Studies (CADS) mit gendermedizinischen Fragestellungen verknüpft.

Mit Fragen der Korpuserstellung und insbesondere der Korpusannotation beschäftigt sich Andrea Lackner, die in ihrem Beitrag über die Entstehung des Österreichischen Gebärdensprachkorpus berichtet. Dieses sich im Aufbau befindende Korpus umfasst Aufnahmen von 50 gehörlosen Muttersprachler*innen der Österreichischen Gebärdensprache (Stand 05/2019) in einem Zeitumfang von ca. 50 Stunden. Zusätzlich werden laufend Daten im Umfang von 30 Stunden für fünf Varietäten der Österreichischen Gebärdensprache hinzugefügt. Der Beitrag schildert ausführlich, welche soziokulturellen und sprachstrukturellen Faktoren für die Datensammlung und Datenaufbereitung zentral sind. Besonders im Fokus steht die Anreicherung des Korpus mit linguistischen Informationen: Mit dem multimodalen Annotationsprogramm ELAN werden sowohl die einzelnen Gebärden (als Tokens in Form von ID-Glossen) als auch satzähnliche Einheiten und nicht-manuelle

Elemente von gehörlosen Muttersprachler*innen annotiert. Berücksichtigt wird dabei insbesondere, dass gebärdensprachliche Elemente nicht nur sequenziell, sondern auch simultan und in Bezug auf den dreidimensionalen Raum realisiert werden. Der Beitrag stellt ferner erste Ergebnisse vor, welche sich einerseits auf die funktionale Verwendung des Gebärdenraums, andererseits auf Interrater-Reliabilität und Variation von gebärdensprachlichen Elementen beziehen.

Bibliographie

Bubenhofer, N./Dreesen, P. (2018): Linguistik als antifragile Disziplin? Optionen in der digitalen Transformation. *Digital Classics Online*, 4(1), 63–75.

Bubenhofer, N. (2018): Wenn „Linguistik" in „Korpuslinguistik" bedeutungslos wird. Vier Thesen zur Zukunft der Korpuslinguistik. In: J. Gessinger/A. Redder/U. Schmitz (Hg.): *Korpuslinguistik*. Duisburg: Universitätsverlag Rhein-Ruhr, 17–30.

Flanders, J./Jannidis, F. (Hg.) (2019): *The shape of data in the digital humanities. Modeling texts and text-based resources*. London, New York: Routledge.

Gregory, I. N./Hardie, A. (2011): Visual GISting: bringing together corpus linguistics and Geographical Information Systems. *Literary and Linguistic Computing* 26(3), 297–314.

Jensen, K. (2014): Linguistics in the digital humanities: (computational) corpus linguistics. *MedieKultur: Journal of Media and Communication Research* 30(57), 115–134.

Klein, L./Gold, M. (2016): Digital Humanities: The Expanded Field. In: M. Gold/L. Klein (Hg.): *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press.

Lemnitzer, L./Zinsmeister, H. (2015): *Korpuslinguistik. Eine Einführung.* 3., überarbeitete und erweiterte Auflage. Tübingen: Narr Francke Attempto.

Losh, E./Wernimont, J./Wexler, L./Wu, H. (2016): Putting the Human Back into the Digital Humanities: Feminism, Generosity, and Mess. In: M. Gold/L. Klein (Hg.): *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press, 92–103.

McEnery, T./Hardie, A. (2012): *Corpus linguistics. Method, theory and practice*. Cambridge: Cambridge University Press.

Nartey, M./Mwinlaaru, I. (2019): Towards a decade of synergising corpus linguistics and critical discourse analysis: a meta-analysis. *Corpora* 14(2), 203–235.

Perkuhn, R./Keibel, H./Kupietz, M. (2012): *Korpuslinguistik*. Paderborn: Fink.

Scherer, C. (2014): *Korpuslinguistik*. 2., aktualisierte Auflage. Heidelberg: Winter.

Iris Jammernegg, Sonja Kuri, Federico Collaoni

Diskursive Implementation der Energiewende in Deutschland: Eine linguistische Analyse zu Diskursorganisation und Wissensaufbau anhand von drei Akteuren des deutschen Energiesektors

1 Einführung[1]

Dieser Beitrag stellt erste Teilergebnisse eines derzeit im Fachbereich Germanistik an der Universität Udine laufenden Forschungsprojekts vor. Untersucht wird, sowohl getrennt als auch vergleichend, ein Ausschnitt des auf die erneuerbaren Energien bezogenen Diskursgeschehens in Deutschland und Italien. Da insbesondere jene in sich verwobenen Diskursstränge von Interesse sind, die Rezeption und Akzeptanz der Inhalte zu steuern versuchen, werden als Korpusbasis schriftliche, multimodale, größtmögliche Breitenwirkung erzielende und/oder an diskursrelevante Zielgruppen gerichtete Online-Kommunikate von Akteuren, die dem Entscheidungszentrum des öffentlichen, profitorientierten und Dritten Sektors[2] zuzuordnen sind, sowie von Multiplikatoren herangezogen.

Dabei kommt einerseits der Analyse von Wissensaufbau bzw. -management innerhalb eines äußerst komplexen, zahlreichen Kontextvariablen unterliegenden Handlungsfeldes, andererseits der diskursorganisierenden und Wissensbestände formenden Rolle des bisher noch wenig erforschten Sach- und Fachwortschatzes zentrale Bedeutung zu. Der multidimensionale, Synergieeffekte aus der konsequenten Verbindung von Fachsprachen-, Text-, Diskurs- und Framelinguistik nützende Analyseansatz ist pragmatisch und handlungsorientiert ausgerichtet.

1 Die Einführung und Abschnitt 2 verfasste Iris Jammernegg, Abschnitt 3 Federico Collaoni und Abschnitt 4 Sonja Kuri, die Konzeption des Artikels sowie Zusammenfassung und Bibliographie werden gemeinsam verantwortet.

2 Darunter ist der Non-Profit-Bereich zu verstehen, also der weder gewinnorientierte noch staatliche Teil einer Gesellschaft bzw. Volkswirtschaft.

Nachstehend wird das Forschungsdesign präsentiert, in weiterer Folge werden erste Ausblicke auf Diskurs und Wissen organisierende Elemente des relevanten Sach- und Fachwortschatzes im Rahmen der zyklisch ineinandergreifenden quantitativen und qualitativen Untersuchung anhand von RWE, des Ministeriums für Energiewende Landwirtschaft Umwelt Natur und Digitalisierung (Schleswig-Holstein) und der Bürgerwerke gegeben. Die die einzelnen Abschnitte prägenden konzeptuell-methodischen Schwerpunkte werden durch entsprechende Korpusanalysen veranschaulicht.

1.1 Korpusgestaltung und Datenerhebung

Sowohl aus den Forschungsfragen resultierende Kriterien als auch die von der linguistischen Diskursanalyse gestellte Forderung nach Repräsentativität (vgl. u.a. Busch 2007) leiteten die Zusammenstellung des Korpus. Angesichts der eingangs resümierten Zielsetzung wurden daher jeweils 14 Akteure für den deutschen bzw. italienischen Sprachraum ausgewählt, deren Webauftritte das Themenfeld in unterschiedlichen Texten, Textsorten und Kommunikationssituationen sowie aus diversen Perspektiven behandeln (vgl. Kämper 2015: 30). Für den öffentlichen Sektor wurden die Ministerien für Wirtschaft und Umwelt, für die Implementation der erneuerbaren Energien auf nationaler oder regionaler Ebene tätige Agenturen sowie gemeinnützige Akteure für Wissenstransfer und Beratung der drei Gesellschaftssektoren bei Energiebelangen selektiert. Den Wirtschaftssektor vertreten jeweils die größten Energieunternehmen, Zulieferunternehmen für die sich wandelnde Energiebranche und Kommunikationsagenturen als Multiplikatoren. Im Non-Profit-Bereich finden sich die Dachverbände der Energiewirtschaft, Dachverbände für Natur- und Umweltschutz, auf nationaler Ebene tätige sowie lokale Energiegenossenschaften. Eine eigene, keinem Gesellschaftssektor vorrangig zuzuordnende Kategorie der Multiplikatoren bildet die Fachpresse für erneuerbare Energien. Das primäre Einreihungsprinzip für die ersten beiden Sektoren war die Funktion des Akteurs, während für den Dritten Sektor hauptsächlich die Organisationsform und die daraus ableitbaren Merkmale wie etwa Nichtstaatlichkeit oder Gemeinnützigkeit ausschlaggebend waren.

Die Datenauswahl innerhalb der somit zur Verfügung stehenden Textmenge erfolgte dann exemplarisch und fokussierte Schlüsseltexte (die konzeptuell-linguistisch Textserien antizipieren oder darauf verweisen) sowie ein Textnetz konstituierende intertextuelle Beziehungen des Einzeltextes[3] (vgl. Fix 2016: 211–213). Als Abgrenzungsinstrument diente dazu die semiotische bzw. medienspezifische Salienzsetzung durch den Sender. Die Daten wurden für die Pilotstudien zwischen August und Oktober 2018, für das Gesamtkorpus vorerst zwischen November 2018 und Mai 2019 erhoben.

Die diskurs- und textlinguistische Analyse erfolgt zuerst quantitativ mithilfe der freien Software AntConc[4] (s. Abschnitt 3) und wird dann anhand der so ermittelten Schwerpunkte qualitativ (s. Abschnitt 1.2 und 4) vertieft. Dabei zu Tage tretende Salienzen werden dem zyklischen Ansatz zufolge weiteren quantitativ-qualitativen Durchgängen unterzogen.

An dieser Stelle sei vermerkt, dass wir uns in Anlehnung an Klug/Stöckl (2016) und Meier (2010) an einem umfassenden Linguistikverständnis orientieren, das notwendigerweise eine Synthese von Text-, Bild- und Audiolinguistik sucht. Demzufolge erfassen wir Sprachtexte, Bildtexte, Sprach-Bild-Texte des schriftlichen und/oder mündlichen Mediums.[5]

1.2 Theoretisch-methodischer Ansatz

Wertvolle Analyseinstrumente bietet die Frame-Semantik, da einerseits sowohl kognitive als auch sprachliche bzw. semiotische Organisationsformen des verstehensrelevanten Wissens sowie ihre diskurssteuernde Handhabung durch die Akteure erfasst werden sollen, sie andererseits gegenüber unterschiedlichen Ansätzen bzw. Techniken anschlussfähig ist (s. Abschnitt 3 und 4). Mit Busse (2012) verstehen wir dabei Frames als jeweilige Wissensrahmen,

3 Dabei berücksichtigten wir salient gesetzte Verweise auf weitere thematisch verbundene Texte sowohl innerhalb der Akteurssite als auch auf externen Seiten.

4 Verfügbar unter https://www.laurenceanthony.net/software/antconc/ [05/05/2019]. Beispielhaft für bisherige Studien unter Einsatz dieses Analysetools nennen wir hier Ziem (2013) und Jakob (2017).

5 Insgesamt handelt es sich um rund 1.000 Dokumente bzw. 685.000 tokens, die sowohl als Word-Dateien als auch im .txt-Format abgespeichert infolge der Genehmigungsabkommen mit den einzelnen Akteuren in diesen Formaten nur dem Forscherteam zugänglich sind.

die bei der aus Senderperspektive erwünschten bzw. von RezipientInnen[6] effektiv gewählten Lesart aktiviert werden und das Verstehen bzw. Erinnern lenken, indem sie den menschlichen Kognitionsmechanismen entsprechend Neues musterhaft über Analogien und Ähnlichkeiten in schon bekanntes Wissen einbetten und als Gedächtnisleistung abrufbar machen. Sie gehören zum semantischen Wissen um das zu verstehende Element (vgl. Busse 2012: 306, 526). Bereits einzelne Wörter – bzw. Zeichen – können diese Schemata im Gedächtnis aktivieren (vgl. ebd.: 332–333).

Die Analyse soll dabei folgende Aspekte erhellen: Welche Wissensbestände bzw. -domänen sind allgemein relevant, um das Thema und die darum kreisenden Ausführungen zu verstehen? In diesem Zusammenhang ist dann zu untersuchen, welche dieser Bestände die einzelnen Sender aktivieren und ob sie dies explizit oder implizit tun. In weiterer Folge sind Strukturen zu ermitteln, die die Rezeption steuern, seien es Verstehenshilfen oder Sendereinstellungen.

Die diversen Frames, ihre Komponenten und Anschlussstellen müssen auf unterschiedlichen Ebenen erfasst werden. So ist für die Bedeutung eines Wortes oder eines anderen semiotischen Zeichens zuallererst die Kontext- bzw. Kontextualisierungsebene zu berücksichtigen. Dazu gehören von angesprochenen Personen oder von Forschern aufgrund ihrer anzunehmenden Vorkenntnisse, Interessen und Erwartungen abgerufene Wissensbestände bzw. jene, die jeweilige Sender durch explizite oder implizite Bezugnahme aktivieren. Da sich die erste Untersuchungsphase auf ein quantitatives korpuslinguistisches Instrument stützt, erscheint es operativ angebracht, *bottom up* anzusetzen und von den eigenen Erwartungen und Intentionen als mögliche Rezipienten und involvierte Forscher ausgehend diese als a priori gesetzte Frames aufzulisten und anschließend im Korpus zu verifizieren. Dazu zählen z.B. Klimawandel, Effizienz, Nachhaltigkeit, Rentabilität, Versorgungssicherheit, nationale und europäische Gesetzeslage. Auf diese Schlüsselwörter, die in Abschnitt 3.2 aufgelistet sind, stützen sich die Analysevorgänge mit AntConc in den Abschnitten 3.3 und 4.2, deren Ergebnisse dank ihrer satz-

6 Im weiteren Verlauf des Artikels verzichten wir auf geschlechterspezifische Formulierungen, einerseits aus konzeptuellen Gründen, wenn Institutionen und Rollen angesprochen werden, andererseits auch aus stilistischen. Selbstverständlich sind in beiden Fällen stets physische Vertreter beider Geschlechter zu verstehen.

bzw. textsemantischen Einbettung (Frequenzen, Konkordanzen, Kollokationen) weitere Aufschlüsse über die Aktivierung bestimmter Wissensbestände geben (vgl. Fraas 2001), wobei Kookkurrenzen stets sowohl ein Mittel der Spezifizierung als auch der Vernetzung von Frames sind (Busse 2012: 508).

Danach ist auf Textebene das Wissen vermittelnde Potenzial der Makrostruktur, der intertextuellen Bezüge sowie von Leitframes zu sondieren. Hier verzahnen sich quantitative und qualitative Methoden, denn im ersten Arbeitsschritt anhand der Analyse mit AntConc bereits ermittelte konzeptträchtige Nomina sowie Prädikationen können mit der auf Layout-Gestaltung beruhenden und navigationsleitenden Salienzsetzung durch den Sender in Verbindung gebracht werden. Diese Stellen sind auf weitere mögliche, noch nicht maschinell erfasste Frame-Elemente hin zu überprüfen, deren weiterführende Vernetzungen dann wiederum über AntConc zu bestimmen sind. Dabei treten eben auch prädikative Verweise auf weiter entfernt liegende Teile desselben Textes bzw. auf fremde Texte in Erscheinung (vgl. ebd.: 512). In dieser Phase lassen sich auch Hypothesen über mögliche Leitframes aufstellen, die die Grundstruktur der Informationsaufbereitung durch den Sender bzw. der erwünschten Informationsverarbeitung durch die Rezipienten bilden und in Abschnitt 2 näher ausgeführt werden.

Auf der Mikroebene werden dann – wiederum im zyklischen Verfahren – einzelne Frameelemente im Satzbereich bzw. in anderen semiotischen Konstrukten sowie die Knotenpunkte zwischen dem sprachlichen und dem visuellen Code untersucht.

Als verstehensrelevant werden auf allen Ebenen die Sendereinstellungen zur Wort- bzw. Strukturbedeutung hinzugerechnet. Diese Konnotationen können z.B. in Form von wertenden Adjektiven, der Auswahl einiger unter allen möglichen Anschlussstellen, einer besonderen Farbgebung oder dem Lichteinfall auf Abbildungen zum Tragen kommen. Transversal werden dazu grammatisch-semantische Frames zu Modalität, konzessiven oder adversativen Beziehungen, aber auch zu Narration und Argumentation erforscht.

Da Frames rekursiv sind und folglich jede ihrer Komponenten ebenfalls einen Frame darstellt, der sich über die sogenannten Anschlussstellen (Slots) in kleinere Einheiten unterteilen oder sich mit anderen bzw. übergeordneten Frames verknüpfen kann, ist eine erste Menge an Slots mithilfe von Fragen

und/oder Prädikationen sowie der diversen quantitativen Recherchen zu definieren. Slots legen die Kriterien fest, denen zufolge konkrete Füllungen (Werte) geeignet sind, den Frame „zu einem epistemisch voll spezifizierten („instantiierten“) Wissensgefüge“ (ebd.: 564) zu machen. Für jeden Akteur, aber auch für die kulturspezifischen Subkorpora[7] sind als besonders forschungsrelevant Standard-Ausfüllungen, strukturelle Invarianten sowie bei der zulässigen Ausfüllung der Slots jeweils geltende Constraints zu erheben. Während erstere aus dem konventionalisierten thematischen Wissen ergänzt werden, wenn keine spezifischen Werte zum Einsatz kommen, zeigt die zweite Kategorie eine mehr oder weniger stabile Beziehung zwischen den Slots eines Frames an, wie etwa zwischen *Angebot* und *Nachfrage* bei dem Konzept *Marktmechanismus*. Constraints stellen Abhängigkeitsverhältnisse dar, die zwischen Slots oder zwischen Werten diverser Slots bestehen (ebd.: 419). Mit diesem Fokus werden im zyklischen Verfahren explizite und implizite Füllungen – im Sinne der versteckten Prädikationen von von Polenz (1985) – erfasst. Zudem wird jeweils die Perspektive geortet, die bei der Aktivierung eines Frames bestimmte Aspekte markiert (vgl. ebd.: 334). Es wird geklärt, ob sie bereits breit konventionalisiert ist oder ob es sich um eine einzelorganisationale Sichtweise handelt, die „Wissens-Facetten“ (ebd.: 517) individuell kombiniert und auf diese Weise dazu beitragen kann, dass sich bestimmte Slots und Werte von der Peripherie zum Zentrum eines Konzepts bewegen oder umgekehrt. Diese Fokussierung ist wiederum an charakteristischen Kollokationen und im weiteren Sinn an allen semiotischen Lexem-Kookkurrenzen ablesbar.

2 Multimodale Leitframes am Beispiel des Web-Auftritts von RWE

Angesichts der strukturellen Beschaffenheit von Hypertext-Knotenpunkten, die im Webseiten-Layout durch Navigationsleisten, Linksammlungen, Verlinkungen im Fließtext sowie als thematische Blöcke gekennzeichnet sind, gehen wir davon aus, dass an jenen Stellen großteils – vor allem übergeordne-

7 Dazu zählen das deutsche und das italienische Korpus, aber auch die jeweils einen Gesellschaftssektor eines der beiden Kulturräume umfassenden Korpora.

te – Frame-Elemente wie eigenständige Frames und Slots bezeichnet werden. Wenn wir als Beispiel die auf der RWE-Homepage angebotene Unterseite zum seit längerem umstrittenen Hambacher Tagebau herausgreifen,[8] finden wir im oberen Navigationsmenü unter anderen auch den Begriff *Renaturierung*. Je nachdem, welche Vorkenntnisse und Erwartungen bzw. welche Informationsquellen Rezipienten mit diesem Konzept verbinden, werden sie eine stärker im Allgemein- oder im Fachwissen fußende, neutral oder kritisch orientierte Lesart aktivieren, die wir anhand der folgenden Definitionen charakterisieren können. So rückt die von Duden online gebotene Erklärung „(eine kultivierte, genutzte Bodenfläche o. Ä.) wieder in einen naturnahen Zustand zurückführen“[9] die Konzepte *freie, unberührte Natur* und *Umkehrungsprozess* in den Vordergrund. Der im Wikipedia-Eintrag angesprochene Wissensbestand gestaltet sich komplexer: Nach der eingehenden Definition als „Wiederherstellung von naturnahen Lebensräumen“ aus „landwirtschaftlichen Flächen, Meliorationsgebieten, aufgelassenen Industrie- und Verkehrsanlagen oder Bergbaufolgelandschaften“ moniert der Artikel, dass Renaturierungsmaßnahmen die erfolgte Schädigung nie vollständig ausgleichen, dass sie folglich den vorbeugenden Schutz der Ökosysteme nicht ersetzen können. Der Eintrag im Gabler Wirtschaftslexikon weist den Begriff als Teil eines innerhalb des Wirtschafts-Fachwissens weiter spezialisierten Bereichs aus und fokussiert die ökologisch vertretbare, positiv bewertete Wirtschaftlichkeit des Prozesses. Auf Rezipientenseite zu erwartende Anschlussstellen sind also „Umkehrungsprozess“, „Natur als (Nah-)Erholungsgebiet“, die Beschreibung der Anwendungsgebiete und Maßnahmen sowie die relativierende Wertung, die der Lesereinstellung entsprechend neutral, positiv oder negativ besetzt sein wird.

Die über den Link *Renaturierung* zu öffnende RWE-Unterseite weist überraschenderweise nur 3 Okkurrenzen des Begriffs auf, wobei die erste, in

8 Die die Diskussion betreffenden Quellen (letzter Zugriff: 24/04/2019) werden aus Platzgründen hier zusammengefasst: https://www.group.rwe/, https://www.hambacherforst.com/, https://www.duden.de/rechtschreibung/renaturieren, https://de.wikipedia.org/wiki/Renaturierung#Anwendungsbereiche, https://wirtschaftslexikon.gabler.de/definition/postwachstumsoekonomie-53487/version-276574 (Paech, Niko (2013), 6.3. e), https://www.hambacherforst.com/renaturierung/. Zur Positionierung von RWE im Energie(wende)diskurs verweisen wir hier auf die Untersuchung von Wieder/Rosenberger (2017).

9 Die Quellenangaben für den gesamten Absatz sind in Fußnote 8 angeführt.

der Gesamtüberschrift „Tagebau Hambach: Ausgleich schaffen durch Renaturierung“ enthaltene bereits dank der Nähe- bzw. Hierarchiebeziehung mit dem unmittelbar im Untertitel folgenden verwandten Konzept der *Rekultivierung* verknüpft wird: „Die Rekultivierungsmaßnahmen von RWE im Überblick“. Diese konzeptuelle Verschränkung wird gleich zu Beginn des Inhalt des Textes sowie Senderposition resümierenden Vorspanns durch wertende Füllung – „aufwendige“ – und Slot-Erweiterung – „im Bereich der Renaturierung und Rekultivierung“ – explizit gemacht: „Der Tagebau Hambach greift in die Landschaft ein – das ist unbestritten. Doch durch aufwendige Maßnahmen im Bereich der Renaturierung und Rekultivierung schafft RWE Power neue Wälder, Felder, Seen und Biotope.“ Am Ende der Seite wird sie im abschließenden Abschnitt „Ein kontinuierlicher Lernprozess“ nochmals, und zwar unter umgekehrten, die sich im Verlauf des Textes herauskristallisierte Hierarchie abbildenden Vorzeichen aufgegriffen: „Die Methoden der Rekultivierung und Renaturierung, die RWE anwendet, haben sich im Laufe eines langen Lernprozesses stets weiterentwickelt.“ Sonst setzt der Sender nur um den spezifischer mit dem Bergbau verbundenen Begriff *Rekultivierung*[10] (9 Okkurrenzen) kreisende Lexeme ein, deren Kookkurrenzen Aufschluss geben über die aktivierten Slots und ihre Füllungen. Die als prädikative Sätze bzw. als Fragen formulierbaren Leerstellen werden mit Informationen ergänzt, respektive etwa „fördernde Begleitumstände?“ – „Lernen“ bzw. „kontinuierlicher Lernprozess“, „wissenschaftliches Arbeiten“ und Forschung, unerlässliche „Prämisse“ für den Tagebau; „Mehrwertergebnisse?“ – „Artenschutz“, „Paradies für Pflanzen, Tier und Mensch“, „neue Wälder, Felder, Seen und Biotope“, „Neues Zuhause für tausende Arten“, „Naherholungsgebiet für die Menschen der Region“, „den Menschen vor Ort etwas zurückgeben“, „neue Lebensräume“, „schnelle Wiedernutzbarmachung“. Wertende Einstellungen zeigen sich in der den bisher geleisteten und weiter in die Zukunft reichenden Einsatz betonenden Zeitdeixis („bislang“, „bis heute“, „kontinuierlicher“), den konnotierten Synonymen für „Lebensraum“ wie „Paradies“ und „Zuhause“, dem Hinweis, dass die Rekultivierungsinitiative von RWE „weltweit von Experten

10 Das Duden Universalwörterbuch (2016) erklärt als Basislemma nur den fachsprachlichen Begriff rekultivieren: „[durch Bergbau] unfruchtbar gewordenen Boden wieder urbar machen“.

gelobt“ wird, dem Adjektiv „schnell“. Der in Teilbeständen des konventionalisierten Wissens als unmöglich angesehene Ausgleich wird in diesem Text nicht nur als erzielbar, sondern sogar als übertreffbar präsentiert. So wurde der „Waldbestand […] nachweislich erhöht“. Der „gezielte[r] Gestaltung“ zu verdankende Mehrwert zieht sich konzeptuell durch die gesamte Website und stellt somit einen Leitframe dieses Akteurs dar.

Was den visuellen Anteil der Seite betrifft, nimmt die erste von fünf Fotografien als einleitendes Banner-Bild horizontal die gesamte Breite ein, während die anderen den verbalen Absätzen gegenüberliegen. Inhaltlich korrelieren sie jeweils mit den sprachlich thematisierten Gegenständen – einer Naturlandschaft, einem naturnahen Lebensraum für schutzbedürftige Tiere und Pflanzen, einem landwirtschaftlichen Nutzungsraum. Von der Ausdrucksseite her stützen sie jedoch – auch unabhängig von den Sprach-Texten – einen in vielen Teilen der Homepage wirkenden Leitframe, der die Perspektivierung und Prozesshaftigkeit fokussiert, die aufgrund spezifischer, in anderen Teiltexten gelieferter Füllungen auch dort, wo diese konkreten Werte fehlen, als Standard-Werte inferiert werden. Wenn man nämlich das Interpretationsschema von Kress und van Leeuwen (2006) darüberlegt, deutet sich auf den Abbildungen durch die Kameraeinstellung eine Bewegung auf den rechten oberen oder unteren Bildrand zu an, was dementsprechend eine Entwicklung zum ideellen oder realen Neuen suggeriert. In diesen Positionen sind hier primär Renaturierungselemente zu sehen. Im Schlussbild vereinen sich die Fluchtlinien darstellenden Anbaufurchen eines bestellten Feldes im rechten Teil des fotografierten Horizonts mit dem von rechts einfallenden Sonnenlicht und veranschaulichen dank der beim westlichen Betrachter aktivierten symbolischen Lesart die im beigestellten verbalen Textteil angesprochenen positiven Perspektiven, die RWE den ansässigen Landwirten nachhaltig bietet.

3 Quantitative Untersuchung anhand des Ministeriums für Energiewende, Landwirtschaft, Umwelt, Natur und Digitalisierung (Schleswig-Holstein)[11]

Die Analyse und Interpretation von linguistischen Mustern anhand des illustrierten theoretischen Rahmens (s. Abschnitte 1.2 und 2) basiert auf einem als *corpus-assisted* bezeichneten Ansatz, welcher die Korpusbefragung bzw. -bearbeitung prägt. Die nächstfolgende Erläuterung dieser ersten Untersuchungsphase einschließlich der jeweiligen Methoden und der damit verbundenen Fragestellungen erfolgt daher ausgehend von der Definition des ausgewählten Ansatzes und dessen Anwendung.

3.1 Der corpus-assisted Ansatz

Bezieht sich der Begriff *corpus-assisted* in der traditionellen Korpuslinguistik (Partington/Duguid/Taylor 2013: 11) auf bestimmte Verfahren der Korpusanalyse, ist dieser mit großem Erfolg auch zur Bezeichnung weiterer Studien angewandt worden (*Corpus-Assisted Discourse Studies*, kurz *CADS*), die über die Methoden der Korpuslinguistik hinausgehen, indem sie diese mit denjenigen der *Critical Discourse Analysis* zusammenführen.

Im Rahmen des Dualismus *corpus-based* bzw. *corpus-driven* versteht man unter *corpus-assisted* „a way in which the two procedures may be combined depending on the researcher's aims and goals" (Bevitori 2010: 51):

> Werden Verfahren prototypisch mit dem Etikett *corpus-driven* als induktive und relativ voraussetzungslose Methoden der Korpusbearbeitung beschrieben [...], so sind Verfahren der *corpus-based approaches* durch die Formulierung von Hypothesen charakterisiert, welche die Analytiker vor der Befragung und Bearbeitung des Korpus aufstellen, um sie anschließend am Datenmaterial empirisch zu überprüfen. (Felder 2012: 124).

11 https://www.schleswig-holstein.de/DE/Landesregierung/V/v_node.html [05/05/2019].

Im Hinblick auf das von Partington (2004) geprägte Etikett *Corpus-Assisted Discourse Studies* weist der Begriff darauf hin, dass die quantitativen, statistisch basierten und computergestützten Verfahren der Korpuslinguistik zwecks der Analyse von linguistischen Mustern im Diskurs im Zusammenhang mit weiteren Ansätzen bzw. Techniken angewandt werden. Der Begriff *corpus-assisted*

> was needed not only to describe the kind of study which incorporate quantitative/statistical methods in the study of discourse types but which also emphasised the eclectic nature of the approach. That corpus techniques were only one of sort among others, and that CADS analysists employ as many as required to obtain the most satisfying and complete results, hence „corpus assisted". (Partington/Duguid/Taylor 2013: 10)

Dahingehend lassen sich laut Partington, Duguid und Taylor (2013: 10) CADS als „a subset of corpus linguistics" definieren, nämlich „the set of studies into the form and/or function of language as a *communicative discourse* which incorporate the use of computerised corpora in their analyses".

Unsere Studie verortet sich in diesem theoretisch-methodologischen Kontext: Die sich aus einer computerunterstützten Korpusanalyse ergebenden Muster werden anhand der im Abschnitt 4 illustrierten Methoden und Kriterien weiter untersucht, wodurch ihre Form/en und Funktion/en im Diskurs z.B. unter Berücksichtigung der multimodalen Kommunikation sowie der Textrezeption auch qualitativ interpretiert werden können.

3.2 Methoden der quantitativen Untersuchung

Die Bestimmung von linguistischen Mustern im Korpus, welche zur Steuerung des ausgewählten Diskurses und zum Aufbau des damit verbundenen Wissens dienen bzw. beitragen können, erfolgte in erster Linie durch die Ermittlung der diskursprägenden Begriffe, die folglich im jeweiligen Verwendungskontext in Betracht gezogen wurden. Zu den Methoden der quantitativen Korpusanalyse, die hierfür angewandt wurden, zählen zunächst die Untersuchung

von Schlüsselwörtern als *Keywords in Context*[12] und deren Konkordanzen, und anschließend die Ermittlung der sich daraus ergebenden Wortcluster bzw. Kookkurrenzen und Kollokationen.
Hinsichtlich der ersten Phase der Korpusbefragung, in deren Mittelpunkt Suchbegriffe wie *Energie*, *Effizienz*, *Klima*, *Mobilität* und *Umwelt* sowie entsprechende Zusammensetzungen, Kollokationen und Unterbegriffe stehen, lässt sich an dieser Stelle festhalten, dass sie nach einem *corpus-based* Verfahren durchgeführt wurde, und zwar ausgehend von einer bestimmten Definition von *keyword* (kulturelle Schlüsselwörter). So wurden die Begriffe

- Energie: „Energiewende", „erneuerbare Energien", „Energiequellen", „Wasser*[13]", „Biomasse", „Solar*", „Sonne*", „Wind*", „Geo*", „Atom*", „Kohle", „fossil*"
- Effizienz: „Enernet", „Passivhaus", „Smart-Grid"
- Klima und Umwelt: „Klimawandel", „Klimaschutz", „Umweltschutz", „öko*", „bio*", „grün*"
- Mobilität: „E-*", „Elektro*"

als Ansatzpunkt für die Analyse des deutschen Korpus nicht aufgrund der Frequenz ihres Vorkommens, sondern aufgrund ihrer Relevanz als *kulturel-*

12 Von nun an KWIC. Damit wird ein „einzelner Kotext zu einem Schlüsselwort" bezeichnet (Lemnitzer/Zinsmeister 2015: 196f.) Hinsichtlich des terminologischen Dualismus Kontext / Kotext wird in der Folge der zweite Begriff bevorzugt, da von den Begriffserklärungen nach Gardt (2005) und Lemnitzer/Zinsmeister (2015) ausgegangen wird: Laut diesen versteht man unter Kotext die „sprachliche Umgebung des Wortes im Text" (Gardt 2005: 152), unter Kontext „die Summe der unmittelbaren Rahmenbedingungen einer Sprachhandlung als das Bezugssystem, innerhalb dessen einer Äußerung eine Funktion zukommt" (Lemnitzer/Zinsmeister 2015: 31). Dahingehend stellt Lautenschläger (2018: 87) fest, dass „KWIC (Keyword(s) in Context) [ ] nach obiger Definition eher ‚keyword(s) in cotext' heißen" müssten. An dieser Stelle wird daher der Begriff Kotext im Zusammenhang mit der Untersuchung der Schlüsselwörter in ihren Konkordanzen verwendet, denn dabei „wird der Kotext um ein Schlüsselwort herum zeilenweise in einer Liste dargestellt" (ebd.), während Kontext in Bezug auf „über den Text hinausgehende Zusammenhänge" (ebd.) Anwendung findet.
Die genaue, hier angewandte Definition von Schlüsselwort wird im unmittelbar folgenden Textteil angegeben.

13 Bei Angabe des Sternchens (*) ermittelt die AntConc-Software alle Okkurrenzen des vorher bzw. nachher angegebenen Clusters einschließlich Zusammensetzungen und flektierter Formen.

le Schlüsselwörter (von nun an *Schlüsselwörter*) a priori ausgewählt: „Keywords are words which are claimed to have a special status, either because they express important evaluative social meanings, or because they play a special role in a text or text-type. [...] Sense 1 is explicitly cultural [...] Sense 2 is statistical" (Stubbs 2010: 21–25; zum Begriff *keyword* als Zugang zur Kultur einer Gesellschaft s. Wierzbicka 1997: 15–16; zum Begriff *cultural keyword* in Bezug auf *climate change* s. auch Bevitori 2010: 19).

Die Ermittlung von Schlüsselwörtern zwecks ihrer Untersuchung als KWICs versteht sich daher als *corpus-based*, soweit sie auf „other sources of information outside our corpus" (Partington/Duguid/Taylor 2013: 10) und insbesondere auf das Forschungsvorhaben zurückgreift, das wir als Grundlage für die Korpusbefragung betrachten. Dieses Forschungskonzept lässt sich beispielsweise anhand des im Abschnitt 3.3 diskutierten Suchbegriffes *Energiewende* erläutern, dessen Brisanz im auf erneuerbare Energien bezogenen Diskurs in zahlreichen Studien hervorgehoben wird (vgl. u. a. Buchan 2012 und Hockenos 2013).

Zielt eine erste Häufigkeitsberechnung darauf ab, die Relevanz der ausgewählten Suchbegriffe im Diskurs anhand quantitativer Daten (Frequenz des Vorkommens im Korpus) zu überprüfen, ist die darauffolgende Analyse der Schlüsselwörter in ihren Konkordanzen bzw. in der „Sammlung von Kotexten eines bestimmten Schlüsselworts" (Lemnitzer/Zinsmeister 2015: 196f.) hinsichtlich der weiteren Korpusbearbeitung zentral. Dadurch wird der Fokus der Untersuchung auf Wortcluster, Kookkurrenzen sowie Kollokationen gelegt, die im Zusammenhang mit den ausgewählten Schlüsselwörtern vorkommen, und die sich sowohl quantitativ als auch nach erfolgter qualitativer Interpretation als relevant für die Steuerung des Diskurses bzw. hinsichtlich des Wissensaufbaus ergeben können.

3.3 Anwendung am Beispiel des Schlüsselwortes „Energiewende"

Die Ermittlung von diskursprägenden, mit Schlüsselwörtern assoziierten linguistischen Mustern im Korpus stellt das Forschungsziel der quantitativ orientierten Untersuchung in ihrer zweiten Phase dar, der die im Abschnitt 4 erläuterte qualitative Analyse folgt. Dabei stellen sich die entsprechenden

Forschungsfragen, welche Muster sich durch ihr systematisches Vorkommen im Diskurs über die erneuerbaren Energien auszeichnen, und inwieweit dieses mit einer diskurssteuernden Funktion im Sinne der Akzeptanz bzw. des Wissenstransfers verbunden ist.

Wie schon erwähnt, verschiebt sich daher der Fokus der Untersuchung von den einzelnen Schlüsselwörtern zu Wortclustern – nämlich zu „Ketten von sprachlichen Einheiten" (Kunze/Lemnitzer 2007: 190) – in ihren Kotexten, die mittels der KWIC-Sortierung identifiziert bzw. durch die Collocate-Funktion der Software AntConc berechnet werden können. Zwecks der Bestimmung diskursrelevanter Muster sind allerdings jene Wortcluster in quantitativer Hinsicht von Bedeutung, die durch „das gemeinsame Vorkommen zweier oder mehrerer Wörter in einem Kontext [= Kotext; Anm. d. V.] von fest definierter Größe [Kookkurrenz]" (Lemnitzer/Zinsmeister 2015: 196f) geprägt sind. Insbesondere „sind Kookkurrenzen dort linguistisch interessant, wo das gemeinsame Auftreten der Wörter häufiger zu beobachten ist, als bei einer Zufallsverteilung aller Wörter zu erwarten wäre" (Kunze/Lemnitzer 2007: 391f.).

Stellt das systematische Auftreten eines bestimmten Wortes bzw. Wortclusters im Zusammenhang mit einem Schlüsselwort ein Signal hinsichtlich der Diskursrelevanz des Ausdrucks dar, sind „linguistisch interpretierte Kookkurrenzen" (ebd.: 391f.), nämlich Kollokationen, laut Baker diesbezüglich zentral:

> When a word regularly appears near another word, and the relationship is statistically significant in some way, then such co-occurrences are referred to as collocates and the phenomena of certain words frequently occurring next to or near each other is *collocation* […]. A collocation analysis [...] gives us the most salient and obvious lexical patterns surrounding a subject, from which a number of discourses can be obtained. When two words frequently collocate, there is evidence that the discourses surrounding them are particularly powerful. (Baker 2006: 95, 114).

Ist die statistische Relevanz von Kookkurrenzen mittels der Collocate-Funktion überprüfbar, wird an dieser Stelle eine qualitative Interpretation zur Be-

stimmung der diskursiven Funktion erforderlich, die einer Kookkurrenz bzw. Kollokation im Korpus entspricht (s. Abschnitt 4).

Anhand des Ministeriums Schleswig-Holstein lässt sich das bisher erläuterte Verfahren am Beispiel des im Diskurs dieses Akteurs zentralen Schlüsselwortes „Energiewende" illustrieren, dessen Okkurrenzen im entsprechenden Subkorpus[14] zunächst berechnet wurden (61). Die KWIC-Sortierung des Begriffes mit der Einstellung +/- 5 Wörter hat anschließend einen ersten Überblick der Kotexte geboten, in denen dieser vorkommt: Am häufigsten lässt sich die Kookkurrenz „Energiewende (und) Klimaschutz*" beobachten (13 Okkurrenzen, 11 davon mit „und"), was in den meisten Fällen mit Verweisen auf die Benennungen „Energiewende- und Klimaschutzgesetz" (5), „Energiewende- und Klimaschutzbericht" (2), „Beirat für Energiewende und Klimaschutz" (2) „Energiewende- und Klimaschutzpolitik" (1), „Energiewende und Klimaschutzprogramm" (1) verbunden ist.

Im Hinblick auf die Diskurssteuerung ist der zweithäufigste Wortcluster „die Energiewende ist" (7) von größerem Interesse, und zwar in doppelter Hinsicht. In erster Linie lässt eine solche prädikative Formulierung (s. 1.2) an eine Begriffserklärung und somit an eine Strategie zum Wissenstransfer bzw. -aufbau denken. Zieht man allerdings die jeweiligen Kotexte in Betracht, ergibt sich, dass operationale Definitionen der Energiewende wie „der Umbau des Stromsektors" bzw. „der Ersatz von Atomstrom durch erneuerbare Energien" lediglich einmal vorkommen, und dass diese von der Komparativform „mehr als" eingeleitet werden:

- „Die Energiewende ist mehr als der Umbau des Stromsektors." (1)
- „Die Energiewende ist mehr als der Ersatz von Atomstrom durch erneuerbare Energien." (1)

Anhand einer gezielten Analyse des Wortclusters „die Energiewende ist" im Korpus konnte hingegen ein linguistisches Muster ermittelt werden, das sich vielmehr durch „important evaluative social meanings" auszeichnet, und das zur Steuerung des Diskurses im Sinne der Akzeptanz dient. Dem Muster „Die

14 https://www.schleswig-holstein.de/DE/Landesregierung/V/v_node.html [05/05/2019].

Energiewende ist + mehr als + operationale Definition“ folgt nämlich im Text eine sozialorientierte Begriffserklärung, in deren Mittelpunkt die lexikalische Einheit „Projekt“ steht: „Die Energiewende ist + sozial konnotierte/s Adjektiv bzw. Bestimmungsform + *Projekt“:

- „Die Energiewende ist ein gesamtgesellschaftliches Projekt“ (1)
- „Die Energiewende ist ein Generationsprojekt“ (1).
- „[Die Energiewende ist darüber hinaus ein] demokratisches Projekt“ (1)
- „[Die Energiewende ist das schleswig-holsteinische] Zukunftsprojekt“ (1)

Die weitere Analyse der Kotexte um die im Korpus identifizierten Muster hat darüber hinaus die Ermittlung der folgenden Formulierungen ermöglicht:

- „Die Energiewende als Generationsaufgabe“ (1)
- „Die Energiewende als Chance“ (1)
- „Die Energiewende ist eine große Herausforderung. [...] Wir möchten Sie als Bürgerinnen oder Bürger auf diesem Weg mitnehmen.“ (1)

Dass die angenommene diskurssteuernde Funktion des Wortclusters „die Energiewende ist“ hier eine tatsächliche Entsprechung bzw. einen konkreten Ausdruck im Text findet, ist hinsichtlich der angewandten sowie der anzuwendenden Verfahren von großer Bedeutung. Im Hinblick auf die durchgeführte quantitative Untersuchung von Wortclustern, Kookkurrenzen und Kollokationen „there is evidence that the discourses surrounding them are particularly powerful“ (Baker 2006: 114). Nichtsdestoweniger erweist sich dabei eine Interpretation der ermittelten Sprachmaterialien im diskurslinguistischen Sinn als erforderlich, um konkrete Schlussfolgerungen dementsprechend ziehen zu können. Dahingehend eignet sich der illustrierte CADS-Ansatz für eine Studie besonders gut, in deren Rahmen eine computergestützte sowie auch eine qualitative Analyse der erhobenen Daten vorgesehen ist. Dieser zweiten Untersuchungsphase ist der nächstfolgende Abschnitt gewidmet.

4 Von der quantitativen zur quantitativ informierten qualitativen Analyse: Wie sich die Bürgerwerke[15] als NPO im agonalen Feld der Stromanbieter positionieren

Quantitative Methoden dienen, wie im vorherigen Abschnitt gezeigt wurde, dazu, „Belege für bestimmte vermutete Phänomene zu finden und deren Distribution im Korpus zu erfassen" (Bubenhofer 2013: 134) sowie deren Kookkurrenzen und damit Salienzen zu ermitteln. Als Kernelemente eines Kommunikats aktivieren Schlüsselausdrücke Frames, die ihrerseits die Interpretation dieser Ausdrücke aktivieren. Die semantischen Beziehungen zwischen den jeweiligen Schlüsselausdrücken konfigurieren schließlich die Sinnstrukturen von Diskursen (Fraas/Meier 2013: 139/140). Die Basiskategorien für die qualitative Analyse stellen hier einerseits als der Thematik inhärente „kulturelle Begriffe" (s. 3.2) die Schlüsselwörter dar, andererseits werden im Laufe der Analyse auch induktiv weitere relevante Kategorien aus dem spezifischen Untersuchungsmaterial erschlossen (vgl. Mayring 2010 und Kuckartz 2012) und miteinander in Beziehung gesetzt.

4.1 Das Untersuchungsdesign

Mit welchen Diskursthemen und in welchem Stil sich die Bürgerwerke versuchen im agonalen Feld der Energiebereitsteller zu behaupten, soll anhand der anschließenden *quantitativ informierten qualitativen Diskursanalyse* erhoben werden. Wir schauen also mit einem „Mikroblick" (Bubenhofer 2013: 110) auf den verbalen Text, aber auch notwendigerweise auf die multimodalen Aspekte als integraler und nicht zu trennender Bestandteil der Nachricht (ebd.: 134), um die „sprachliche[n] und visuelle[n] Artefakte in ihrer bedeutungsstiftenden Korrespondenz zu erfassen" (Fraas/Meier 2013: 137). Das Vorgehen ist, wie in der Einleitung bereits erwähnt, immer theorie- und materialgeleitet zyklisch von quantitativ zu qualitativ, Salienzen des Mikroblicks werden wiederum mittels AntConc einer quantitativen Untersuchung unterworfen, um hier das Vorkommen zu überprüfen, was wiederum zu weiteren Salienzen

15 https://buergerwerke.de/ [05/05/2019].

führt. Unter Punkt 4.3 werden die so erhobenen Daten interpretiert und als Dissens-Artikulationen zusammengefasst.

Die vorliegende Analyse basiert auf mehrmaligen Erhebungen im Zeitraum September 2018 und Mai 2019. In diesem Zeitraum sind auf der Internet-Seite des Akteurs *Bürgerwerke* in einigen Bereichen substanzielle Veränderungen feststellbar: So die Erweiterung des Betätigungsfeldes des Akteurs durch das Angebot von Bio-Gas und damit einer neuen Kategorie, stärkere Profilierung in Hinblick auf Umwelt und ethisches Handeln mit der Einführung des Abschnitts „Nachhaltige Empfehlungen", im April 2019 war noch ein „Wirkungsbericht" verfügbar, auch wurden Fotos getauscht; nunmehr gibt es im Vergleich auffallend weniger Fotos mit dem Sujet Windräder. Feststellbar sind auch redaktionelle Bearbeitungen der Texte nicht nur hinsichtlich der geschäftlichen Neuausrichtung, sondern auch stilistisch durch Verminderung der sprachlichen Komplexität durch die Reduzierung komplexer Syntagmen; in Bezug auf die Schlüsselwörter und deren Kookkurrenzen sowie Konzeptualisierungen sind bei manchen keine einschneidenden, bei manchen doch substantielle Veränderungen feststellbar, wie in weiterer Folge gezeigt wird. Insgesamt werden damit Charakteristika sowie Vor- und Nachteile von Analysen netzbasierter Kommunikate deutlich und zeigen die Notwendigkeit mehrmaliger Erhebungen, wie u.a. von Fraas/Meier (2013: 136) festgestellt.

Das der Untersuchung zugrunde gelegte Material konstituiert sich aus der Startseite sowie der jeweiligen 2. Navigationsebene, Material der 3. Ebene wurde nur in Betracht gezogen, wenn dies von der Forscherin als relevant für den Erkenntnisgewinn beurteilt wurde; so wurden davon nur zwei Pressemitteilungen aus dem Untersuchungszeitraum für das Untersuchungskorpus erfasst; textbegleitende Grafiken wurden als Bilder erfasst, darin enthaltene sprachliche Elemente sind daher nicht in die statistischen Zahlen einbezogen. Insgesamt umfasst das Textmaterial 21.194 *tokens*, die Anzahl der *types* erscheint mit 3.327. Da für das vorliegende Erkenntnisziel nicht relevant, wird die händische Nachbearbeitung, also der erste Schritt von der *quantitativen* zur *quantitativ informierten qualitativen* Analyse, um alle Homonyme in ihren unterschiedlichen Bedeutungen zu erfassen und damit die Angabe der type-Anzahl zu korrigieren, nicht unternommen. Bei der Untersuchung der die

KWICs begleitenden und sich als salient erweisenden Attribute wurden die Okkurrenzen-Werte hinsichtlich ihrer konkreten Semantik im Kontext überprüft und gegebenenfalls korrigiert. Im Zuge der zukünftigen Erfassung der sprachlichen Argumentationsstrukturen wird die Differenzierung bspw. für die Lexeme „damit" und „so" als finale Konjunktionen oder Adverbien relevant.

4.2 Lexeme, Schlüsselwörter, Salienzen

Ausgehend von den KWICs „Energiewende" und „erneuerbar* / erneuerbar* Energie*" wird hier untersucht, wie dieser NPO-Akteur die Energiewende konzeptualisiert und sich in diesem agonalen Feld von, nach eigenen Angaben „über 1.100 Energieerzeugern allein in Deutschland", positioniert. Die Aktionsfelder des Akteurs sind die Energiebereitstellung aus Sonne, Wind und Wasser, seit Januar 2019 auch die Gaserzeugung aus organischen Abfällen eines Zuckerrübenverarbeitungsbetriebs, die E-Mobilität sowie der Kundenservice mit besonderem Fokus auf den Wechselprozess vom bisherigen Energiebereitsteller zu Genossenschaftsmitgliedern.

Das Lexem „Energiewende*" kommt im Untersuchungskorpus insgesamt 142-mal vor. Die Energiewende wird definiert als:

- „eine (erfolgreiche) Energiewende ist eine Energiewende von unten" (4)
- „Die Energiewende ist eine große Gemeinschaftsaufgabe" (1)
- „Die Energiewende muss von uns allen gemeinsam umgesetzt werden" (1)
- „Die Energiewende wird aktuell vor allem durch Bürger vorangebracht" (2)

Direkt verbunden ist der Schlüsselbegriff mit Verben, die Dynamik und Aktivität ausdrücken: „voranbringen" (11), „vorantreiben" (4), „umsetzen" (3). Mit „ambitioniert fortsetzen" (1) und „weiter" (2), „weiterhin" (1) werden die aktuellen Bemühungen und Aktivitäten hervorgehoben.

Mit der Energiewende als „Grundsatzentscheidung über die gesellschaftliche, technologische und kulturelle Entwicklung Deutschlands“ ist der „Ausbau der erneuerbaren Energien und einer entsprechenden Infrastruktur“, neben der „Steigerung [der] Energieeffizienz“ sowie der „intelligente Umgang mit Energie“ (BMU 2012: 5), damit untrennbar verbunden. Jeder Akteur des Energiebereitstellungssektors hat sich nunmehr damit auseinanderzusetzen. Der als Fachsyntagma geprägte Begriff „erneuerbare Energie“ hat im Laufe seiner Verbreitungskarriere eine hohe Symbolkraft erhalten, dabei aber seinen Fachinhalt verloren und ist zu einem Fahnenwort geworden. Von den 121 Okkurrenzen für „erneuerbar*“ entfallen nur 33 auf die usuelle Bildung „erneuerbar* Energie*“, auf „erneuerbar* Quellen“ (1); auf verschiedene Schreibformen von Erneuerbare-Energie-Anlagen (22), auf Erneuerbare-Energie-Projekte (4); auf „erneuerbar produzierter Strom“ (1) und auf „erneuerbarer Strom“ (1). Die uneingeschränkte positive Konnotation des Attributs „erneuerbar*“ wird hier besonders auch für die Bildung weiterer, nicht-usueller Syntagmen genutzt, wobei diese auf das Nomen des neuen Syntagmas übergehen soll, wie in den folgenden Beispielen: „erneuerbare Stromherkunft“ (1), „erneuerbare Energieversorgung“ (7); das Syntagma „erneuerbares Ladenetz“ (1) ist nur durch den Kontext in seiner gemeinten Bedeutung erschließbar. Hier zeigt sich deutlich die Verwendung des Lexems „erneuerbar*“ als Fahnenwort, dasselbe gilt für das Synonym „regenerativ*“ in den folgenden 2 Beispielen: „regenerative Energiewende“ (1), „der Weg in das regenerative Zeitalter“ (1).

Unter der den Bildschirm ausfüllenden Fotostrecke mit insgesamt 13 Fotos der Startseite, dem Logo des Akteurs im Zentrum und rechts unten dem Logo von *Next Economy Award* ist zu lesen: „Versorgen Sie sich mit Ökostrom und Ökogas aus Deutschland – von Bürgern für Bürger. Wir machen die Energiewende in Bürgerhand: Erneuerbar – Regional – Unabhängig.“ Diese Salienzsetzung durch den Akteur wird konsequent durchgezogen und durch die quantitativen Werte bestätigt: „erneuerbar*“ (121) gemeinsam mit „regenerativ*“ (8), „regional*“ (43), ergänzt durch „vor Ort“ (26) und „lokal*“ (14), schließlich „unabhängig*“ (33). Als salient erweisen sich auch „dezentral*“ (28), das „regional*“ stützt, und die synonymischen Ausdrücke „gemeinsam*“ (42), „gemeinschaftlich*“ (11), „zusammen“ (13), die das Personalpronomen „wir“ als Inklusiv-Plural zu verstehen erlauben.

Als neues KWIC erweist sich das in vielfältigen Kombinationen und Kompositionen sowie Neologismen auffallende Lexem „Bürger“. Der Name *Bürgerwerke* ist ein Kompositum aus dem Hochwertwort „Bürger“ als Bestimmungswort und „Werke“ als Grundwort als Anspielung auf *Stadtwerke* mit der Assoziation „regionaler Stromanbieter“[16], der Akteur ist ein 2013 gegründeter Dachverband für private Energieerzeuger, gemeinsam mit der Gesellschaftsform „eingetragene Genossenschaft“ stellt der Name bereits eine erste Positionierung in diesem Diskursfeld dar. Ermittelt man mittels AntConc die Häufigkeit, Verteilung und Kookkurrenzen der KWICs, erweist sich das Lexem „Bürger“, gestützt durch Fotos mit fröhlichen Menschen aller Altersgruppen, die entweder mit orangefarbenen Händeattrappen winken und/oder Sonnenpaneele oder Windräder in der Hand halten oder vor diesen in schöner, lebenswerter Landschaft abgebildet sind, als wesentlicher Leitframe für das Online-Kommunikat und repräsentiert gleichzeitig eine erste „Machtgestaltung durch Sprache qua Nomination“ (Warnke 2013: 79 nach Reisigl 2007). Insgesamt kommt das Lexem 804-mal vor; davon besonders auffällig: als Name „Bürgerwerke“ (265), in der Komposition „Bürgerstrom“ (139), „Bürgerenergie*“ (52), „BürgerÖkogas“ (59), in Syntagmen wie „von Bürgern für Bürger“ (11), „von Bürgerinnen und Bürgern“ (3), „aus/in Bürgerhand“ (74), „Teil der Bürgerenergie-Bewegung“ (14), „Bürgerenergieanlagen“ (6) sowie in der Komposition „Energiebürger/Energiebürgern“ (30) als Grundwort.

Diese Wiederholungen der Signalwörter in vielfältigen Syntagmenvariationen mit wiederum variationsreichen Komposita – einerseits usuellen Komposita und andererseits Neologismen aus den Lexemen *Bürger*, *Strom*, *Energie* – erweisen sich als die diskurssteuernden verbalen Mittel, die durch die Korrespondenz mit den visuellen Mitteln wie Fotografien, Grafiken, dem Vi-

16 Diese Information stammt aus dem Antwortschreiben von Christopher Holzem, Energiebotschafter der Bürgerwerke, der sich als sehr interessiert an der Untersuchung gezeigt hat (12/10/2018). An dieser Stelle möchten wir uns beim Ministerium Schleswig-Holstein und den Bürgerwerken für die offizielle Genehmigung bedanken, alle Elemente von ihren Seiten herunterladen, in verschiedenen Formaten abspeichern, sie mittels Software analysieren sowie auch Screenshots der Seiten für Analysen und Publikationen verwenden zu dürfen. Da RWE zum Publikationszeitpunkt noch nicht auf unser Ansuchen geantwortet hatte, wurde für die betreffenden Seiten auf Data Mining verzichtet.

deo sowie der Farbsymbolik von Orange als Symbol für Freude, Leben, Licht, Sonne und damit einer wichtigen Energiequelle des Akteurs bedeutungswirksam sind. Mit diesen Basiskonzepten wird dann jeweils jedes weitere Diskursthema verbunden.

So rückt bspw. mit dem Begriff „Wertschöpfung“ (18) ein aus volkswirtschaftlicher Sicht argumentativ schlagendes Diskursthema ins Zentrum der Aufmerksamkeit, z. B. in „Wertschöpfung für die dezentrale Energiewende“ (2), „Wertschöpfung für die Gemeinschaft und den Ausbau der erneuerbaren Energien“ (1), „Wertschöpfung verbleibt vor Ort“ (2) „Wertschöpfung in den Regionen“ (1), „Wertschöpfung in Ihrer Region“ (2).

Andere Diskursthemen rücken vergleichsweise eher vom Zentrum in die Peripherie. Ein Vergleich mit den Analyseergebnissen vom Oktober 2018 macht offensichtlich, wie sich der Akteur durch die Erweiterung des Geschäftsfeldes neu ausrichtet: Die Attribute „sicher“, „sauber“ und „bezahlbar“ als Aufnahme der Programmatik des BMWi[17] in Bezug auf Stromgewinnung und -versorgung im Rahmen der Energiewende zeigen folgendes Bild: „sauber“ (22) und „sicher“ (22), beide bezogen auf Versorgung, Preis und Korrektheit der Abrechnung, reduzieren sich im Mai 2019 bei „sauber” auf 15 Okkurrenzen; die Kategorie „sicher“ (4) sowie explizit einmal „Versorgungssicherheit“ erscheint nunmehr als eher untergeordnet. Der Aspekt „bezahlbar“ wird im Gegenzug nicht nur über den auf allen Seiten und Unterseiten noch prominenter platzierten Tarifrechner implizit als Aufforderung, sich vom lohnenden Wechsel zu überzeugen, verhandelt, sondern erscheint auch explizit als Diskursthema im neuen Bereich Ökogas. So wird das Thema Finanzierbarkeit der Energiewende propositiv behandelt: „der Beitrag beträgt nur 0,3 Cent pro Kilowattstunde“ (1), „gesicherter Preis“ (1). Implizites Signal: Die Kosten sind nicht so hoch, wie das anderswo kommuniziert wird.

17 https://www.bmwi.de/Redaktion/DE/Dossier/energiewende.html [05/05/2019].

4.3 Artikulationsformen des Dissenses

In aufgeklärten Industrienationen gehören die Energiewende und die damit verbundenen Diskursthemen, wie bspw. das Wissen um Umweltthemen und Klimawandel, zum vergesellschafteten Wissen. Die Akteure mit ihren Interessen bilden ein agonales Feld, in dem versucht wird, die eigenen Interessen und die Gültigkeit von Aussagen mittels sprachlicher und multimodaler Beiträge durchzusetzen.

Thematisch und sprachlich ist der Web-Auftritt der Bürgerwerke Teil des Textnetzes Energiewende und zeigt dementsprechend auch explizite und implizite intertextuelle Bezüge. Der Akteur ist einerseits den legistischen Rahmenbedingungen unterworfen und muss sich mit den von der Gesetzgebung gesetzten Diskursthemen auseinandersetzen, andererseits muss er glaubhaft erklären können, worin seine Spezifizität im Unterschied zu den anderen Energiebereitstellern besteht. Dabei muss notwendigerweise definiert werden, was er unter *Energiewende* versteht, was für ihn *erneuerbare Energien* sind und wie er dementsprechend agiert.

Die Bürgerwerke nehmen die im deutschen Energiewendediskurs diskutierten Themen auf, konzipieren die Energiewende, wie andere Akteure auch, als ambitioniertes Gesellschaftsprojekt bzw. als „eine große Gemeinschaftsaufgabe“ (1 unter „Jobs“), wobei im „Wettkampf um Einfluss, Geltung und Hörbarkeit“ (Warnke 2013: 76) zur Schärfung des eigenen Profils explizite und implizite Artikulationen des Dissenses feststellbar sind. Die folgende Tabelle erfasst zusammenfassend diese Profilierungsstrategien:[18]

18 Entgegen der üblichen Vorgehensweise im Rahmen der qualitativen Inhaltsanalyse, Kategorien zusammenzufassen und zu paraphrasieren, werden hier auch wörtliche Zitate aus dem Korpus wiedergegeben, da diese jeweils Schlüsselwörter oder saliente Lexeme enthalten, die ihrerseits grundlegende Kategorien des Energiewendediskurses repräsentieren.

Bürgerwerke	versus	Andere
„wir Bürger*“ (7) – als Umdrehung des konventionellen Schemas Oben-Unten und der konventionell damit verbundenen Werte; „gemeinsam“ (42), „gemeinschaftlich“ (11), durch Fotos Repräsentation einer schönen Welt, in der Mensch und Energieversorgungstechnologie gemeinsam existieren, wobei sich die Technologie in der Hand der Menschen befindet und damit beherrschbar ist	versus	„Stromkonzerne/Energiekonzerne“ (3), „fossile Energiewirtschaft“ (1), „fossile Unternehmen“ (1); nur kurzzeitig explizit RWE als Referenz im Ticker mittels Spiel mit Intarsia durch Integration des Konkurrenten im Appellativ #StromanbieteRWEchsel[19] (bis 15/10/2018), sonst Erwähnung der Rodungen im Hambacher Forst in einer Pressemitteilung unter „Presse“;
Umweltschutz und Klimaschutz	versus	Umweltzerstörung, besonders durch die Anfangsbilder im Video, die rauchende Schlote in düsterer Stimmung zeigen
Öko-Strom ausschließlich aus Solar-, Wind- und Wasseranlagen von bekannten Energieerzeugern Öko-Gas/Biogas aus Abfallprodukten, aber „weder Energiepflanzen noch Produkte aus der Tierhaltung“ (1)	versus	Strom aus Atom-, Kohle- oder Biogasanlagen „Stromanbieter, die Gülle aus Massentierhaltungen verwenden“ (2)
Transparenz Betreiber sind als Mitbürger namentlich und visuell erkennbar; die bewusste Verwendung der Farbe Orange mit ihrer Farbsymbolik	versus	Graustrom mit einem „grünen“ Etikett „Etikettenschwindel“ durch den Zukauf von Zertifikaten
„dezentral*“ (29), „regional*“ (44), „lokal*“ (14), „vor Ort“ (30) in Deutschland	versus	zentral, international
Re-Investition „Wertschöpfung“ in der Region (30)	versus	Gewinnmaximierung

19 Die Anspielung ist anlassbezogen und verweist auf die Rodungen im Hambacher Forst, wurde aber nicht vertieft. Dieser Ticker wurde nach dem 19/09/2018 (1. Erhebung) angebracht, bei der 2. am 27/09/2018 erfasst, und war mindestens bis 15/10/2018 vorhanden (3. Erhebung). Dieser Aspekt findet sich bei der Abfassung dieses Beitrags nur in einem Pressemitteilungstext unter „Presse“ wieder. Vgl. dazu die Verwendung dieser Strategie durch RWE während der Kampagne *voRWEggehen*, abgedruckt in Janich (2013), Abb. 25, S. 208.

unabhängige Aktivitäten und Projekte der Bürgerwerke hinsichtlich der Energiewende „unbeirrt von politischen Querschüssen“ (1)	versus	Energiewendegesetz (EEG) (1) und die aktuellen „politischen Rahmenbedingungen“ (2)
ein im Laufen befindliches „Gemeinschaftsprojekt“, das die Bürgerwerke jetzt „vorantreiben“/ „voranbringen“ (17)	versus	ein Zukunftsprojekt (u.a. RWE) ein Generationenprojekt (BMWi)

Tabelle 1: Artikulationen des Dissenses zur Stärkung des Profils im Online-Kommunikat der Bürgerwerke

5 Zusammenfassung

Beim vorliegenden Beitrag handelt es sich um einen Ausschnitt eines größeren Forschungsprojektes zum Energiewendediskurs in Deutschland und Italien aus kontrastiver Sicht, mit je 14 Akteuren aus Deutschland und analogen Akteuren aus Italien, das sich zum Ziel setzt, die textsemantische Steuerung von Rezeption und Akzeptanz der Inhalte zu erforschen, wobei besonderes Augenmerk auf den Wissensstrukturen und Diskurs überformenden Sach- und Fachwortschatz gerichtet wird.

Analysiert wurden hier exemplarisch die Webauftritte von drei deutschen Akteuren der Energiebereitstellungswirtschaft als Repräsentanten der drei Gesellschaftssektoren, um einerseits Einblick in den methodischen Ansatz des zyklisch, pragmatisch und handlungsorientiert angelegten Forschungsvorhabens zu geben und andererseits erste Ergebnisse für Deutschland zu präsentieren: Anhand des Webauftritts von RWE wurde unter Einbeziehung der multimodalen Perspektive das *framing* untersucht und am Beispiel „Renaturierung“ gezeigt, wie es RWE gelingt, im betreffenden agonalen Feld eine glaubhafte Gegenposition zu Teilbeständen des konventionalisierten Wissens aufzubauen. Sowohl auf Inhalts- als auch auf Ausdrucksebene werden aufeinander verweisende verbale und visuelle Elemente mit denselben Werten und Konnotationen verknüpft, die der im Diskursgeschehen heftig kritisierten Tagebau-Aktivität des Akteurs einen ökologisch-sozialen Mehrwert im Ver-

gleich zur vorher vorhandenen Landschaft zuschreiben. Dieser Mehrwert wird schließlich auch in anderen Teilen des Web-Auftritts zu einem der dominanten Frames bei der Abgrenzung zu Konkurrenten und Gegnern. Was die Glaubhaftigkeit bzw. Überzeugungskraft angeht, verfügt der Durchschnitts-Rezipient nicht über die erforderlichen Sach- bzw. Fachkenntnisse, um Vollständigkeit und Wahrheit der angeführten Prämissen zu prüfen. Je nach Erfahrungshorizont wird er folglich die Schlüssigkeit argumentativ aufgebauter Teil-Ganzes-Beziehungen oder der Verallgemeinerung bzw. Postulierung spezieller Ursache-Wirkung-Relationen unterschiedlich bewerten. Anhand des Webauftritts des Ministeriums für Energiewende, Landwirtschaft, Umwelt, Natur und Digitalisierung (Schleswig-Holstein) als Beispiel der gesetzgebenden Ebene wurde durch die Ermittlung quantitativ relevanter Wortcluster um das zentrale Schlüsselwort *Energiewende* gezeigt, wie die Begriffserklärung selbst ein Mittel zur Steuerung des Diskurses sein kann, soweit der Akteur eher eine sozial konnotierte Definition zwecks der Akzeptanz der Energiewende als eine rein operationale Begriffserklärung bevorzugt. Hinsichtlich des angewandten theoretischen Rahmens erweist sich das Teilergebnis als besonders relevant und vielversprechend für eine vertiefte Untersuchung von Sprachmustern der sozialorientierten Argumentation im politischen Diskurs. Aus einem *discourse-historical* Blickwinkel lässt sich z.B. die Konnotationsänderung des Wortes *Projekt* – vom technisch-politischen Bereich zur sozialen Domäne – in der Versprachlichung des Energiewende-Prozesses beobachten.

Schließlich wurde mittels quantitativ informierter qualitativer Diskursanalyse eruiert, wie der Anbieter Bürgerwerke als ein Akteur aus dem Non-Profit-Sektor vergesellschaftetes Wissen nutzt, um die eigene Position im agonalen Feld des Energiewendediskurses und konkreten -prozesses zu schärfen, und versucht, diese zur Durchsetzung zu bringen.

Damit fügt sich diese Studie in die *Corpus-Assisted Discourse Studies* ein, verwendet einen *quantitativ informierten qualitativen Analyseansatz* unter Einbeziehung der *multimodalen Stil- und Frameanalyse* und gibt Einblick in die Komplexität solcher Forschungsvorhaben. Diese Analysen von drei Akteuren unterschiedlicher Gesellschaftsbereiche sind auch als Pilotstudie zur Methodenwahl und Auslotung der Adäquatheit und Leistungsfähigkeit der angewandten Ansätze hinsichtlich Fallstricken und Trugschlüssen für das

Gesamtprojekt anzusehen, wie sie bspw. von Reisigl/Wodak (2016: 21ff.) anhand des Klimadiskurses exemplifiziert wurden.

Bibliographie

Links zu den Internet-Seiten der analysierten Akteure und zur freien Software AntConc:

https://buergerwerke.de [05/05/2019]
https://www.group.rwe [24/04/2019]
https://www.hambacherforst.com/ [24/04/2019]
https://www.hambacherforst.com/renaturierung/ [24/04/2019]
https://www.schleswig-holstein.de/DE/Landesregierung/V/v_node.html [05/05/2019]
https://www.laurenceanthony.net/software/antconc/ [05/05/2019]

Referenzliteratur

Baker, P.(2006): *Using Corpora in Discourse Analysis*. London: Continuum.

Bevitori, C. (2010): *Representations of Climate Change. News and opinion discourse in UK and US quality press: a Corpus-Assisted Discourse Study*. Bologna: Bonomia University Press.

BMU (Hg.) (2012): *Energiewende. Zukunft made in Germany*. Hamburg: KNSK GmbH. Verfügbar unter: www.bmu.de/themen/klima-energiewende [05/05/2019].

BMWi (Hg.) (2019): *Unsere Energiewende: sicher, sauber, bezahlbar*. Verfügbar unter: https://www.bmwi.de/Redaktion/DE/Dossier/energiewende.html [05/05/2019].

Bubenhofer, N. (2013): Quantitativ informierte qualitative Diskursanalyse: Korpuslinguistische Zugänge zu Einzeltexten und Serien. In: K. S. Roth/C. Spiegel (Hg.), *Angewandte Diskurslinguistik: Felder, Probleme, Perspektiven*. Berlin: Akademie Verlag, 109–134. https://doi.org/10.1524/9783050061054.109.

Buchan, D. (2012): *The Energiewende – Germany's Gamble*. Oxford: Oxford Institute for Energy Studies.

Busch, A.(2007): Der Diskurs: ein linguistischer Proteus und seine Erfassung – Methodologie und empirische Gütekriterien für die sprachwissenschaftliche

Erfassung von Diskursen und ihrer lexikalischen Inventare. In: I. H. Warnke (Hg.), *Diskurslinguistik nach Foucault. Theorie und Gegenstände*. Berlin/Boston: De Gruyter, 141–164.

Busse, D. (2012): *Frame-Semantik: Ein Kompendium*. Berlin/Boston: De Gruyter.

Dudenredaktion (2016[8]): *Duden – Deutsches Universalwörterbuch: Das umfassende Bedeutungswörterbuch der deutschen Gegenwartssprache*. Berlin: Bibliographisches Institut GmbH.

Dudenredaktion (2019): *Duden – Onlinewörterbuch*. Berlin: Bibliographisches Institut GmbH. Verfügbar unter: https://www.duden.de/rechtschreibung/renaturieren [24/04/2019].

Felder, E. (2012): Pragma-semiotische Textarbeit und der hermeneutische Nutzen von Korpusanalysen für die linguistische Mediendiskursanalyse. In: E. Felder/M. Müller/F. Vogel (Hg.), *Korpuspragmatik. Thematische Korpora auf Basis diskurslinguistischer Analyse*. Berlin/Boston: De Gruyter, 115–174.

Fix, U. (2016): Diskurslinguistik und literarische Texte. *Tekst i dyskurs – text und diskurs* 9, 207–241.

Fraas, C. (2001): Usuelle Wortverbindungen als sprachliche Manifestation von Bedeutungswissen. Theoretische Begründung, methodischer Ansatz und empirische Befunde. In: N. Henrik/R. Drescher (Hg.), *Lexikon und Text*. Vaasa: Vaasan yliopisto, 41–66.

Fraas, C./Meier, S. (2013): Multimodale Stil- und Frameanalyse – Methodentriangulation zur medienadäquaten Untersuchung von Online-Diskursen. In: K. S. Roth/C. Spiegel (Hg.), *Angewandte Diskurslinguistik. Felder, Probleme, Perspektiven*. Berlin: Akademie Verlag, 135–161.

Gardt, A. (2005): Begriffsgeschichte als Praxis kulturwissenschaftlicher Semantik: die Deutschen in Texten aus Barock und Aufklärung. In: D. Busse/T. Niehr/M. Wengeler (Hg.), *Brisante Semantik. Neuere Konzepte und Forschungsergebnisse einer kulturwissenschaftlichen Linguistik.* Tübingen: Niemeyer, 151–168.

Hockenos, P. (2013): *Why California is to blame for the Energiewende*. Washington, D.C.: Heinrich Böll Stiftung.

Jakob, K. (2017): Diskursive Kehrtwenden in der Energiepolitik: Wer dreht hier eigentlich welches Fähnchen wie im Wind? Eine diskurslinguistische Untersuchung. In: N. Rosenberger/U. Kleinberger (Hg.), *Energiediskurs. Perspektiven*

auf Sprache und Kommunikation im Kontext der Energiewende. Bern: Peter Lang, 199–224.

Janich, N. (2013): *Werbesprache. Ein Arbeitsbuch.* 6. durchgesehene und korrigierte Auflage. Tübingen: Narr Francke Attempto.

Kämper, H. (2015): Diskurslexikografie als gesellschaftsbezogene Wortforschung. Vorstellung eines Wörterbuchkonzepts. In: J. Eckhoff/J. Killian (Hg.), *Deutscher Wortschatz – beschreiben, lernen, lehren: Beiträge zur Wortschatzarbeit in Wissenschaft, Sprachunterricht, Gesellschaft*, Frankfurt a.M./Berlin/Bern/Bruxelles/New York/Oxford/Wien: Peter Lang, 21–38.

Klug, N.-M./Stöckl, H. (Hg.) (2016): *Handbuch Sprache im multimodalen Kontext*. Berlin: de Gryuter.

Kress, G./van Leeuwen, T. (2006[2]): *Reading Images: The Grammar of Visual Design*. London: Routledge.

Kuckartz, U. (2012): *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*. Weinheim und Basel: Beltz Juventa.

Kunze, C./Lemnitzer, L. (2007): *Computerlexikographie. Eine Einführung*. Tübingen: Narr [E-Book].

Lautenschläger, S. (2018): *Geschlechtsspezifische Körper- und Rollenbilder. Eine korpuslinguistische Untersuchung*. Berlin/Boston: De Gruyter.

Lemnitzer, L./Zinsmeister, H. (2015[3]): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Mayring, P. (2010): *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. 11., aktualisierte und überarb. Aufl. Weinheim: Beltz.

Meier, S. (2010): Bild und Frame – Eine diskursanalytische Perspektive auf visuelle Kommunikation und deren methodische Operationalisierung. In: A. Duszak/J. House/Ł. Kumięga (Hg.), *Globalization, Discourse, Media: In a Critical Perspective/ Globalisierung, Diskurse, Medien: eine kritische Perspektive*. Warschau: Warsaw University Press, 369-389.

Paech, N. (2013): Postwachstumsökonomie. In: *Gabler Wirtschaftslexikon*, Wiesbaden: Springer Gabler, 6.3. e. Verfügbar unter: https://wirtschaftslexikon.gabler.de/definition/postwachstumsoekonomie-53487/version-276574 [24/04/2019].

Partington, A. (2004): Corpora and Discourse, a most congruous beast. In: A. Partington/J. Morley/L. Haarman (Hg.), *Corpora and Discourse*, Bern/Berlin/Bruxelles/Frankfurt a.M./New York/Oxford/Wien: Peter Lang, 11–20.

Partington, A./Duguid, A./Taylor, C. (2013): *Patterns and Meaning in Discourse. Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam/Philadelphia: John Benjamins.

Polenz, P. von (1985): *Deutsche Satzsemantik. Über die Kunst des Zwischen-den-Zeilen-Lesens*. Berlin/New York: de Gryuter.

Reisigl, M. (2007): *Nationale Rhetorik in Fest- und Gedenkreden. Eine diskursanalytische Studie zum „österreichischen Millenium" in den Jahren 1946 und 1996*. Tübingen: Stauffenburg.

Reisigl, M./Wodak, R. (2016): The Discourse-historical Approach (DHA). In: R. Wodak/M. Meyer (Hg.), *Methods of Critical Discourse Studies*. London: Sage, 23–61.

Stubbs, M. (2010): Three concepts of keywords. In: M. Bondi/M. Scott (Hg.), *Keyness in Texts*. Amsterdam/Philadelphia: John Benjamins, 21–42.

Warnke, I. H. (2013): Diskurslinguistik und die ‚wirklich gesagten Dinge' – Konzepte, Bezüge und Empirie der transtextuellen Sprachanalyse. In: E. Felder (Hg.), *Faktizitätsherstellung in Diskursen: Die Macht des Deklarativen*. Berlin/Boston: De Gruyter, 75–98. https://doi.org/10.1515/9783110289954.75.

Wieder, R. /Rosenberger, N. (2017): Akzeptanz durch Organisationskommunikation – Positionierung des Energieunternehmens RWE im Energiediskurs. In: N. Rosenberger/U. Kleinberger (Hg.), *Energiediskurs. Perspektiven auf Sprache und Kommunikation im Kontext der Energiewende*. Bern: Peter Lang, 99–122.

Wierzbicka, A. (1997): *Understanding Cultures Through Their Key Words: English, Russian, Polish, German, Japanese.* Oxford/New York: Oxford University Press.

Wikimedia Foundation Inc. (Hg.) (2019): *Renaturierung*. Verfügbar unter: https://de.wikipedia.org/wiki/Renaturierung#Anwendungsbereiche [24/04/2019].

Ziem, A. (2013): Krise im politischen Wahlkampf: linguistische Korpusanalysen mit AntConc. In: F. Liedtke (Hg.), *Die da oben: Texte, Medien, Partizipation*. Bremen: Hempen, 69–90.

Rosemarie Lühr

Diskursforschung an alten Sprachen

1 Problemstellung

Die von der Deutschen Forschungsgemeinschaft geförderten Projekte *Die Informationsstruktur in älteren indogermanischen Sprachen* und *Informationsstruktur in komplexen Sätzen – synchron und diachron: Erarbeitung eines diachronen Kontrastkorpus* (Leitung R. Lühr)[1] haben ein geparstes Korpus mit insgesamt ca. 125.970 Wörtern (Latein, Indisch, Griechisch, Avestisch, Hethitisch/Luwisch/Lykisch) erzeugt. Das Suchwerkzeug ist ANNIS, die Datenbank LAUDATIO (**L**ong-term **A**ccess and **U**sage of **D**eeply **A**nno**t**ated **I**nforma**tio**n). ANNIS, das für „ANNotation of Information Structure" steht, sollte ursprünglich den Zugriff auf die Daten des SFB 632 – „Informationsstruktur: Die sprachlichen Mittel zur Strukturierung von Äußerungen, Sätzen und Texten" ermöglichen. Das Werkzeug wurde seitdem von zahlreichen Projekten genutzt, die eine Vielzahl von Phänomenen untersuchen (https://korpling.org/annis3/#_c=SVNBSVNfMS4w).

Das Textkorpus besteht im Einzelnen aus Ausschnitten aus folgenden Texten:

Avestisch 14.510 Wörter	Hom Yašt; Yasna 29; Yasna Haptaŋhāiti; Yasna 44; Zamyād Yašt; Mihr Yašt
Alt-, Mittelindisch 32.330 Wörter	Rigveda; Brihad-Āraṇyaka-Upaniṣad; Aitareya-Brāhmaṇa; Chāndogya-Upaniṣad; Śathapatha-Brāhmaṇa; Taittirīya-Saṃhitā; Pañcatantra; Bhagavadgītā; Hitopadeśa; Tantrākhyāyikā; Kathāsaritsāgara; Mahābhārata; Jātaka; Mahāvaṃsa

1 DFG-Projektnummer 5465771; DFG-Zeichen: LU 341/5-3, 1999-2006; DFG-Projektnummer 109055449; DFG-Zeichen: LU 341/22-1; LU 341/22-2; 2009-2016; DFG-Projektnummer 199843560; DFG-Zeichen: LU 341/27-1; LU 341/27-2; 2011-2017.

Latein 38.700 Wörter	Caesar; Cato; Cicero; Horaz; Tacitus; Vergil; Sallust; Seneca; Livius; Ovid; Petron; Sueton; Ammianus; Boethius
Altgriechisch 27.700 Wörter	Homer; Isokrates; Longos; Plutarch; Thukydides; Antiphon; Gorgias; Aristoteles; Herodot; Hesiod; Platon; Nonnos; Prokop
Hethitisch, Luwisch, Lykisch 12.730 Wörter	Hethitisch: Gerichtsprotokoll CTH 293; Muwatalli CTH 381; Ritual CTH 443; Ritual CTH 447; Telipinu CTH 19; Apologie Hattušili III CTH 81; Illuyanka CTH 321; Parabeln; Ullikummi CTH 345 Luwisch: Briefe (Assur e, Assur f), Gebäude (Hama I, Hama II, Karkamikiš A 11 b, c; Karkamiš A 1 b); Schale (Babylon III); Stelen (Babylon I; Bohca; Karkamiš A 4 b; Karkamiš A 4 d; Kululu 4; Maras 1; Qal'at El Mudiq) Lykisch: Trilinguen von Letoon

Das Besondere an diesem Projekt ist, dass größere Textabschnitte in ihrer Gesamtheit behandelt wurden, d. h. die ausgewählten Abschnitte wurden Satz für Satz analysiert. Dieses Verfahren erlaubt erstens statistische Aussagen. Zweitens können mehrere Texte einer Sprache autoren- und textspezifische Anwendungen und diachrone Entwicklungen aufzeigen.

Texte alter Sprachen, die mit diesem Werkzeug untersucht werden können, sind auf jeden Fall Diskurse, hier Prozesse der Verbindung von sprachlichen Zeichen, „Aussagen mit Dingen“ (Nonhoff 2004: 64). So sind unter den einheitlich getaggten informationsstrukturellen Werten (*tags*) Parameter wie *saliency*, *I[nformation]-particle*, TOPIC, FOCUS, *discourse*, *style* einschlägig:

Attribute	Werte (*tags*)
[TOP]	(type of topic): Con-T (= Continuing Topic), S-T (= Shifting Topic), C-T (= Contrastive Topic), I-T (= Intonational Topic)
[discourse]	narration, explanation, elaboration, direct speech etc.
[style]	*hyperbaton*, *tmesis* etc.

Vgl. die Gesamtheit von Attributen und Inhalten:

Table (1): Annotated IS parameters

	Attribute	Content
1	[text]	Word token
2	[lem]	Lemma
3	[glos]	Glossing
4	[pos]	Part of speech
5	[saliency]	Animacy: human, animate, concrete, abstract etc.
6	[givenness]	Accessibility: given, new, world-knowledge etc.
7	[definiteness]	Definiteness, indefiniteness
8	[context]	Identity, anaphora, deictic reference etc.
9	[frame]	Scheme according to *Frame Theory*
10	[WPosition]	Position for Wackernagel particles, deficient pronouns, auxiliaries
11	[I-particle]	Particle which is relevant for information structure, foregrounding particles, backgrounding particles etc.
12	[shift]	Continue, retain, smooth shift, rough shift
13	[TOP]	Kind of topic: continuing, shifting, contrastive topic
14	[position-T]	Topic position
15	[F-domain]	Focus domain
16	[NFocus]	New-information focus
17	[CFocus]	Contrastive focus
18	[position-F]	Focus position
19	[discourse]	Narration, explanation, elaboration, direct speech etc.
20	[style]	Stylistic devises, e.g., hyperbaton, tmesis
21	[orig]	Original sentence
22	[transl]	German translation
23	[MC/SCclause-st]	Main clause status, subordinated elements
23	[MC/SCgrfunct]	Subject, object, attribute, predicate, adverbials
25	[MC/SCsyl_no]	Syllable number of phrases
26	[MC/SCword-order]	Verb first, verb second, verb end, enclitics etc.

Da man auf diese Weise Nähediskurse Distanzdiskursen gegenüberstellen kann, ist die übergeordnete Frage, ob dieses heute weit verbreitete polare Konzept (Biber/Conrad 2009: 229) sich auch in alten Sprachen in der Wahl sprachlicher Mittel widerspiegelt (Koch/Oesterreicher 2007; Ágel/Hennig 2006; 2006a; Schnelle 2017). Der Ausdruck von Emphase z. B. gilt als Kennzeichen eines Nähediskurses.

Dabei ist die Art von Konfigurationalität in den altindogermanischen Sprachen zu berücksichtigen. Gilt Diskurskonfigurationalität oder Syntaxkonfigurationalität?[2] Diskurskonfigurationalität wird eher der gesprochenen Sprache zugewiesen. Daher könnten direkte Reden Merkmale dieser Art von Konfigurationalität aufweisen. D.h., Sätze wären nach den pragmatischen oder informationsstrukturellen Einheiten *Topik* und *Fokus* organisiert und nicht nach den syntaxkonfigurationalen oder syntaktischen Einheiten *Subjekt* und *Objekt* (Lühr 2015).

Als Erstes ist daher von Interesse, wo sich Topiks im Satz befinden. Am Satzbeginn könnten sie in einer syntaxkonfigurationalen Sprache Ausdrucksmittel für Emphase, also eine Variante des *Contrastive Topic*[3], sein (Zimmermann 2008). Eine zweite, die Wortstellung betreffende emphatische Struktur bildet das Hyperbaton, die Trennung zusammengehöriger Wörter. In der antiken Rhetorik gilt diese Figur zwar als Stilmittel der gehobenen Sprache, also des Distanzdiskurses, das zur Verstärkung kommunikativer Absichten eingesetzt ist.[4] Doch erscheint das Hyperbaton auch in der direkten Rede und so im Nähediskurs. Für die Analyse dieser rhetorischen Figur werden Texte ausgewählt, die Dialoge enthalten und ältere und jüngere Sprachzustände

2 Obwohl das Indogermanische als SOV-Typ bestimmt wird, überwiegt im Altgriechischen der SVO-Typ, und auch im vedischen Sanskrit findet man diese Stellung. Man sieht hier Reste von Diskurskonfigurationalität, da vor allem morphologiereiche Sprachen, wie es die altindogermanischen Sprachen sind, zu Diskurskonfigurationalität neigen (Hale 1983).

3 Vgl. Kiss (1998: 245): Der *Identificational Focus* steht für "a subset of [a] set of contextually or situationally given elements", während der *Emphatic Focus* gilt, wenn kein solches „subset", sondern eine andere Art von Fokus effektiv wird. Vgl. auch Fanselow 1987: 106; Krisch 1998; Lühr 2010; De Kuthy 2002; Fanselow/Féry 2006; Krifka (2007) erklärt Strukturen wie „[Wild HORses]$_F$ wouldn't drag me there." als *Emphatic Focus*. Die Alternativen bilden ein *ordered set*.

4 Das Hyperbaton steht im Dienste der *compositio* und ist eine Stilfigur der gehobenen Prosa, aber vor allem der Poesie (Lausberg 2008: 357–359, 428).

der jeweiligen Sprache repräsentieren: Dialoglieder des *Rigveda* (Altindisch), *Jātakas* (Mittelindisch), Homer, Nonnos (Griechisch), Cicero, Boethius (Latein). Das Hethitische bleibt außer Betracht, weil so gut wie keine Hyperbata vorkommen. Diese Sprache duldet keine Extraktionen aus einer Phrase; das Fehlen des Hyperbatons ist also grammatisch begründet (Kozianka/Zeilfelder 2016; Lühr 2016). Von den verschiedenen Redeformen wird nur die direkte Rede untersucht, weil das Altindische (wie das Hethitische) – anders als das Griechische und Lateinische – allein diese Art von Rede aufweist. Zunächst geht es um die Anzahl der initialen Topiks oder Subjekte. Darauf folgt die Analyse von Hyperbata.

2 Initiale (Subjekte)

Bei den initialen Topics wird zwischen *Continuing*, *Shifting* und *Contrastive Topics* unterschieden. Von diesen nimmt das *Contrastive Topic* eine Sonderstellung ein. Denn nach Büring (1999) wird beim *Contrastive Topic* oder *I[nonational]-Topic* eine Frage nicht komplett beantwortet:

(1) F: Was trugen die Popstars?
A: Die /weiblichen Popmusiker trugen \Kaftane.

Es bleibt offen, was die männlichen Popmusiker trugen. Semantisch bezeichnet das *Contrastive Topic* ein "subset of a given discourse" (Lühr & Zeilfelder 2011: 115). Gegenüber den anderen Topiks ist es im Deutschen durch ein spezifisches Tonmuster gekennzeichnet (Jacobs 1997). Da es sich beim *Contrastive Topic* zudem stets um ein neu etabliertes Topik handelt (Breindl 2008), hat es besonderes Gewicht.

2.1 Altindisch

2.1.1 *Rigveda* (3.063 getaggte Tokens)[5]

Der *Rigveda*, das älteste indische Literaturwerk, vermutlich aus dem 2. Jt. v. Chr., enthält unter anderem Dialoglieder, die sogenannten *Saṃvāda*-Hymnen, die teilweise aus Dialogen oder sogar gänzlich aus Rede und Gegenrede bestehen. Auch gibt es unter den Dialogliedern einige mit erzählenden Passagen (Schnaus 2008: 1, 454).

Im *Rigveda* ergibt die Abfrage für Topiks in initialer Stellung 67 Belege. Davon sind 40 gleichzeitig auch Subjekt des Satzes. Für die Unterscheidung von Nähe- und Distanzdiskursen ist nun von Interesse, wie häufig welche Art von Topik initial erscheint. Auch die Anzahl von initialen Subjekten ist dabei zu berücksichtigen.

Categories		Categories	
Initial Topics	67 (a)	Thereof Subjects	38 (b)
Initial Continuing Topics	23 (c)	Thereof Subjects	13 (d)
Initial Contrastive Topics	20 (e)	Thereof Subjects	11 (f)
Initial Shifting Topics	23 (g)	Thereof Subjects	14 (h)
Initial Intonational Topics	- (i)		

Die Suchanfragen lauten:

(a) TOP!=/#/ _=_ position-T=/init.*/
(b) TOP!=/#/ _=_ position-T=/init.*/ _i_ MEgrfunct=/subj.*/
(c) TOP=/Con-.*/ _=_ position-T=/init.*/
(d) TOP=/Con-*/ _=_ position-T=/init.*/ _i_ MEgrfunct=/subj.*/
(e) TOP=/C-.*/ _=_ position-T=/init.*/

5 Agastya RV 1.170; Agni RV 10.51–52; Flussüberschreitung RV 3.33; Indra und sein Affe RV 10.86; Indras Geburt RV 4.18; Königsweihe RV 4.42; Lopamudra RV 1.179; Marut 1.165; Rätsellied RV 10.28; Sarama RV 10.108; Urvaśi RV 10.95; Yama und Yami RV 10.10; Abendlied RV 2.38; Froschlied RV 7.103; Nachtlied RV 10.127; Soma-Rausch RV 10.119; Waffensegnung RV 6.75; Spieler-Lied RV 10.34.

(f) TOP=/C-.*/ _=_ position-T=/init.*/ _i_ MEgrfunct=/subj.*/
(g) TOP=/S-.*/ _=_ position-T=/init.*/
(h) TOP=/S-.*/ _=_ position-T=/init.*/ _i_ MEgrfunct=/subj.*/
(i) TOP=/I-.*/ _=_ position-T=/init.*/

Direkte Reden sind Merkmale von Nähediskursen. Von den unterschiedlichen Topikarten sind hier vor allem *Contrastive Topics* zu erwarten:

Categories		**Categories**	
Initial Topics	56 (a)	Thereof Subjects	30 (b)
Initial Contrastive Topics	18 (c)	Thereof Subjects	9 (d)

(a) discourse=/.*direct.*/ _i_ TOP!=/#/ _=_ position-T=/init.*/

(b) MEclause-st=/.*/ _i_ discourse=/.*direct.*/ _i_ TOP!=/#/ _=_ position-T=/init.*/ _i_ MEgrfunct=/subj.*/

(c) MEclause-st=/.*/ _i_ discourse=/.*direct.*/ _i_ TOP=/C-.*/ _=_ position-T=/init.*/

(d) MEclause-st=/.*/ _i_ discourse=/.*direct.*/ _i_ TOP=/C-.*/ _=_ position-T=/init.*/ _i_ MEgrfunct=/subj.*/

Von den 20 Belegen für ein initiales *Contrastive Topic* entfallen im Rigveda also 18 auf die direkte Rede. Vgl. folgendes getaggte Beispiel aus dem Dialoglied *Indra und sein Affe*. Der Affe Vṛṣakāpi stellt der Frau des Gottes Indra nach. Nach Schnaus (2008: 343) spricht die Frau des Affen, die Äffin:

(2) RV 10,86,11 Indra und sein Affe

[text]	**indrāṇ́ím**	āsú	nā́riṣu
[glos]	Indrāṇī(F): ACC.SG	dieser: LOC.F.PL	Frau(F): LOC.PL
[saliency]	proper/human		human
[givenness)	giv		access-situational
[definiteness]	def	def-i	def-i[6]
[context]		situational-deictic	
[shift]	rough shift		
[TOP]	**C-T**		
[position-T]	**initial**		

6 Attribut und Nucleus werden mit dem gleichen Index –i bezeichnet.

[text]	subhágām	ahám	aśravam
[glos]	mit gutem Los: ACC.F.SG	ich: NOM.SG	hören: AOR.IND. ACT1SG
[saliency]		pr1	
[givenness]		access-situational	
[definiteness]		def	
[context]		personal-deictic	
[F-domain]	fd		
[NFocus]	nf		
[position-F]	final		
[discourse]	direct speech/turn speaker4/narrative		
[original]	indrāṇ̍īmāsú nā́riṣu subhágāmahámaśravam /		
[transl]	‚Ich habe unter diesen Frauen hier von Indrāṇī gehört als einer, die ein gutes Los hat.‘		
[literal]	‚Von Indrāṇī unter diesen Frauen hier als einer, die ein gutes Los hat, habe ich gehört.‘		

Das *Contrastive Topic*, der Eigenname *indrāṇ̍īm*, ist definit und gegeben. Der partitive Genitiv *āsú nā́riṣu* fungiert als Teilmenge, aus der eine Entität hervorgehoben ist.

2.1.2 *Jātakas* (9.011 getaggte Tokens)[7]

Die mittelindischen *Jātakas*, eigtl. ‚Geburtsgeschichten‘, moralisch lehrreiche Geschichten im Sinne von Erzählungen aus dem Leben des Buddha[8], sind unser größtes Korpus und enthalten zahlreiche direkte Reden, in Hauptsätzen und Nebenstrukturen bis zur 3. Ebene (ME: main-clause level, SE: sub-clause level/sub-clause-like structure).

7 Andabhuta ch. 62; Garahita ch. 219; Kharaputta ch. 386; Kurungamiga ch. 21; Maha-Ummaga ch. 546: Manisukara ch. 285; Nigrodhamiga ch. 12; Sasa ch. 316; Sihacamma ch. 189; Silanisamsa ch. 190; Sujata ch. 269; Sumsumara ch. 208; Supparaka ch. 463; Sussondi ch. 360; Vanarinda ch. 57; Vedabbha ch. 48.

8 Ursprünglich umfasste der Begriff nur Geschichten aus dem Leben des historischen Buddha, Siddhartha Gautama.

direkte Rede
ME 412
SE1 268
SE2 120
SE3 12

Bleibt man bei den Hauptsätzen, so sind die Belegzahlen für Subjekte und Topiks:

ME Initial Topics	298
ME Thereof Subjects	253
ME Initial Subjects	316
ME Initial Shifting Topics	205
ME Initial Continuing Topics	77
ME Initial Contrastive Topics	14

Die direkte Rede ist in den allermeisten Fällen eingebettet und befindet sich dann auf einer SE-Ebene. Für Hauptsätze ergeben sich so wenige Belege:

ME Direct Speech	
ME Initial Topics	16
ME Thereof Subjects	6
ME Initial Contrastive Topics	4

Doch kommen auch hier *Contrastive Topics* vor. Ein Beispiel für ein solches Topik der Ebene SE1 mit einem Vokativ ist (3). Das *Contrastive Topic* erscheint zu Beginn eines *locativus absolutus*:

(3) *Nigrodhamiga-Jātaka* 26

[1]

[text]	**dvīhi**	abhaye	laddhe	**avasesā**
[glos]	zwei: INSTR.M.PL	Immunität(N): LOC.SG	erlangt: LOC.N.SG	übrig: NOM.M.PL
[saliency]	animate	abstract		animate
[givenness]	access-situational	giv		set-relation
[definiteness]	def			spec.indef
[context]		identity.anapher		
[shift]	rough shift			smooth shift
[TOP]	C-T			C-T
[position-T]	initial			middle

[2]

[text]	kim	karissanti	narinda	ti
[glos]	was: ACC.N.SG	machen: FUT. IND.ACT3PL	Männerindra (M): VOC.SG	QUOT
[saliency]	pr3.interrog/abstract		human	
[givenness]	new		access-situational	
[definiteness]	indef		def	
[context]	kata.ref		personal.deictic	
[discouse]	direct speech/turn speaker2/question			
[original]	Dvīhi abhaye laddhe avasesā kiṁ karissanti narindā ti.			
[transl]	‚Nachdem von zweien Immunität erlangt wurde, was werden die übrigen machen, Männerherrscher?‘			
[literal]	… die übrigen, was werden sie machen, Männerherrscher?			

Adjektve wie *avasesā* ‚übrige‘ gehören zu einer *partly ordered set*-Relation. Es liegt ein Bezug auf eine Alternativmenge vor, wodurch beiden Entitäten Nachdruck verliehen wird: zwei – die übrigen (der Menge) (Umbach 2001: 177; 2003).

2.2 Griechisch[9]

2.2.1 Homer (2.748 Tokens)

Im Griechischen bei Homer ist die Verteilung bei den rund 50 initialen Topiks:

Initial Topics	48
Thereof Subjects	23
Initial Subjects	31
Initial Continuing Topics	8
Initial Contrasting Topics	14
Initial Shifting Topics	26

Direct Speech	
Initial Topics	21
Thereof Subjects	10
Initial Contrastive Topics	9

Zu *Contrastive Topics* in direkter Rede (im Folgenden nur noch mit Glossierung und Angabe der relevanten informationstrukturellen Funktionen) vgl.:

(4) Ilias 1,17-19

Atreΐ́dai	te	kaì	álloi	
Atride(M): VOC.PL	etwa	auch	anderer: VOC.M.PL	
eüknḗmides		Achaioí		
wohlbeschient: VOC.M.PL		Achaier(M): VOC.PL		
humῖn	mèn	theoì		doῖen
ihr: DAT.PL	freilich	Gott(M): NOM.PL		gestatten: AOR.OPT.ACT3PL
C-T initial				
Olúmpia	dṓmata		échontes	

9 Ilias 1–187; Odyssee 1–177.

olympisch: ACC.N.PL	Haus(N): ACC.PL	habend: NOM.M.PL	
ekpérsai	Priámoio		
austilgen: AOR.INF.ACT	Priamos(M): GEN.SG		
pólin	eũ	dè	oíkade
Stadt(F): ACC.SG	wohlbehalten	und	nach Hause
hikésthai			
zurückkehren: PRS.INF.ACT			
paîda dè	emoì	lúsaite	
Kind(F): ACC.SG aber	ich: DAT.SG	losgeben: AOR.OPT.ACT2PL	
C-T			
intial			
phílēn			
lieb: ACC.F.SG			

Ἀτρεΐδαι τε καὶ ἄλλοι ἐϋκνήμιδες Ἀχαιοί, / ὑμῖν μὲν θεοὶ δοῖεν Ὀλύμπια δώματ‘ ἔχοντες /
ἐκπέρσαι Πριάμοιο πόλιν, εὖ δ‘ οἴκαδ‘ ἱκέσθαι • παῖδα δ‘ εμοὶ λύσαιτε φίλην •
‚Atriden und auch andere wohlbeschiente Achaier, **euch** freilich mögen die Götter mit ihren olympischen Wohnsitzen gestatten, Priamos’ Stadt auszutilgen und wohlbehalten nach Hause zurückzukehren; (**mein) Kind** aber gebt mir los, das liebe‘

In der Satzreihe folgt auf die zwei initialen *Contrastive Topics*, pronominales *humĩn* ‚euch‘ und nominales *paîda* ‚Kind‘, zwar der Dativ *emoì*, eine pronominale betonte Form. Wegen der parallelen Stellung von *humĩn* und *paîda*, jeweils vor einer adversativen Partikel, dürfte jedoch das zweite *Contrastive Topic* nicht *emoì,* sondern *paîda* sein.

2.2.2 Nonnos (4.543 Tokens)[10]

Das große Vorbild des byzantinischen Dichters *Nonnos* (5. Jh. n. Chr.), des Verfassers der *Dionysiaka* (Διονυσιακά), des letzten großen Epos der Antike, ist Homer. Die *Dionysiaka* beschreiben in 48 Gesängen und annähernd 21.300 Hexametern den Siegeszug des Dionysos nach Indien.

10 *Dionysiaca* Buch 1, 1–443.; Buch 32. 1–299.

Auf der Hauptsatzebene kommen vor:

Initial Topics	33
Thereof Subjects	29
Initial Subjects	35
Initial Contrastive Topics	18
Initial Shifting Topics	11
Initial Continuing Topics	4

Direct Speech	
Initial Topics	8
Thereof Topics	4
Initial Contrastive Topics	4

Vgl. eine Stelle mit den Personalpronomina ‚du' und ‚ich' als *Contrastive Topics*:

(5) Nonnos, *Dionysiaka* 1,443
στήσω δ', ἢν ἐθέλῃς, φιλίην ἔριν:
‚Ich werde mich aber, wenn du es wünschen solltest, einem freundlichen Streit stellen'

allá	**sù**	mélpōn	
aber	du: NOM.SG	singend: NOM.M.SG	
	C-T		
	initial		
pémpe		mélos	donakõdes
schicken: PRS.IMP. ACT2SG		Gesang(N): ACC.SG	vom Rohr erzeugt: ACC.N.SG
egỏ		brontaĩon	arássō
ich: NOM.SG		donnernd: ACC.N.SG	schlagen: PRS.IND. ACT3SG
C-T			
initial			

ἀλλά σὺ μέλπων / πέμπε μέλος δονακῶδες, ἐγὼ βρονταῖον ἀράσσω: /
‚aber du schicke singend einen vom Rohr erzeugten Gesang; ich schlage einen donnernden (Gesang)'

Im Griechischen folgen *Contrastive Topics* oftmals der Konjunktion ἀλλά ‚aber' wie in *allá* ***sù*** ‚aber du'. Voraus geht ein Satz mit kovertem ‚ich'.

2.3 Latein

2.3.1 Cicero (3.272 Tokens)[11]

Ciceros erste Rede gegen Catilina gilt als Meisterstück der antiken Rhetorik und hat maßgeblich zu seinem Ruf als bedeutendster römischer Redner beigetragen. Wie zu erwarten, enthält diese Rede ebenso *Contrastive Topics*.

Initial Topics	25
Thereof Subjects	13
Initial Subjects	29
Initial Contrastive Topics	11
Initial Continuing Topics	11
Initial Shifting Topics	3

Direct Speech	
Initial Topics	25
Thereof Subjects	13
Initial Constrastive Topics	11

Vgl. folgendes Beispiel mit Prolepse des Pronomens der 2. Person:

(6) Cicero, *In Catilinam* 1,9

te	ut	ulla	res
du: ACC	dass	irgendeiner: NOM.F.SG	Sache(F): NOM.SG

11 *In Catilinam (prima oratio).*

C-T
initial
frangat
erweichen: PRS.SUBJ.ACT2SG
te ut ulla res frangat?
‚damit dich irgendeine Sache erweicht?‘

tu	ut	umquam	te	corrigas
du: NOM	dass	jemals	du: ACC	bessern: PRS.SUBJ. ACT2SG

C-T
initial
tu ut umquam te corrigas? …
‚Damit du dich jemals besserst? …‘

Die Spitzenstellung vor einem Finalsatz dient der Emphase.

2.3.2 Boethius (2.943 Tokens)[12]

Die *Consolatio Philosophiae* aus den zwanziger Jahren des 6. Jhd.s ist ein Dialog zwischen dem Autor und der personifizierten Philosophie, die ihn tröstet und belehrt. Der Text besteht hauptsächlich aus direkten Reden.

Initial Topics	25
Thereof Subjects	12
Initial Subjects	29

Direct Speech	
Initial Topics	25
Thereof Subjects	12
Initial Constrastive Topics	2

12 2,3; 2,4; 3,3; 3,4; 3,5; 3,6; 3,7; 3,8.

In dem Dialog zwischen dem Autor und der personifizierten Philosophie, in dem die irdischen Güter auf die Frage hin behandelt werden, ob sie zur Glückseligkeiten verhelfen können, heißt es:

(7) Boethius, Consolatio philosophiae 3,3
Atqui, inquam, libero me fuisse animo, quin aliquid semper angerer, reminisci non queo.
Nonne quia vel aberat, quod abesse non velles, vel aderat, quod adesse noluisses? Ita est, inquam.
Allerdings, sage ich, kann ich mich nicht erinnern, dass ich von freiem Geist war, ohne dass ich mich in irgendeiner Beziehung immer geängstigt hätte. Weil entweder fehlte, was zu fehlen du nicht wolltest, oder da war, was da zu sein du nicht gewollt hättest, nicht wahr?
‚So ist es!' entgegnete ich.

illius	igitur	praesentiam	**huius**
jener: GEN.N.SG	also	Anwesenheit(F): ACC.SG	dieser: GEN.N.SG
C-T		**C-T**	
initial		**initial/post-compound sentence**	
absentiam		desiderabas	
Abwesenheit(F): ACC.SG		wünschen: IPF.IND.ACT2SG	

Illius igitur praesentiam, huius absentiam desiderabas?
‚Von jenem (Ding oder Zustand) also wünschtest du dir die Anwesenheit, von diesem die Abwesenheit?'

Durch die Verwendung der initialen *Contrastive Topics illius* und *huius* signalisiert die Philosophie offensichtlich, dass der Angesprochene nicht weiß, welches Gut er von den vielen Gütern eigentlich will.

Die Auszählung aller Topikarten hat nun ergeben: Initiale Topiks sind oftmals Subjekte. Pragmatik und Syntax stimmen so überein, so dass Diskurskonfigurationalität eher nicht gegeben ist. Für diese Annahme spricht auch, dass die Anzahl der initialen Subjekte ohne gleichzeitige Topikfunktion hoch ist. Das *Contrastive Topic* erscheint aber, auf die Anzahl der Belege berechnet, häufiger in der direkte Rede und ist so tatsächlich Kennzeichen des Nähe-

diskurses. Insbesondere die Gegenüberstellung von Sprecher und Adressat, ‚ich' und ‚du', sind für solche Diskurse typisch. Bemerkenswert ist dabei, dass die ermittelten kontrastiven Strukturen ausnahmslos, wie die Übersetzungen belegen, auch im heutigen Deutsch ohne weiteres möglich sind. Satzinitiale *Contrastive Topicity* scheint so zumindest in den indogermanischen Sprachen universalen Charakter zu haben.

3 Hyperbaton

Beim Hyperbaton kommen nun aber tatsächlich diskurskonfigurationale Merkmale ins Spiel; vgl. folgende Gegenüberstellung:

diskurskonfigurational	**vs.**	**syntaxkonfigurational**
„freie" Wortstellung		„feste" Wortstellung
diskontinuierliche Konstituenten		keine diskontinuierlichen Konstituenten
keine NP-Bewegungsoperationen		NP-Bewegungsoperationen[13] (Luraghi 2013; Viti 2015: 269; Lühr 2019)

Zunächst liegt kein Hyperbaton vor, wenn in den altindogermanischen Sprachen zweigliedrige Strukturen durch Wackernagel-Partikeln getrennt sind. Diese Partikeln sind unbetont; vgl. bei Cicero:

(8) Cicero, Brutus 12

Populus	se	**Romanus**	erexit
Volk	sich	Römisch	erhob sich

‚Das Römische Volk erhob sich'

(8)(b) Cicero, Brutus 10

Marcus	ad	me	**Brutus**	venerat

13 NP-Bewegung ist die Bewegung einer NP in eine Argument-Position (A-Position), d. h. in eine syntaktische Position, in welcher eine Thematische Rolle (Theta-Rolle) zugewiesen werden kann. Die NP-Bewegung hinterlässt eine Spur (t), z.B. [John$_i$ seems t$_i$ have won].

Marcus zu mir Brutus war gekommen
‚Marcus Brutus war zu mir gekommen' (Lühr 2016)[14]

Auch in (9) ist kein Hyperbaton gegeben. Zwar ist Latein eine *pro-drop*-Sprache, d.h. eine Nullsubjektsprache. Das bedeutet, dass dann, wenn Cicero das Subjektpronomen setzt, dieses eine Betonung trägt. Doch erscheint im Lateinischen das Fragepronomen in der Regel zu Beginn des Fragesatzes, so dass für SprecherInnen der damaligen Zeit die Wortfolge nicht ungewöhnlich gewesen sein dürfte:

(9) Cicero, *In Catilinam* 1,6

cui	tu	**adulescentulo**
welcher: DAT.M.SG	du: NOM.SG	Jüngelchen(M): DAT.SG
quem	corruptelarum	illecebris
welcher: ACC.M.SG	Verführung(F): GEN.PL	Verlockung(F): ABL.SG

irretisses
fangen SUBJ.PLQ.ACT2SG

Cui tu adulescentulo, quem corruptelarum illecebris irretisses, non aut ad audaciam ferrum aut ad libidinem facem praetulisti?
‚Welchem Jüngelchen, das du mit der Verlockung der Verführungen gefangen hattest, hast du nicht entweder zur Waghalsigkeit das Schwert oder zur Begierde die Fackel vorangetragen?'

Ist aber ein Adverb, das sich auf ein Verb bezieht, in eine Substantivgruppe eingeschoben, ergibt sich ein Hyperbaton:

(10) Cicero, *Ad familiares* 3,9,3

Tuis	incredibiliter	**studiis**
dein: ABL.N.PL	unglaublicherweise	Studien(N):

14 Auch Fälle von geschlossener Wortstellung mit einer adverbialen Bestimmung, die sich auf das Adjektiv bezieht, gehören nicht hierher: *Tarquinius ex gravi vulnere aeger* ‚der aufgrund einer schweren Wunde kranke Tarquinius' (Menge 2008: 335f.).

ABL.PL
delector
sich erfreuen: PRS.IND.MED1SG
‚Ich erfreue mich unglaublich an deinen Studien‘
wörtl.: ‚An deinen unglaublicherweise Studien erfreue ich mich‘

Aus dem Englischen werden in diesem Zusammenhang ungrammatische Sätze mit *discontinuous syntax* genannt:

(11)(a) *Which has he invited friend to dinner?

Diese verstoßen gegen Ross’ *Left Branch Condition*[15]. Der linke Zweig einer Nominalphrase kann nicht *in situ* verbleiben:

(11)(b) *The RED he bought car last week, the BLUE he has had car for years.
(Devine/Stephens 2000: 4f.)

Vgl. dazu einen griechischen Beleg aus unserem Korpus. Voraus geht:

(12)(a) Ilias 1,503f.Ζεῦ πάτερ εἴ ποτε δή σε μετ᾽ ἀθανάτοισιν ὄνησαἢ ἔπει ἢ ἔργῳ
‚Vater Zeus, wenn je ich dir nützte unter den Göttern, sei es mit Wort oder Werk‘
(12)(b) Ilias 1,504

tóde	moi	
dieser: ACC.N.SG	ich: DAT.SG	
krḗēnon		**eéldo̅r**
erfüllen: AOR.IMP.ACT2SG		Wunsch(N): ACC.SG

τόδε μοι κρήηνον ἐέλδωρ
‚erfülle mir diesen Wunsch‘
wörtl.: ‚diesen mir erfülle Wunsch‘

15 Die Extraktion betreffende Beschränkung besagt, dass keine Nominalphrase auf dem linken Zweig einer anderen Nominalphrase aus dieser Nominalphrase herausbewegt werden kann (Ross 1967).

Ist ‚dieses' als Attribut und nicht als Substantiv gebraucht, wie dem vorausgehenden Kontext zu entnehmen ist, erwarten AdressatInnen ein im Akkusativ Singular Neutrum kongruierendes Substantiv. Ähnlich funktioniert (13). Solche Sätze haben Devine/Stephens zu der Äußerung veranlasst: das Hyperbaton ist "perhaps the most distinctively alien feature of Latin word order". (2006: 524)

(13) Cicero, *In Catilinam* 1,21

quorum		ego	vix	abs	te
welcher: GEN.M.PL		ich: NOM	kaum	von	du: ACC
iam	diu	**manus**		ac	tela
schon	lange	Hand(F): ACC.PL		und	Waffe(N): ACC.PL
contineo					
fernhalten: PRS.IND.ACT1SG					

quorum ego vix abs te iam diu manus ac tela contineo.
‚Ich kann schon lange kaum noch deren Waffen und Hände von dir zurückhalten.'
wörtl.: ‚deren ich kaum von dir schon lange Hände und Waffen zurückhalte'

Ähnlich urteilen antike Rhetoriker. So gefährde das Hyperbaton die *perspicuitas* und kann zur *obscuritas* führen (Lausberg 2008: 358).

Aber auch die Begriffe Grammatikalität und Künstlichkeit spielen beim Hyperbatun eine Rolle: Nach Menge (2000: 580) gilt im Klassischen Latein das Hyperbaton als grammatisch, wenn „syntaktisch zusammengehörige Wörter getrennt werden, ohne den Eindruck der Künstlichkeit hervorzurufen".

Wenn aber ein Redner wie Cicero einen Satz wie (13) äußert, kann man davon ausgehen, dass er für die RömerInnen grammatisch fehlerfrei war. Versetzt man sich daher in AdressatInnen, die als erstes Wort *quorum* hören und dieses als relativen Satzanschluss verstehen, baut sich eine Erwartungshaltung auf. Da ein partitiver Genitiv oftmals naheliegt, wird ein substantivischer Bezugsausdruck evoziert. Er erscheint dann auch nach den sechs relativ unbetonten Wörtern *ego vix abs te iam diu*.

Festzuhalten ist, dass ein Hyperbaton in den alten Sprachen auch dann, wenn es für uns unnatürlich wirkt, grammatisch korrekt ist, solange es nur verstanden wird. So ist das „gekünstelte" Hyperbaton ein Merkmal der home-

rischen Formelsprache. Das ganze Problem erscheint in verschärfter Form bei Nonnos, der ein äußerst idiosynkratisches Griechisch mit zahlreichen Hyperbata schreibt. Auch im Avestischen sind wie im Altindischen raffinierte Hyperbata Bestandteile spezieller Dichtersprachen.[16] Dennoch zeigen Hyperbata je nach Textkohärenz unterschiedliche Grade von Künstlichkeit.

Die eigentliche Leistung des Hyperbaton wird aber darin gesehen, dass es dem einfachen Satz die zyklische Spannung zwischen auflösungsbedürftigen und auflösenden Bezugsgliedern [verleiht und es] so als einer Periode gleichwertig erscheinen [lässt] (Lausberg 2008: 357).

Trifft dies zu, so liegt beim Hyperbaton eine spezielle Art von Verarbeitung vor. Man hat hier einen Fall von inkrementeller Syntax. Eine solche Syntax folgt psycholinguistischen Befunden: In der gesprochenen Sprache basiert sprachliche Produktion und Rezeption auf dem Phänomen der Projektabilität, das an die Zeitlichkeit der Entfaltung von Sprache im Gebrauch gebunden ist. Bei einer inkrementellen, d.h. schrittweise erfolgenden Sprachverarbeitung, haben AdressatInnen gleich nach den ersten Worten im Satz eine Ahnung davon, worum es im Folgenden geht, und zwar unabhängig vom Kontext. Denn die fortlaufenden Projektionen über den Verlauf einer emergenten syntaktischen Struktur erlauben es den HörerInnen, den entstehenden Redebeitrag ohne Verzögerung zu prozessieren (Auer 2007; 2009; 2015; Lühr 2018).[17]

Hier kommt auch die *Scenes*- und *Frames*-Semantik ins Spiel: Eine *Scene* ist ein sich wiederholender, zusammenhängender Ausschnitt der Realität, mit dem Menschen durch eigene Erfahrung oder die anderer vertraut sind, während ein *Frame* ein System sprachlicher Wahlmöglichkeiten darstellt und mit der jeweiligen *Scene* verbunden ist (Fillmore 1977: 63).[18] Dabei genügen, vergleichbar dem Hyperbaton, einige wenige Signale, die die Erwartungshaltung von RezipientInnen steuern (Wengeler/Ziem 2018). Da Hyperbata und *Frames* so Gemeinsamkeiten aufweisen, wird bei den folgenden Typen von Hyperbata auch der jeweilige *Frame* oder die zugehörige *Scene* mit einbezogen.

16 Einem anonymen Gutachter verdanke ich viele wertvolle Hinweise.

17 Vgl. dazu auch Hopper's (1987; 2001; 2011) *Emergent Grammar*.

18 Minsky (1975: 212) versteht unter *Frame* „a data structure for representing a stereotyped situation".

Die Gliederung dieser Typen richtet sich nach dem Gesamtkonzept, das sich aus dem auflösenden Bezugsglied ergibt.

4 Typen von Hyperbata

4.1 Funktionale Begriffe

Funktionale Begriffe kennzeichnen ihren Referenten über eine Relation, in der stets nur ein Referent zu einem gegebenen *Possessor* steht. Prototypische Beispiele sind innerhalb der Verwandtschaftsbegriffe ‚Mutter' und ‚Vater' (Löbner 2005a: 4).

Das folgende indische Beispiel mit einem Demonstrativpronomen stammt aus dem schon angeführten Dialoglied *Indra und sein Affe*. Die Äffin preist Indrāṇī als glücklich. Die Begründung folgt:

(14) RV 10,86,11 Indra und sein Affe

nahí	**asyāḥ**	aparám	caná
denn nicht	dieser: GEN.F.SG	künftig	selbst
jarásā	márate		**pátiḥ**
Altersschwäche(M): INSTR.SG	sterben: AOR.SUBJ.MED3SG		Gatte(M): NOM.SG

nahyàsyā aparáṃ caná jarásā márate pátir

‚Denn selbst künftig stirbt ihr Mann nicht an Altersschwäche.'

wörtl.: ‚denn nicht deren selbst künftig an Altersschwäche wird sterben der Gatte.'

Die Negation am Satzbeginn lässt für auflösungsbedürftiges possessives *asyāḥ* ‚deren' (Indrāṇis) mit dem im Satzzusammenhang genannten Substantiv *jarásā* ‚Alterschwäche' nur einen Bezug auf Indra, Indrāṇis Gatten, erwarten. In der Tat folgt der Funktionalbegriff *pátiḥ* ‚Gatte' am Schluss des Satzes, der seinem Argument einen eindeutigen Referenten zuweist (Löbner 1985; Lühr 1990).[19]

19 Ein funktionales Konzept ist ein relationales Konzept, dessen Referent sich in Abhängigkeit von einem *Possessor* eindeutig bestimmt.

Ein Beleg aus dem Griechischen mit Possessivpronomen und dem Funktionalbegriff ‚Vater' ist (15). Wie Apollo soll Phoibos auf seinen Vater Zeus achtgeben. Es droht Gefahr von der Mondgöttin Semele.

(15) Nonnos, *Dionysiaka* 1,330-333

hōs	Nómios		klutótokse
als	Herr der Weiden(M): NOM.SG		berühmter Bogenschütze(M): VOC.SG
teòn		poímaine	**tokẽa**
dein: ACC.M.SG		hüten: PRS.IMP.ACT2SG	Vater(M): ACC.SG

ὡς Νόμιος, κλυτότοξε, τεὸν ποίμαινε τοκῆα
‚Als Herr der Weiden, berühmter Bogenschütze, hüte deinen Vater'
wörtl.: ‚Als Herr der Weiden, berühmter Bogenschütze, deinen hüte Vater'

Die Bezugsglieder des Hyperbatons sind nur durch ein Verb getrennt.

4.2 Relationale Begriffe

4.2.1. Verwandtschaftsbegriffe

Nach (Löbner 2005a: 3) bestimmen „relationale Begriffe ihren potenziellen Referenten primär darüber, dass sie zu einer gegebenen anderen Entität (oder mehreren anderen) in einer bestimmten Beziehung stehen."
Vgl. (16) mit ‚Mutter':

(16) Cicero, *In Catilinam* 1,7

nunc	te	patria	quae
nun	du: ACC.SG	Vaterland(F): NOM.SG	welcher: NOM.F.SG
communis		est	**parens**
gemeinsam: NOM.F.SG		sein: PRS.IND.ACT3SG	Mutter(F): NOM.SG
omnium		nostrum	odit
alle: GEN.M.PL		wir: GEN.M.PL	hassen: PF.IND.ACT3SG

et	metuit
und	fürchten: PF.IND.ACT3SG

nunc te patria, quae communis est parens omnium nostrum, odit ac metuit
‚Nun aber hasst und fürchtet dich das Vaterland, das unser aller gemeinsame Mutter ist'
wörtl.: ‚die gemeinsame ist Mutter aller'

communis … omnium nostrum verhält sich wie ein relationales Adjektiv. Allein unbetontes *est* erscheint zwischen den zwei Teilen des Hyperbatons, einem qualifizierenden Adjektiv und dem relationalen Substantiv.[20]

4.2.2. Körperteilbezeichnungen

Dem folgenden Beleg liegt eine *Scene* mit einem Teil-von-Begriff zugrunde:

(17) RV 10,52,5 Gespräch zwischen den Göttern und Agni.

á́	**bāh(u)vóḥ**	vájram	**índrasya**
hinein	Arm(M): LOC.DU	Vajra(M): ACC.SG	Indra(M): GEN.SG

dheyām
legen: AOR.OPT.ACT1SG
á́ bāhvórvájramíndrasya dheyām
‚In die beiden Arme Indras möchte ich den Vajra (Donnerkeil) legen.'
wörtl.: ‚In die beiden Arme den Vajra Indras möchte ich legen.'

In (17) wird zwar der Donnerkeil als Attribut Indras genannt. Der Genitiv ‚Indras' ist aber Attribut zu dem Dual *bāh(u)vóḥ* ‚in die beiden Arme'.

Auf eine *Scene* referiert auch (18). Es ist das für die Götter ausgerichtete griechische Opfermahl mit Tieropfern. *Frames* mit Bezeichnungen von Opfertieren und Teil-von-Begriffen sind hier typisch.

20 *est* ist wohl kein Wackernagel-Element. Es steht nicht an der zweiten Stelle im (Neben-)Satz.

(18) Homer, Ilias 1,40

ḕ	ei	dḗ	poté	toi	katà	**píona**
oder	wenn	schon	einmal	du: DAT.SG	völlig	fett: ACC.N.PL
mēría				ékēa		
Schenkel(N): ACC.PL				brennen: AOR.IND.ACT1SG		
taúrōn			**ēdè**	**aigōn**		
Stier(M): GEN.PL			und	Ziege(F): GEN.PL		

ἢ δή ποτέ τοι κατὰ πίονα μηρί‘ ἔκηα / ταύρων ἠδ‘ αἰγῶν,oder ob ich dir schon einmal fette Schenkel von Stieren und Ziegen verbrannt habe‘

wörtl.: ‚oder ob ich dir schon einmal fette Schenkel verbrannt habe von Stieren und Ziegen‘

‚Fette Schenkel von Stieren und Ziegen‘ enthält sortale Merkmale. Wie auch sonst oftmals steht ein Verb zwischen den beiden Teilen des Hyperbatons.

4.3 Ereignisbegriffe

Auch einen Ereignisbegriff, der zeitliche Gebundenheit signalisiert (Härtl 2015: 159 Anm. 1), findet man in *Frames*. In Boethius’ *Consolatio philosophiae* ist die *Scene* das Ende des Lebens:

(19) Boethius, Consolatio philosophiae 2,3,18

Nam etsi rara est fortuitis manendi fides,

‚Denn wenn auch selten Zufälligkeiten die Zuverlässigkeit des Bleibens haben‘

ultimus	tamen	vitae
jenseitig: NOM.M.SG	dennoch	Leben(F): GEN.SG
dies	mors	quaedam
Tag(M): NOM.SG	Tod(F): NOM.SG	ein gewisser: NOM.F.SG
fortunae	est	etiam
Glück(F): GEN.SG	sein: PRS.IND.ACT3SG	auch
manentis		
bleibend: GEN.F.SG		

ultimus tamen vitae dies mors quaedam fortunae est etiam manentis

‚ist dennoch der letzte Tag des Lebens eine Art von Tod selbst für ein dauerhaftes Glück‘
wörtl.: ‚der letzte dennoch des Lebens Tag ein Tod ein gewisser des Glücks ist auch des dauernden‘

4.4 Abstrakte Begriffe

Des Weiteren erzeugen abstrakte Begriffe *Scenes* mit entsprechenden *Frames*. Beleg (20) hat einen mythologischen Hintergrund. Das Lied handelt von Agnis Verschwinden. Agni ist vor seinen Brüdern geflohen.

(20) RV 10,51,6 Agnis Verschwinden

gauráḥ	ná	**kṣepnóḥ**	avije
Büffel(M): NOM.SG	wie	Schnellen(M): ABL.SG	fliehen: IPF.IND.MED1SG

j(i)yā́yāḥ
Bogensehne(F): GEN.SG
gauró ná kṣepnóravije jyā́yāḥ
‚Wie ein Büffel vor dem Schnellen der Bogensehne floh ich.‘
wörtl.: ‚Wie ein Büffel vor dem Schnellen floh ich der Bogensehne.‘

Das Substantiv ‚Schnellen‘ löst den *Frame* ‚Abschnellen einer Bogensehne‘ aus. Es hat eine Ereignis-Lesart (vgl. dazu Dölling 2015: 50).

Ein Sprecher kann mit einem Hyperbaton auch Ironie verbinden; vgl. dazu das auflösende Bezugsglied *gratia* ‚Dank‘, eine Bezeichnung eines Resultatszustands:

(21) Cicero, *In Catilinam* 1,11

praeclaram	vero	populo	Romano
herrlich: ACC.F.SG	freilich	Volk(M): DAT.SG	römisch: DAT.M.SG
initial/focus-split			
refers		**gratiam**	
zurückgeben: PRS.IND.ACT2SG		Dank(F): ACC.SG	

praeclaram vero populo Romano refers gratiam, qui te, hominem per te cognitum,

nulla commendatione maiorum tam mature ad summum imperium per omnis honorum gradus extulit …
‚Du gibst wirklich dem römischen Volk einen herrlichen Dank zurück, das dich, einen Mann, der nur durch sich bekannt ist, ohne Empfehlung der Vorfahren so frühzeitig in die höchste Machtposition durch jeden Rang der Ehrenämter emporgehoben hat.‘
wörtl.: ‚einen herrlichen freilich dem römischen Volk stattest du ab Dank‘

Da es Cicero in seinen Reden gegen Catilina um die Aufdeckung, Verfolgung und Bestrafung der zweiten Catilinarischen Verschwörung ging, kann ein evaluatives Adjektiv der Bedeutung ‚herrlich‘ in diesem Kontext nur ironisch gemeint sein. Die zugrundeliegende Phrase ist *gratiam referre alicui*.

4.5 Sortale Begriffe

Sortale Begriffe beschreiben in der Regel Bündel von Eigenschaften bzw. Merkmalen und damit eine Kategorie von potenziellen Referenten (Löbner 2015a: 3). Ein *Frame* mit einem solchen Begriff und einem Hyperbaton erscheint gleich zu Anfang der *Ilias*. Agamemnon will Chryses’ Tochter nicht herausgeben:

(22) Ilias 1,26

mḗ	se	géron	**koílēisin**
nicht	du: ACC.SG	Alter(M): VOC.SG	hohl: DAT.F.PL
egṑ		**parà**	**nēusì**
ich: NOM.SG		bei	Schiff(F): DAT.PL
kichánō			
antreffen: AOR.SUBJ.ACT1SG			

μή σε, γέρον, κοίλῃσιν ἐγὼ παρὰ νηυσὶ κιχείω
‚Alter, nicht will ich Dich bei den hohlen Schiffen antreffen‘
wörtl.: ‚nicht dich, Alter, den hohlen ich bei den Schiffen will ich antreffen.‘

Da die *Scene* am Meer spielt, evoziert das qualifizierende Adjektiv ‚hohl‘ das Konzept ‚Schiff‘. ‚Hohl‘ ist ein häufiges Epitheton zu diesem Wort bei Homer und erfüllt auch die metrischen Erfordernisse des Hexameters. Homer

kann aber deswegen Hyperbata wie κοίλῃσιν … νηυσὶ bilden, weil in seinem hochartifiziellen epischen Idiom bestimmte Adjektive als *epitheta ornantia* fest mit festen Substantiven verbunden sind. Wie bemerkt (3), sind Hyperbata ein Merkmal der homerische Formelsprache.

Ein *Frame* mit einem sortalen Begriff wird auch in einem Beleg aus den *Jātakas* aufgerufen. Das *Sihacamma-Jātaka* handelt von einem Esel, der sich für etwas Besseres ausgibt. Er trägt die Haut eines Löwen:

(23) *Sihacamma-Jātaka* 189

na	etam	sīhassa	naditam	na
nicht	dieser: NOM.N.SG	Löwe(M): GEN.SG	Ruf(N): NOM.SG	nicht
vyagghassa	na	dīpino	pāruto	
Tiger(M): GEN.SG	nicht	Leopard(M): GEN.SG	bedeckt: NOM.M.SG	
sīhacammena		**jammo**		
Löwenhaut(N): INSTR.SG		elend: NOM.M.SG		
nadati		**gadrabho**	ti	
rufen: PRS.IND.ACT3SG		Esel(M): NOM.SG	QUOT	

Na'etaṁ sīhassa naditaṁ na vyagghassa na dīpino, pāruto sīhacammena jammo nadati gadrabho ti.

‚Das ist nicht der Ruf eines Löwen, nicht eines Tigers, nicht eines Leoparden, bedeckt mit einer Löwenhaut ruft ein elender Esel!'

wörtl.: ‚bedeckt mit einer Löwenhaut ein elender ruft Esel'

Auf das situationsbezogene Adjektiv[21] ‚elend' kann nur das Substantiv ‚Esel' folgen. Das evaluierende Adjektiv ist wie schon in anderen Fällen durch ein finites Verb von seinem Bezugswort getrennt.

Anders als im Griechischen und Lateinischen, wo Hyperbata auch in den späteren Sprachstufen auftreten, sind die Belege in den *Jātakas* vereinzelt. Man kann so annehmen, dass im Mittelindischen das Hyperbaton im Schwinden begriffen war.

21 Zu situationsbezogenen Modifikatoren vgl. Schäfer 2015: 150f.; Fortmann et al. 2015: 4f.

Hyperbata in *Frames* mit sortalen Begriffen sind auch bei Boethius bezeugt:

(24) Boethius, *Consolatio philosophiae* 2,4,17

anxia	enim	res	est
ängstigend: NOM.F.SG	nämlich	Sache(F): NOM.SG	sein: PRS.IND. ACT3SG
humanorum		condicio	**bonorum**
menschlich: GEN.N.PL		Lage(F): NOM.SG	Gut(N): GEN.PL

Anxia enim res est humanorum condicio bonorum
‚Denn die Beschaffenheit menschlicher Güter gibt Anlass zur Sorge.'
wörtl.: ‚beängstigend nämlich die Sache ist der menschlichen Lage Güter'

Zu der *Scene* ‚Lebensumstände der Menschen' gehören seine Güter. Das relationale Adjektiv ‚menschlich' im Genitiv Plural Neutrum kongruiert mit ‚Güter'.

Wie ein relationales Adjektiv verhält sich das Adjektiv ‚innerlich'. *intestinam* ‚inneres' ist sinngemäß mit dem sortalen Begriff *rei publicae* ‚des Staates' verbunden: ‚Verderben gegen das Innere des Staats'; vgl. *in medias res* ‚mitten in die Dinge'.

(25) Cicero, *In Catilinam* 1,2
Eorum autem castrorum imperatorem ducemque hostium intra moenia
‚Von ihrem Lager aber (seht ihr) den Befehlshaber und Anführer der Feinde innerhalb der Stadtmauern'

atque	adeo	in	senatu	videtis
und	gar	in	Senat(M): ABL.SG	sehen: PRS.IND. ACT2PL
intestinam		aliquam		cotidie
innerer: ACC.F.SG		irgendein: ACC.F.SG		täglich
perniciem			rei	publicae

Verderben(F): ACC.SG Sache(F): GEN.SG öffentlich: GEN.F.SG molientem
planen: PRT.PRS.ACT.ACC.M.SG
atque adeo in senatu videtis intestinam aliquam cotidie perniciem rei publicae molientem.
‚und sogar im Senat seht ihr, wie er täglich irgendein Verderben gegen das Innere des Staates ausbrütet.‘
wörtl.: ‚und sogar im Senat seht ihr (ihn) ein inneres irgendein täglich Verderben des Staates ausbrütend‘

Die *Scene* ist wieder die Gerichtsverhandlung gegen Catilina.

Die folgende *Scene* in Nonnos’ *Dionysiaka* ergibt sich dagegen aus dem mythologischen Kontext. Voraus geht: Hera ist eifersüchtig auf die Mondgöttin Semele, eine besonders schöne Königstochter, die Zeus als Mutter seines Sohnes Dionysos auserkoren hatte. Phoibos soll auf seinen Vater Zeus aufpassen; vgl. (15). Hera kann sich nicht zurückhalten:

(26)(a) Nonnos, *Dionysiaka* 1,326–329.
καὶ Κρονίδην ὁρόωσα πόθῳ δεδονημένον Ἥρη
ζηλομανὴς γελόωντι χόλῳ ξυνώσατο φωνήν:/
Φοῖβε, τεῷ γενετῆρι παρίστασο, μή τις ἀροτρεὺς
Ζῆνα λαβὼν ἐρύσειεν ἐς ἐννοσίγαιον ἐχέτλην.
αἴθε λαβὼν ἐρύσειεν, ὅπως Διὶ τοῦτο βοήσω: ὅπως Διὶ τοῦτο βοήσω: / τέτλαθι διπλόα κέντρα καὶ ἀγρονόμων καὶ Ἐρώτων.
‚Und als Hera den Kronossohn sah, von Verlangen erschüttert, erhob sie rasend vor Eifersucht mit spottendem Zorn die Stimme: Phoibos, steh deinem Erzeuger bei, damit nicht irgendein Pflüger Zeus ergreifen und ihn in seinen erderschütternden Pflug zerren kann! Wenn er [Phoibos] ihn [Zeus] doch ergreifen und zerren würde, damit ich Zeus dies zurufen könnte: Ertrag die zweifache Stachelknute, sowohl (die) der Landmänner als auch (die) der Liebesgötter!‘

Semele ist äußerst fruchtbar: Sie hat mit ihrem schönen ewig jungen Liebhaber, dem Hirten Endymion, 50 Kinder, obwohl er immer schläft:

(26)(b) Nonnos, Dionysiaka 1,330–333
μὴ Κρονίδην ζεύξειε βοῶν ἐλάτειρα Σελήνη
‚damit den Kronossohn nicht Selene, die Treiberin der Rinder einjoche'
mḕ léchos Endumíōnos
damit nicht Bett(N): ACC.SG Endumion(M): GEN.SG
ideῖn speúdousa
sehen: AOR.INF.ACT eilend: PRT.PRS.ACT.NOM.F.SG
nomẽos Tsēnòs hupostíkseien
Hirte(M): GEN.SG Zeus(M): GEN.SG einritzen:AOR.OPT.ACT3SG
apheidéi nõton **himásthlēi**
schonungslos: DAT.F.SG Rücken(N): ACC.SG Peitsche(F): DAT.SG
μὴ λέχος Ἐνδυμίωνος ἰδεῖν σπεύδουσα νομῆος / Ζηνὸς ὑποστίξειεν ἀφειδέι νῶτον ἱμάσθλῃ.
‚damit sie nicht, wenn sie eilt, um das Bett des Hirten Endumion zu sehen, den Rücken des Zeus mit schonungsloser Peitsche einritze.'
wörtl.: ‚damit sie nicht … einritze mit schonungsloser den Rücken Peitsche.'

Eine Peitsche hinterlässt blutige Striemen. Ein Verb wie ‚einritzen' kann so den in (26)(b) genannten *Frame* mit dem evaluativen Adjektiv ‚schonungslos' und dem sortalen Nomen ‚Peitsche' verursachen.

Mythologisches Wissen ermöglicht weiterhin die Deutung der folgenden Stelle aus den *Dionysiaka*, die gleich zwei Hyperbata, die sich auf das Individualnomen *Kadmos* beziehen, enthält. Kadmos' Schwester Europa wurde von Zeus entführt. Kadmos sucht nach ihr und hilft Zeus im Kampf gegen Typhon, der diesem die Blitze geraubt hat. Zeus fordert Kadmos auf, mit seiner Hirtenflöte den Sinn des Typhon zu verwirren. Dafür macht Zeus Kadmos zum Beschützer und Beischläfer der *Harmonia*. Auch hier kommt ein Wort für einen Schmerzen verursachenden Gegenstand, ‚Stachel', vor:

(27) Nonnos, Dionysiaka 1,404–407
Kadmeíēs dè echétō
zu Kadmos gehörig: GEN.F.SG aber in sich haben: PRS.IMP.ACT3SG

phrenothelgéos oĩstron
das Herz bezaubernd: GEN.F.SG Stachel(M): ACC.SG
aoidēs
Gesang(F): GEN.SG
Καδμείης δ’ ἐχέτω φρενοθελγέος οἶστρον ἀοιδῆς
‚den Stachel aber des das Herz bezaubernden Gesanges des Kadmos soll er [Typhon] in sich haben‘
wörtl.: des Kadmos aber soll er haben des Herz bezaubernden den Stachel Gesangs‘
ὅσσον ἐγὼ πόθον ἔσχου ἐς Εὐρώπης ὑμεναίους
‚wie sehr ich [Zeus] das Verlangen in mir hatte nach Europas Hochzeitsgesängen‘

Die zusammengehörigen Wörter ‚des Kadmos Gesang‘ sind durch vier Wörter getrennt und ‚des Herz bezaubernden Gesanges‘ durch ein Wort, ‚Stachel‘. Der Name *Kadmos* fungiert als *genitivus subjectivus* am Ende des Satzes. Es ist bekannt, dass Kadmos’ Gesang die Menschen berührt hat, so dass voran stehendes ‚des Herz bezaubernden‘ ebenfalls auf das Substantiv ‚Gesang‘ hindeutet.

4.6 Distribution von Hyperbata in Nähe- und Distanzdiskursen

Nachdem nur Belege für Hyperbata in direkten Reden näher untersucht wurden, bleibt noch die Frage nach der Distribution dieser rhetorischen Figur in Nähe- und Distanzdiskursen insgesamt. Die Suchanfragen haben für Hyperbata in direkten Reden und narrativen Strukturen folgende Verteilung ergeben (wobei in Klammern die Gesamtzahl der narrativen Strukturen insgesamt angegeben wird):

Hyperbata		
Altindisch	Direkte Rede 137	Narration 15 (77 narrative Strukturen)
Mittelindisch	Direkte Rede 8	Narration 0 (489 narrative Strukturen)
Altgriechisch	Direkte Rede 57	Narration 22 (41 narrative Strukturen)
Nonnos	Direkte Rede 70	Narration 316 (80 narrative Strukturen)
Latein	Direkte Rede 55	Narration 4 (17 narrative Strukturen)
Spätlatein	Direkte Rede 114	Narration 0 (0 narrative Strukturen)

Die Suchanfragen sind:

discourse=/.*direct.*/ _i_ style=/.*hyperbaton.*/
discourse=/.*narration.*/ _i_ style=/.*hyperbaton.*/
discourse=/.*narrative.*/ _i_ style=/.*hyperbaton.*/

In der Regel übertrifft die Anzahl der Hyperbata in der direkten Rede die der narrativen Strukturen. Eine Ausnahme bildet Nonnos' Dionysiaka. Sein „Reichtum [an] ornamentaler Durchbildung des Stils" ist ein Sonderfall.[22]

5 Fazit

Die eingangs gestellte Frage nach der Unterscheidung von Distanz- und Nähediskursen wurde anhand von Subjekten und Topiks am Satzanfang untersucht und mit der Art von Konfigurationalität verknüpft. Da initiale Topiks oftmals als Subjekte erscheinen, weist diese Doppelfunktion ebenso wie die hohe Anzahl initialer Subjekte ohne Topikfunktion in Richtung Syntaxkonfigurationalität. Das *Contastive Topic* kommt aber häufiger in der direkten Rede vor und wurde so als ein Phänomen des Nähediskurses gewertet. Anders verhält es sich beim Hyperbaton. Diskontinuierliche Konstituenten sind generell ein Merkmal von Diskurkonfigurationalität. Dabei ist auch die Anzahl der Hyperbata in direkten Reden höher als in narrativen Strukturen. (Mit seinem hochartifiziellen idiosynkratischen Stil bildet Nonnos eine Ausnahme. Damit ist z.B. die Sonderstellung in Sachen Hyperbaton, die dem „Mittelgriechischen" im Fazit zugeschrieben wird, illusionär. Nonnos ist so nicht repräsentativ für die Entwicklungsstufe des Griechischen, sondern ausschließlich für sich selber.–) Zudem repräsentiert das Hyperbaton eine inkrementelle Syntax. Da diese Art von Syntax in der gesprochenen Sprache ihren Ursprung hat, könnte das Hyperbaton als syntaktische Variante auch in den Distanzdiskurs übernommen worden sein. Es wäre

22 Friedländer 1912: 49.

so möglicherweise ein Relikt aus der oralen Kultur (dazu Lühr 2019).[23] Am leichtesten sind Hyperbata im Falle von funktionalen und relationalen Begriffen und wohl auch Ereignisbegriffen aufzulösen, dann folgen abstrakte und sortale Begriffe. Alle diese Begriffe können *Frames* zugeordnet werden. Die zugrundeliegenden *Scenes* sind zuweilen der Mythologie entnommen. Auch ein Beispiel für Ironie wurde gefunden. Selbst wenn nur ein Wort oder wenige Wörter zwischen den beiden Gliedern eines Hyperbatons erscheinen, gibt erst die Auflösung des zweiten Bezugsgliedes Aufschluss über den tatsächlichen Gesamtbegriff dieser Redefigur. AdressatInnen sind gezwungen, genau mitzulesen oder, noch besser, zuzuhören.

Bibliographie

Ágel, V./Hennig, M. (2006): Theorie des Nähe- und Distanzsprechens. In: V. Ágel/M. Hennig (Hg.), *Grammatik aus Nähe und Distanz. Theorie und Praxis am Beispiel von Nähetexten 1650 –2000*. Tübingen: Niemeyer, 3–31.

Ágel, V./Hennig M. (2006a): Praxis des Nähe- und Distanzsprechens. In: V. Ágel/M. Hennig (Hg.), *Grammatik aus Nähe und Distanz. Theorie und Praxis am Beispiel von Nähetexten 1650 –2000*. 33–74.

Ágel, V./Hennig, M. (Hg.) (2006b): *Grammatik aus Nähe und Distanz. Theorie und Praxis am Beispiel von Nähetexten 1650-2000*. Tübingen: Niemeyer.

Aringer, N. (2012): Kadmos und Typhon als vorausdeutende Figuren in den Dionysiaka. Bemerkungen zur Kompositionskunst des Nonnos von Panopolis. *Wiener Studien* 125, 85-105.

Auer, P. (2007): Syntax als Prozess. In: H. Hausendorf (Hg.), *Gespräch als Prozess. Linguistische Aspekte der Zeitlichkeit verbaler Interaktion*. Tübingen: Narr, 95–124.

Auer, P. (2009): On-line syntax: thoughts on the temporality of spoken language. *Language Sciences* 31, 1–13.

23 Orale Kulturen nehmen gerne die gesprochene Sprache als Grundlage ihrer Grammatik, während des Schreibens kundige Kulturen dazu die geschriebene Sprache nutzen (Devine/Stephens 1999: 207f.). Markovic (2006) nimmt an, dass "one of the most salient functions of hyperbaton in Greek literary language is to signal or reinforce the end of a colon or sentence."

Auer, P. (2015): The temporality of language in interaction: projection and latency. In: A. Deppermann/S. Günthner (Hg.), *Temporality in Interaction*. Amsterdam: Benjamins, 27–56.

Biber, D./Conrad, S. (2009): *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Breindl, E. (2008): Die Brigitte nun kann der Hans nicht ausstehen. Gebundenes Topik im Deutschen. In: E. Breindl/M. Thurmair (Hg.), *Erkenntnisse vom Rande. Zur Interaktion von Prosodie, Informationsstruktur, Syntax und Bedeutung. Zugleich Festschrift für Hans Altmann zum 65. Geburtstag. Deutsche Sprache (Themenheft)* 2, 27–49.

Büring, D. (1999): Topik. In: P. Bosch/R. von der Sandt (Hg.), *Focus, Linguistic, Cognitive and Computational Perspectives*. Cambridge: Cambridge University Press, 142–165.

Devine, A. M./Stephens, L. D. (2000): *Discontinuous syntax. Hyperbaton in Greek.* New York/Oxford: Oxford University Press.

Dölling, J. (2015): Sortale Variation der Bedeutung bei *ung*-Nominalisierungen. In: C. Fortmann/A. Lübbe/I. Rapp (Hg.), *Situationsargumente im Nominalbereich.* Berlin/Boston: de Gruyter, 49–91.

Fanselow, G. (1987): *Konfigurationalität. Untersuchungen zur Universalgrammatik am Beispiel des Deutschen*. Tübingen: Niemeyer.

Fanselow, G./Féry, C. (2006): Prosodic and Morphosyntactic Aspects of Discontinuous Noun Phrases: a Comparative Perspective (http://user.uni-frankfurt.de/~cfery/publications/Prosodic_and_morphosyntactic_aspects_of_discontinuous_NPs.pdf).

Fillmore, C. J. (1977): Scenes and Frames Semantics. In: A. Zampolli (Hg.), *Linguistic Structure Processing*. Amsterdam: North-Holland Pub. Co., 55–81.

Fortmann, C./Lübbe, A./Rapp, I. (2015): Einführung. In: C. Fortmann/A. Lübbe/I. Rapp (Hg.), *Situationsargumente im Nominalbereich.* Berlin/Boston: de Gruyter, 1–9.

Fortmann, C./Lübbe, A./Rapp, I. (Hg.) (2015): *Situationsargumente im Nominalbereich. Linguistische Arbeiten.* Berlin/Boston: de Gruyter.

Friedländer, P. (1912): Die Chronologie des Nonnos von Panopolis. *Hermes*, 471, 43–59.

Härtl, H. (2015): Situationsargumente von Nicht-Köpfen: Verb-Nomen-Komposita im Zusammenspiel von Morphologie, Syntax und Pragmatik. In: C. Fortmann/A.

Lübbe/I. Rapp (Hg.), *Situationsargumente im Nominalbereich*. Berlin/Boston: de Gruyter, 159–184.

Hale, K. (1983): Warlpiri and the Grammar of Non-Configurational Languages. *Natural Language and Linguistic Theory* 1, 5-47.

Hopper, P. J. (1987): Emergent grammar. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* 13, 139–157.

Hopper, P. J. (2001): Grammatical constructions and their discourse origins: prototype or family resemblance? In: M. Mütz/S. Niemeier/R. Dirven (Hg.), *Applied Cognitive Linguistics I: Theory and Language Acquisition*. Berlin/Boston: de Gruyter, 109–129.

Hopper, P. J. (2011): Emergent grammar and temporality in interactional linguistics. In: P. Auer/S. Pfander (Hg.), *Constructions: Emerging and Emergent*. Berlin/Boston: de Gruyter, 22–44.

Jacobs, J. (1997): I-Topikalisierung. *Linguistische Berichte* 169, 91–133.

Kiss, K. É. (1998): Identificational Focus versus Informational Focus. *Language* 74, 2, 245–273.

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.

Kozianka, M./Zeilfelder, S. (2016): Das Hyperbaton in altindogermanischen Sprachen. In: R. Lühr (Hg.), *Idiosynkrasie. Neue Wege ihrer Beschreibung. Unter Mitarbeit von Satoko Hisatsugi*. Wiesbaden: Reichert, 71–80.

Krifka, M. (2007): Basic notions of information structure. In: C. Féry/G. Fanselow/M. Krifka (Hg.), *Interdisciplinary studies on information structure* 6. Potsdam: Universitätsverlag Potsdam.

Krisch, Th. (1998): Zum Hyperbaton in altindogermanischen Sprachen. In: W. Meid (Hg.), *Sprache und Kultur der Indogermanen. Akten der X. Fachtagung der Indogermanischen Gesellschaft, Innsbruck, 22*. –28. September 1996. Innsbruck. Institut für Sprachwissenschaft, 351–384.

De Kuthy, K. (2002): *Discontinuous NPs in German. A Case Study of the Interaction of Syntax, Semantics, and Pragmatics*. Stanford: CSLI Publications.

Lausberg, H. (2008): *Handbuch der Literarischen Rhetorik*. [4]Stuttgart: Franz Steiner.

Löbner, S. (1985): Definites. *Journal of Semantics* 4: 279–326.

Löbner, S. (2005): Funktionalbegriffe und Frames – Interdisziplinäre Grundlagenforschung zu Sprache, Kognition und Wissenschaft. In: A. Labisch (Hg.), *Jahrbuch der Heinrich-Heine-Universität Düsseldorf 2004*. Düsseldorf: Heinrich-Heine-Universität, 463–477.

Löbner, S. (2005a): FFF. Forschergruppe „Funktionalbegriffe und Frames“ [https://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Forschung/FFF/Allgemein/Antrag_FFF_gesamt.pdf].

Löbner, S. (2011): Concept Types and Determination. *Journal of Semantics* 28, 279–333.

Lühr, R. (1990): Adjazenz in komplexen Nominalphrasen. In: G. Fanselow/F. Sascha (Hg.), *Strukturen und Merkmale syntaktischer Kategorien*. Tübingen: Narr, 33–50.

Lühr, R. (2010): Fokuspartikeln im Althochdeutschen. In: Y. Desportes/F. Simmler/C. Wich-Reif (Hg.), *Mikrostrukturen und Makrostrukturen im älteren Deutsch vom 9. bis zum 17. Jahrhundert: Text und Syntax. Akten zum Internationalen Kongress an der Université Paris Sorbonne (Paris IV) 6. bis 7. Juni 2008*. Berlin: Weidler Verlag, 103–120.

Lühr, R. (2015): Traces of discourse configurationality in older Indo-European languages? In: C. Viti (Hg.), *Perspectives on Historical Syntax*. Amsterdam: John Benjamins, 203–232.

Lühr, R. (2016): Discontinuous Syntax: Hyperbaton in older Indo-Europepan Languages. In: H. Marquardt/S. Reichmuth/J. Virgilio García Trabazo (Hg.), *Studia linguistica in honorem Johannes Tischler septuagenarii dedicata*. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck, 153–166.

Lühr, R. (2019): Konfigurationale Merkmale im Anatolischen. In: S. Schaffner (Hg.), *Gedenkschrift Hoffmann* (in press).

Lühr, R. (2020): Satzanfänge im Hethitischen. In: R. J. Kim/J. Mynářová/P. Pavúk (Hg.): *Hrozný and Hittite. The First Hundred Years*. Leiden: Brill (in press).

Lühr, R./Zeilfelder, S. (2011): Zur Interdependenz von Diskursrelationen und Konnektoren in indogermanischen Sprachen: Kontrast und Korrektur. In: E. Breindl/G. Ferraresi/A. Volodina (Hg.), *Satzverknüpfungen. Zur Interaktion von Form, Bedeutung und Diskursfunktion*. Berlin/Boston: de Gruyter, 107–148.

Luraghi, S. (2013): The rise (and possible downfall) of configurationality. In: S. Luraghi/V. Bubenik (Hg.), *The Bloomsbury Companion to Historical Linguistics*. London: Bloomsbury Publishing (US), 219–229.

Markovic, D. (2006): Hyperbaton in the Greek Literary Sentence. *Greek, Roman, and Byzantine Studies* 46, 127-146.

Menge, H. (2000): *Lehrbruch der lateinischen Syntax und Semantik*. Völlig neu bearb. von Th. Burkard/M. Schauer. Darmstadt: Wissenschaftliche Buchgesellschaft.

Nonhoff, M. (2004): Diskurs. In: G. Göhler/M. Iser/I. Kerner (Hg.), *Politische Theorie. 25 umkämpfte Begriffe zur Einführung*. Wiesbaden: VS Verlag, 63–79.

Reisigl, M./Ziem, A. (2014): Diskursforschung in der Linguistik. In: J. Angermuller/M. Nonhoff/E. Herschinger/F. Macgilchrist/M. Reisigl/J. Wedl/D. Wrana/A. Ziem (Hg), *Diskursforschung. Ein interdisziplinäres Handbuch (2 Bde.)*. Bielefeldt: Transcript Verlag, 70–110.

Ross, J. R. (1967): *Constraints on Variables in Syntax*. Doctoral dissertation, Massachusetts Institute of Technology.

Schäfer, M. (2015): Nominalmodifikation im Englischen und Ereignisargumente. Zwei Fallstudien. In: C. Fortmann/A. Lübbe/I. Rapp (Hg.), *Situationsargumente im Nominalbereich.* Berlin/Boston: de Gruyter, 133–158.

Schnaus, S. (2006): Die Dialoglieder im altindischen Rigveda. Kommentar unter besonderer Berücksichtigung textlinguistischer Kriterien. Hamburg: Dr. Kovač.

Schnelle, G. (2017): *Funktional bedingte Variation in der Evangelienharmonie Otfrids von Weißenburg. Eine methodische Annäherung an eine variationistische korpusbasierte Registerstudie des Althochdeutschen*. Masterarbeit Humboldt Universität zu Berlin.

Umbach, C. (2001): Restriktion der Alternativen. In: A., Steube (Hg.), *Kontrast – lexikalisch, semantisch, intonatorisch*. Linguistische Arbeitsberichte 77, 165–198.

Umbach, C. (2003): Anaphoric restriction of alternative sets: On the role of bridging antecedents. In: M. Weisgerber (Hg.), *Proceedings of „Sinn und Bedeutung VII". Konstanz (Konstanz Linguistics Working Papers 114)*, 310–323.

Viti, C. (2015): *Variation und Wandel in der Syntax der alten indogermanischen Sprachen*. Tübingen: Narr.

Wackernagel, J. (1892): Über ein Gesetz der indogermanischen Wortstellung. *Indogermanische Forschungen* 1, 333–436.

Wengeler, M./Ziem, A. (Hg.) (2018): *Diskurs, Wissen, Sprache. Linguistische Annäherungen an kulturwissenschaftliche Fragen*. Berlin/Boston: de Gruyter.

Zimmermann, Malte (2008): Contrastive Focus and Emphasis. *Acta Linguistica Hungarica* 55,3–4, 347–360.

Tony McEnery and Gavin Brookes

Register, belief and violence: A multi-dimensional approach

1 Introduction

This chapter examines the linguistic features of texts promoting extremist violence. According to the Government of the United Kingdom – the context in which our work is situated – *extremism* can be defined as 'the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty and the mutual respect and tolerance of different faiths and beliefs.' (HM Government 2015: 9). The Government also notes that it 'regard[s] calls for the death of members of our armed forces as extremist' (ibid.). As this definition indicates, in legal terms at least, extremism is a broad concept capable of encapsulating a wide range of actions, including linguistic (i.e. 'vocal') behaviours, which could be deemed to threaten any of the Country's major institutions or to impinge upon the rights and safety of its citizens. In this chapter, we examine texts which promote ideologically motivated extremist violence, focusing in particular on texts found in the possession of convicted violent jihadists. Note that, as a result, our use of quotations from the texts in question will be sparing – we have no wish to further the promotion of violence ourselves. Accordingly, where we do, for illustrative purposes, provide brief examples, we will not produce any which explicitly promote violence.

Conceptualisations of the relationship between extremism and Islam are, as Baker and Vessey (2018: 257) point out, both 'complex and contested'. A relevant distinction here being that between the adjectives 'Islamic' and 'Islamist', where 'the former has been popularly used in western media (and implies a form of extremism connected to Islam generally) and the latter refers to extremism connected more specifically to politically motivated Islam' (ibid.). The related term, 'Islamism', they contend, refers to 'the desire to impose a

version of Islam over society while other terms like "militant Islam", "radical Islam" or "fundamentalist Islam" complicate definitions further and are also often found in western news media, implying that Islam is an intolerant and insensitive religion' (see also: Kramer 2003; Baker, Gabrielatos and McEnery 2013; Esposito 2014). As well as providing a definition of extremism, the British Government also offers a definition of Islamist extremism specifically, which it describes as 'any form of Islam that opposes democracy, the rule of law, individual liberty and mutual respect and tolerance of different faiths and beliefs' (HM Government 2013: 1). However, given the clear and understandable sensitivities regarding the use of terms such as Islamism and Islamist, in this chapter we will refer instead to *violent jihadism* when discussing the motivations of the people from whom the texts under study were acquired.

In this chapter we analyse the register features of violent jihadist texts by drawing on techniques from corpus linguistics. Corpus linguistics refers to a group of methods that use specialist computer programs to study language in large bodies of naturally occurring, machine-readable language (see McEnery and Hardie 2012). Such a body of data is known as a *corpus* (pl. *corpora*) – the Latin word for 'body'. Corpora are usually assembled in a principled manner with the aim of representing a language or particular linguistic variety on a broad scale (Biber, Conrad and Reppen, 1996). Based on a corpus containing texts collected by violent jihadists, we apply an analytical framework known as Multi-Dimensional Analysis (introduced in the next section) to ascertain the extent to which the texts in our corpus constitute a single register or whether they form a range of registers. Furthermore, we also investigate whether the register features of these texts differ according to the degrees of extremism that they are judged to encode and propagate. As will become clear, the answers to these questions are not only linguistically interesting, but also raise considerations for those working within organizations, such as counter-terrorism forces, which are responsible for identifying such texts both quickly and reliably.

We should note at this point that while we are focusing on violent jihadists here, we do not perceive Islam to be a religion that is in any way characteristic of extremism or particularly likely to motivate violence. As such, we could have studied extremist language in texts produced by and for members affil-

iating with other religions and ideological groups such as fascists (as studied by Richardson, 2017) or white supremacists (as studied by Brindle, 2017). We study violent jihadists here because we were provided with a privileged insight into the texts collected by members of this ideological group, as will be discussed in Section 3. Before considering that data, however, the next section provides a theoretical and methodological backdrop to this study by introducing the central concept of register and the aforementioned Multi-Dimensional Analysis approach to register.

2 Register and Multi-Dimensional Analysis

The perspective of register combines the analysis of linguistic characteristics that are frequent within a text variety with the analysis of the real-world situations in which that variety is used. The assumption underlying analyses of register is thus that linguistic features are functional, with particular features bearing an association with texts' communicative purposes and situational contexts. Registers can therefore be considered groupings of texts that are defined by factors that are external to the texts themselves, such as the social or situational conditions of their medium, their contexts of production, or their purpose. Biber and Conrad (2009: 6) point out that the description of a register covers three major components: 'the situational context, the linguistic features, and the functional relationships between the first two components' (ibid.). They elaborate:

> Registers are described for their typical lexical and grammatical characteristics: their linguistic features. But registers are also described for their situational contexts, for example whether they are produced in speech or writing, whether they are interactive, and what their primary communicative purposes are. […] [L]inguistic features are always functional when considered from a register perspective. That is, linguistic features tend to occur in a register because they are particularly well suited to the purposes and situational context of the register. Thus, the third component of any register description is the functional analysis.
>
> (Biber and Conrad 2009: 6)

As noted, to examine the register features of the extremist texts in our data, we subject those texts to a Multi-Dimensional Analysis (MDA). MDA is an approach to register analysis developed by the linguist Douglas Biber during the 1980s (Biber 1984, 1988) as a way of identifying the major linguistic parameters along which textual registers vary in English. MDA is driven by a lexico-grammatical account of register variation, since Biber's argument is that registers are formed by distinct combinations of words and grammatical categories. MDA rests upon the use of factor analysis to identify the co-occurrence of particular linguistic features in a text or group of texts. Biber (1988) uses 67 of these, which are grouped into 16 broader categories including, *inter alia*, tense and aspect markers, place and time adverbials and pronouns and pro-verbs (see also: Conrad and Biber 2001: 18–19). These 67 features form the basis for the investigation presented here.[1]

To give an example of how these features help to identify groups of linguistic features relevant to identifying a register, Biber (1988) observed a pattern whereby texts with a high frequency of, *inter alia*, private verbs (e.g. *believe*, *think*) are also likely to exhibit a high frequency of *that*-deletion and contractions, as well as the lower frequency of features such as nouns, prepositions and attributive adjectives. These features combine in different ways to form different 'dimensions', or 'sets of syntactic and lexical features that co-occur frequently in texts' (Biber 1989: 5), along which registers place themselves – so. The features bundle to create a number of dimensions and the registers place themselves in distinct configurations along those dimensions.

The dimensions involved in MDA are interpreted and labelled in terms of their perceived functions. Biber (1988) proposed six major dimensions of variation in English, to which he assigns the following functional labels: Dimension 1: Involved v. Informational; Dimension 2: Narrative v. Non-Narrative; Dimension 3: Elaborated v. Situation-Dependent; Dimension 4: Overt Expression of Argumentation; Dimension 5: Abstract v. Non-Abstract; and Dimension 6: Online Informational Elaboration. In MDA, each text comprising the data can be situated along each dimension in accordance with the di-

1 Readers interested in finding out more about the features can find a useful crib sheet here: http://corpora.lancs.ac.uk/stats/docs/multidimensional.pdf.

mension score assigned to it. The mean dimension score assigned to a group of texts can then be used to characterise its discourse. For example, Biber's Dimension 1, 'Involved v. Informational Production', comprises 25 features with high scores, including, for example, the use of private verbs, *that*-deletion, contractions, present tense verbs, and second-person pronouns, meanwhile features with low negative loadings include nouns, word length, prepositions, and attributive adjectives (Berber Sardinha 2018: 127). In this case, the frequent cooccurrence of the features along the positive pole results in the texts in question being interpreted as having an 'involved production' communicative function. The cooccurrence of features along the negative pole, on the other hand, indicates a shared communicative function of 'informational production'. Because the features along the positive and negative poles tend not to occur with similar frequency within the same texts, the presence of the features of one (in this case, 'involved production') usually indicates that the features of the other (e.g. 'informational production') are largely absent. Other dimensions, such as 2 and 4, have a single pole, meaning these dimensions are characterised by the frequency or absence of a single set of linguistic features. For example, for Dimension 2, texts can be characterised either as containing narrative or non-narrative features, and Dimension 4 as containing either overt or non-overt persuasion (Berber Sardinha 2018: 128).

When carrying out MDA, each text in a corpus is simultaneously scored for each dimension using standardised counts of the relevant features. This means that each text will be assigned a score for each dimension. Analysts typically proceed by calculating the mean dimension scores for each of the registers represented in their data, leading to the characterisation of registers in terms of the aforementioned dimensions. For example, on the basis of these dimensions, Biber (1988) identified the following registers: General narrative exposition; Imaginative narrative; Involved persuasion; Learned exposition; and Scientific exposition. The characteristics of individual registers become more salient, and are rendered more apparent, when their mean dimension scores are compared against each other, essentially illuminating the most salient linguistic characteristics of each.

Since the development of the MDA approach, research carried out by Biber, his colleagues and others has focused on extending this method and/

or on applying it to new areas of inquiry. Indeed, the MDA framework has been applied to the analysis of an impressive and growing range of registers and discourse domains, as well as to an increasing number of languages, where the patterns of register variation originally put forward by Biber (1988) have proven to be a useful starting point for such investigations. Although language-wide studies have been carried out, the majority of MDA research focuses on language used in specific contexts, producing accounts of the registers that are characteristic of domains as diverse as university writing (Biber 2006), televised dialogue (Quaglio 2009), call centre interactions (Friginal 2008), pop song lyrics (Bértoli-Dutra 2014), medical encounters (Staples 2015) and political tweets (Clarke and Grieve 2019), to name just a few (see Biber 2019 for a comprehensive overview). This chapter presents a study of a relatively under-researched type of discourse which has yet to be analysed using the MDA framework; extremist language.

Ours is not the first linguistic study concerned with the discourse of Islamist extremism. For example, Droogan and Peattie (2016) examined shifts in the themes in Al Qaeda's *Inspire* magazine using a modified form of thematic network analysis. Wignell, Tay and O'Halloran (2017) utilised a multimodal approach to discourse analysis in order to explore the use of image and text in the magazine of the so-called Islamic State of Iraq and Syria (ISIS), *Dabiq*, as well as in online materials produced by an affiliated British group, Rayat al Tawheed. These authors were particularly interested in how multimodal ensembles functioned to rally assumed reader-supporters. Also adopting a multimodal perspective, Ingram (2017) compared both of the aforementioned publications, *Inspire* and *Dabiq*, in a study which analysed how each publication used language and imagery in their attempts to appeal to and radicalise their readerships. A small body of work in this area has also employed more quantitative, including corpus linguistic, methods in order to study the language associated with Islamist extremism. Prentice et al. (2011) utilised the corpus analysis tool *Wmatrix* (Rayson 2008) to analyse the persuasive strategies emergent from texts that they claimed incited violence in the context of the Gaza conflict. In a later study, Prentice et al. (2012) used this same tool in order to compare the key semantic domains emerging from corpora representing pro- and counter-extremist texts. More recently, Conoscenti (2016) used

the corpus technique of collocation analysis in order to analyse the communicative strategies of *Dabiq*, while Baker and Vessey (2018) combined quantitative corpus methods with more qualitative, manual discourse analysis in their comparative study of the discursive themes and linguistic strategies employed in the aforementioned English-speaking *Inspire* and *Dabiq* magazines and ISIS's French-speaking magazine, *Dar al Islam*. Corpus methods have also been adopted in studies examining texts that can be viewed as more indirectly relating or contributing to extremist discourse, for example McEnery, McGlashan and Love's (2015) corpus-assisted discourse analysis comparing press and social media representations of the murder of Lee Rigby by two violent jihadists in London in 2013.

Despite a seeming surge in linguistic interest in the language associated with violent jihadism in recent years, to our knowledge such texts have yet to be subjected to register analysis. As such, we are currently unfamiliar with the register features of such texts and cannot be sure, in empirical terms at least, of the extent to which texts produced to incite jihadist extremist violence constitute a register distinct from those used in more general writing about Islam. Accordingly, we do not know the extent to which there are linguistically quantifiable differences between moderate, fringe and extreme texts about Islam which may break them into distinct registers. In this chapter, we set out to answer these questions and the next section outlines the data and methodological approach that we use to do it. That section will also present the definitions of *moderate*, *fringe* and *extreme* used in this chapter.

3 Methodology

The texts analysed in this chapter derive from an ongoing project examining jihadist discourse (see: Baker, Vessey and McEnery, 2021; Brookes and McEnery, 2020). The titles of texts associated with 11 successful terrorist prosecutions were supplied to us by contacts working in counter-extremism at the British Home Office or the London Metropolitan Police. The texts had been found on the hard drives of violent jihadists successfully prosecuted for terrorist offences in the UK. All the texts were in some way centrally con-

cerned with Islam. Based on the data provided to us, we were able to find the texts in question and build the corpus used in this study.

So far, we have referred to the texts in our data simply as 'extremist texts'. However, we can be more specific; experts in counter-terrorism research categorised these texts and made their categorizations available to us (see Holbrook (2015) for details of the coding of the texts). Based on their expert close reading of the texts, they classified each as either 'moderate', 'fringe' or 'extreme'. Moderate texts do not promote social isolation or violence, fringe texts do not overtly promote violence but they do promote alienation from mainstream society, while extremist texts openly advocate and facilitate violence. Table 1 provides a breakdown of the number of texts and words belonging to each of these categories.

Table 1: Corpus composition, by ideological category

Corpus	Texts	Words
Extreme	170	1,775,340
Fringe	54	486,650
Moderate	51	1,721,442

The extreme sub-corpus contains numerous articles from the aforementioned ISIS magazines, *Inspire* and *Dabiq*, as well as a variety of other texts including transcripts of interviews and lectures, biographies, political treatises, statements released by groups such as ISIS and Al Qaeda, how-to guides which contain advice on topics such as computer encryption, bomb-making and engaging in combat, and articles written in the style of news reports. The fringe texts contain more ambiguous messages which could be interpreted as advocating violence, but which do so implicitly and as such also have potential for non-violent readings. Such texts advocate withdrawal or segregation from civil society. Thus, while the extreme texts were likely crafted with the intention of advocating and enabling violence, the purpose of the fringe texts is less clear, but they could operate as 'gateway' texts designed to subtly nudge readers towards more extreme positions. Alternatively, we could interpret the

fringe texts as having been written strategically to advocate violence but in a way that evades the attention of legal authorities. Texts categorised as 'moderate' were those which referred to Islam in some way, but which were not categorised as 'extreme' or 'fringe'. Such texts typically involved scholarly discussion of religious topics in books and other sources. Some of these texts were written by people associated with terrorism or incitement to violence in other contexts. However, these texts did not incite violence or terrorism themselves, neither implicitly nor explicitly. The texts were judged by the experts rating them to advocate a more tolerant view of Islam based on co-operation and/or more peaceful co-existence with non-Muslims. Other texts in this category were written by non-Muslims, although they all had Islam as a central theme. Given that all of the owners of these texts spoke English as a first language, texts belonging to this final category are potentially useful for identifying aspects of language which are likely to be familiar to English-speaking Muslims, though they are not necessarily associated with extremist discourse. Nonetheless, we had to consider the possibility that such aspects might also be present in the 'extreme' texts too, perhaps because they are well-known to all Muslims but could also function as a means of making extremist language more familiar and persuasive for their intended Muslim readerships.

These categorisations are useful for our analytical purposes, as they help us to frame the texts in terms of ideology, given that central to each of the categories is a perceived attempt, on behalf of their authors, to propagate a particular worldview. For these purposes, we use the term *ideology* to describe 'the way in which what we say and think interacts with society', whereby ideology 'derives from the taken-for-granted assumptions, beliefs and value systems which are shared collectively by social groups' (Simpson 1993: 5). Similarly, at the level of reception, these texts were also perceived to have contributed to the formation of the worldviews of the individuals who had been radicalised to the extent that they had committed or planned to commit terrorist crimes in pursuit of the ideologies propagated by at least some of these texts. Although this categorisation was not carried out by linguists, and as such was based on non-linguistic characterisations of ideology, we are of the view that language is nevertheless likely to be significant to the ways in which those ideologies are expressed (ibid.). Our analysis therefore sets out to explore the extent to

which these categories, of 'moderate', 'fringe' and 'extreme', are *linguistically* meaningful, and to find out whether distinct forms of language might be associated with each of them in such a way that these categories align with distinct registers.

When assessing a corpus, a key consideration is representativeness. According to Biber (1993: 244), the representativeness of a corpus is determined by 'the extent to which a sample includes the full range of variability in a population'. This dataset can be described as representing language associated with violent jihad, since in all 11 cases the texts' possessors were convicted terrorists. Beyond this, a principled approach to corpus compilation might involve the use of a sampling frame, obtaining extremist texts which provide a balanced view of a range of known terrorist organisations, ideologies and different countries and languages. Our corpus does not follow such a sampling frame. Instead, all available texts from the context studied were included. The advantage of working with the texts that we have is that we know that they have been acquired, and presumably read, by people who have been legally judged to have become violent jihadists. The texts were written almost exclusively in English, reflecting the fact that English was at least one of the first languages of the people who were convicted, who had thus sourced and read the materials, although the texts also contained some evidence of code-switching to Arabic. As such, the corpus is maximally representative of the context studied – texts collected by and found in the possession of 11 convicted extremists.

Finally, due to its unusual nature, there are certain details about this corpus which cannot be shared with readers. Although some extremist texts, such as those published in the *Inspire* and *Dabiq* magazines, have been reported on by mainstream media, others are less widely known. To prevent other such publications from gaining public prominence, we will not provide a list of the names of the texts in our corpus; thus the usual claims pertaining to replicability in corpus linguistics research (see: Leech 1992) cannot apply in this case. Further details on corpus compilation and cleaning, including discussion of the legal and ethical issues attending to the use of extremist materials in linguistic research, can be found in Baker, Vessey and McEnery (2021).

Having introduced MDA in some depth in the previous section, we will not go too deeply into the workings of the technique here. However, on a

practical level, we analysed each of the corpora set out in Table 1 (respectively containing the 'extreme', 'fringe' and 'moderate' texts) using MDA. We began by using the 'Analyze Corpus' function of the *CQPweb* tool (Hardie 2012) to generate the frequency information necessary to undertake MDA. This frequency information then formed the basis of the MDA which was carried out using the freely available online *Lancaster Stats Tools*.[2] This analysis is reported in the next section.

4 Analysis

4.1. Initial findings: Registers

The MDA approach is useful in determining the extent to which the texts within each of the three classifications expressed some degree of linguistic homogeneity with regard to register. If we look at the three classifications and, on the basis of the MDA undertaken, ask the question, 'What registers do the texts in the corpus most closely resemble?', the answer should reveal differences and similarities between the classifications. Table 2 below shows, for each classification, which register is the closest match for the texts in that section of the corpus, using the aforementioned registers established by Biber (1988) for reference.

Table 2: Correspondence of each sub-corpus with the registers established by Biber (1988)

Register	Moderate		Fringe		Extreme	
	Count	Percentage	Count	Percentage	Count	Percentage
General narrative exposition	28	54.90	27	50.00	118	67.82
Imaginative narrative	0	0.00	1	1.85	2	1.15

2 See http://corpora.lancs.ac.uk/stats/toolbox.php. This website is best used in concert with Brezina (2018).

Involved persuasion	17	33.33	14	25.93	33	18.97
Learned exposition	6	11.76	6	11.11	10	5.75
Scientific exposition	0	0.00	6	11.11	11	6.32

Of note in this table is the overall difference between the moderate texts and the rest; moderate texts are most similar to general narratives, involved persuasion and learned exposition. Only the fringe and extreme texts approximate to imaginative narrative and scientific exposition. Similarly, as we move across the table another difference is apparent – the relative proportion of texts which are most similar to general narrative increase markedly in the extreme category. Overall, narrative is clearly important to all three categories – 54.90% of moderate, 51.85% of fringe and 68.97% of extreme texts are most similar to either general or imaginative narrative.

Unless readers are familiar with the system of analysis used by Biber, the results of it can appear to be somewhat abstract. However, the labels produced by the system are typically good guides to the content of the texts thus classified. Below is a brief example, drawn from the corpus of general narrative exposition, by way of illustration. It is an extract from the book, *My Life with the Taliban* (Zaeef, 2010).

> We were still waiting by the road when I saw the tanks coming, firing flares into the sky. Burning debris fell all around us, hitting cars here and there. They pointed their guns at the cars along the road, screaming at people like animals.
>
> This was the first time I had seen a convoy in Kandahar. It was very strange, and worrying. I asked my friend whether it was always this bad. 'Today was a good day', he said. 'This is our daily routine, and many times lives are lost when they pass through the city'.

The work in question is indeed a general narrative – an autobiographical account of life in the Taliban. The short passage quoted shows some of the key

features of narrative within the Biber model that places this text into the narrative category – for example, the use of third person pronouns (e.g. *he*), past tenses (*were*, *saw*, *fell*, *was*) and perfect aspects (*pointed*, *asked*). Across this text, such features appear in a preponderance such that the factor analysis used to carry out our analysis determines that the set of such features is so relatively frequent in the text that it is placed, on a continuum between non-narrative and narrative, on the narrative part of the continuum.

4.2 Initial findings: Authors

A possibility this raises is that, if we consider the author, publisher or organization linked to the text in question, we may find a distinctive style emerge – some authors may prefer a narrative style to learned exposition, for example. The dataset we are using allows us to approach that problem – while the majority of authors and organizations represented in the text typically provide but one text, we can look at 10 cases in the corpus where there are five or more texts from a single author or organization. The results of that analysis are shown in Table 3. In the Table, the register that each author produces most frequently is emboldened.

Table 3: Correspondence of texts produced by each author with the registers established by Biber (1988)

Register	ISIS	Awlaki	Bin Laden	Maqdiss	Yahya	Hizb	Azzam	Oadah	Muhajiroun	Deedat
General narrative exposition	**29**	3	**19**	**4**	2	**6**	**6**	2	2	**5**
Scientific exposition			3	2					**3**	
Imaginative narrative			1							

Involved persuasion		**6**	1	3	**6**	1	2	**4**		
Learned exposition	1	1			1	1			1	
Total	30	10	24	9	9	8	8	6	6	5

While the volume of data in this table is not substantial enough to allow meaningful statistical analysis, several points can be made, informally, about it. Firstly, General Narrative Exposition, the dominant register in Table 2, is the only register produced by all authors in Table 3. For most authors (6 in total), it is also the register that they produce most frequently. Of the authors who do not produce this register most frequently, three (Awlaki, Yahya and Oadah) produce Involved Persuasion most frequently, while Muhajiroun produce Scientific Exposition most frequently. ISIS stands at the opposite extreme to authors like Awlaki in that their documents are almost exclusively General Narrative Exposition. So, without examining at this point *why* different authors and organizations seem to make different choices about which registers to use, we can certainly see some evidence in our dataset that while General Narrative Exposition is indeed a very dominant register in the corpus in general and in extremist writing in particular, there are authors whose writing is represented in other styles and some of those are more prevalent for them in the corpus than General Narrative Exposition. It is very tempting, at this point, to ascribe choice to the author in this. However, while our data can allow us to say that for a specific document an author did draw upon a particular register, we cannot, for example, in the case of Awlaki, argue that he prefers an authorial style based upon Involved Persuasion. This is because we are studying what was collected by the violent jihadists, not what, in general, Awlaki wrote. If we were studying a comprehensive corpus of Awlaki's writings and found that he had an overwhelming preference for Involved Persuasion, we might begin to make a claim about what style he prefers. On the basis of the evidence we have, we might make this inference, but similarly we might make the inference that the violent jihadists, for some reason, prefer his writing in this register and that they downloaded it accordingly, ignoring other registers that he wrote in.

4.3 Initial findings: Mode of production

A quite different way of approaching the data would be to consider the different modes of production in the text; in other words, was the data written or a transcription of discourse that was originally spoken? This may explain some of the findings presented so far. For example, if we find that the Extremist material has within it more speech than writing and subsequently discover that the spoken data falls more commonly into the General Narrative category, we have an explanation, based on choice of mode of production, which could account for the apparent preference for this register in the extremist data.

In order to categorise our data, we were able to rely both on notes made by the analysts who had provided us with the titles that formed our corpus and our own readings of the text to determine when we were dealing with what claimed to be transcribed speech. Of course, we were not able to determine whether the speech was spontaneous or if it was text which had been spoken, but this is precisely the type of distinction that we could expect the MDA to identify, should it be present in the data, as shown in previous studies (Biber 1984). Applying the perspective of mode of production split the data into 237 written texts and 38 transcribed spoken texts. The initial coding of the texts as written and spoken revealed a marked skew in the data, as Table 4 shows.

Table 4: Number of texts in the corpora, categorised by mode (percentages in brackets)

	Written	Spoken
Moderate	51 (100%)	0 (0%)
Fringe	49 (90.74%)	5 (9.26%)
Extreme	137 (80.58%)	33 (19.42%)

So, the hypothesis that the data may, in fact, be prone to skew depending on mode of production does seem worth investigating. In order to do so, the data was subject to MDA once again, this time with the texts categorised as speech and writing alone – if the spoken language across all categories (moderate, fringe, extreme) distributes differently to the written language, this would give

some initial indication that any difference in register preferences of the three categories (moderate, fringe, extreme) may be a consequence of there being different proportions of speech and writing across the categories. The results of placing the spoken and written data on Biber's six dimensions of variation are given in Table 5.

Table 5: Variation in the corpus along Biber's dimensions of variation

Dimension	Results[3]
1. Involved v. Informational	$F = 2.2$; $p = 0.1353808$, $r^2 = 0.8\%$
2. Narrative v. Non-Narrative	$F = 5$; $p = 0.02569213$, $r^2 = 1.8\%$
3. Elaborated v. Situation-Dependent	$F = 1.5$; $p = 0.2286787$, $r^2 = 0.5\%$
4. Overt Expression of Argumentation	$F = 2.9$; $p = 0.09232212$, $r^2 = 1\%$
5. Abstract v. Non-Abstract	$F = 10.4$; $p = 0.001425511$, $r^2 = 3.7\%$
6. Online Informational Elaboration	$F = 2.1$; $p = 0.1524262$, $r^2 = 0.7\%$

These results are not encouraging for this hypothesis. Only one of the results, Abstract v. Non-Abstract, has scores which would lead us to have a reasonable degree of confidence that we are seeing a difference worth reporting. The result certainly looks significant in statistical terms, but in descriptive terms it is not significant – the r^2 score suggests that what we are looking at along this dimension accounts for only 3.7% of variation between the spoken and written texts. So, if we have found a difference, it is a small one. Otherwise, speech and writing in the corpus appear very similar. This is suggestive of the speech in the corpus being scripted or, perhaps in the case of interviews in newspapers, heavily edited before publication. The example below, a part of a transcription of an interview with Ayman Zawahiri, is a good example of this:

3 In the results column of this table, and each table following, F is a measure of variation between means for values taken from the texts analysed, the higher the value, the more scattered from the mean the actual values analysed are. The significance value p gives the confidence we may have of the result not being the product of chance (where a value of 0.05 is at the 95% confidence level, 0.01 is at the 99% confidence level, for example) and r^2 is a measure that shows how much of the variation observed may be accounted for by the dimension in question.

INTERVIEWER: We are happy to interview you four years after the New York and Washington raids [of 9/11].

DR. AYMAN AL-ZAWAHIRI: I, too, am happy to address our Muslim umma through you during this critical stage of its history. I would like to take this opportunity to thank you, and pray that Allah Almighty will reward you for publicizing the word of truth in the midst of the Crusader campaign and global war that is being waged against Islam and the Muslims.

INTERVIEWER: Dr. Ayman, how do you view the Crusader campaign, four years after it began?

DR. AYMAN AL-ZAWAHIRI: The new Crusader campaign is failing, just like the previous ones, by the grace of Allah. America and its Crusader allies have not accomplished a single thing-except for throwing their armies into the battlefield to take blows on a daily basis, to have their soldiers killed on a daily basis, and to have their economies bled on a daily basis.

What did they accomplish in Afghanistan? They evicted the Taliban government from Kabul, but it centered itself in the villages and mountains-where the real power of Afghanistan lies. Northern Afghanistan and Kabul have become a scene of chaos, pillaging, looting, defiling [women's] honor, and drug trafficking, which have flourished and thrived under the American occupation. Then they held elections, which resembled a masquerade more than anything else- since the country's periphery is controlled by highway bandits and warlords, and since the international committees monitoring the elections - or rather, those who bear false witness - could not (even if they really wanted to) cover more than ten voting districts; and since transferring the ballot boxes takes fifteen days, under the control of the warlords and highway bandits, and then under the control of the occupation forces; and since any resistance, or anything resembling resistance or opposition, is met with bombardment, missiles, the burning of villages, and the killing of hundreds.

> Then, after all this, they obtained the false testimony of the U.N. [United Nations], which saw nothing for it to bear witness about-except for a few theatrics in some [voting] districts in the cities. This is but one of many examples of the U.N:s hypocrisy, which they claim to be the symbol of their international legitimacy.

Brief as this example is, it is notable in lacking any of the markers of spoken interaction that one might intuitively expect – there are no markers of hesitation, no repairs and the syntax is somewhat complex (consider the coordinated infinitival clauses in the final sentence of the example). In terms of Biber's analytical scheme, we might expect informational interaction to have positive scores on the involved vs. informational discourse dimension (for example, Nini, 2019). Yet the features we would expect for the text to be placed at the involved part of the continuum are largely absent – for example, *that*-deletion ('pray that Allah'), present tenses (this example starts with present tenses, but shifts to the past and, in the fuller text, the past tenses are dominant) and second-person pronouns (the first- and third-person predominate in the pronouns used in this quote).

4.4 Exploring combinations and new dimensions

It may, of course, be the case that the way in which speech and writing is used across moderate, fringe and extreme texts is not smoothly distributed, as the previous analysis presumed. So, what would happen if we looked at the six categories created by combining speech and writing with moderate, fringe and extreme? The results of undertaking that analysis according to Biber's six dimensions are given in Table 6.

Once again, the results seem to suggest that variation within the dataset is low. Only one of these results is statistically significant at the 99% level (Narrative v. Non-Narrative) but once again the volume of variation explained by this factor is low (6.9%), as we might expect given that we know there is a preponderance of general narrative material in all categories in our data. Abstract v. Non-Abstract is significant at the 95% level but explains only 3.9% of variation. Figure 1 shows the results for the Narrative v. Non-Narrative dimension compared with the findings of Biber (1988).

Table 6: An MDA of moderate, fringe and extreme speech and writing using Biber's six dimensions

Dimension	Results
1. Involved v. Informational	F = 1; p = 0.4325968, r^2 = 1.4%
2. Narrative v. Non-Narrative	F = 5; p = 0.0006705594, r^2 = 6.9%
3. Elaborated v. Situation-Dependent	F = 0.8; p = 0.553918, r^2 = 1.1%
4. Overt Expression of Argumentation	F = 1.5; p = 0.2010924, r^2 = 2.2%
5. Abstract v. Non-Abstract	F = 2.7; p = 0.02957478, r^2 = 3.9%
6. Online Informational Elaboration	F = 2.1; p = 0.0875549, r^2 = 2.9%

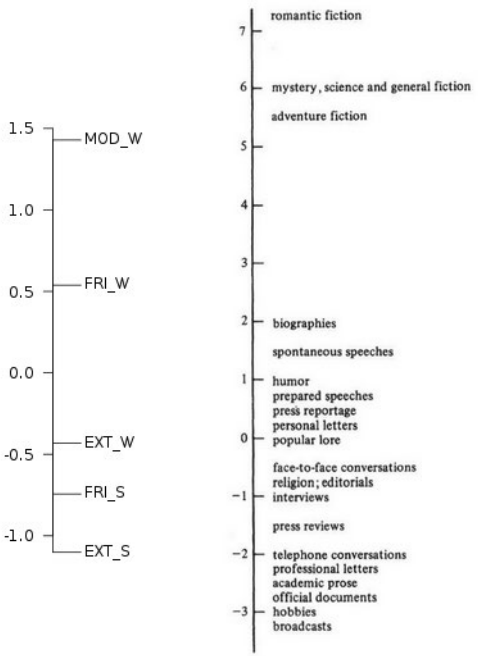


Figure 1: Moderate, fringe and extreme speech and writing on the Narrative v. Non-Narrative dimension (left) compared to Biber's findings for this dimension (right)

The dimensions are very densely grouped, relative to the scale of variation that Biber observed, and the data in our corpus displays a much lower degree of variation on both of these dimensions, clustering around 0. For Narrative, we may say that fringe and moderate writing exhibit a slight preference for this relative to extreme writing, fringe speech and extreme speech. However, relative to some of the forms of speech and writing that Biber studied, none of the data explored has as strong a preference for Narrative as Biographies or Fiction. Similarly, none of them are as non-narratively oriented as he found telephone conversations, professional letters, academic prose, official documents, writing about hobbies and broadcasts to be.

One further way in which we may seek some principled difference between the files in the data that might align with register variation would be to look at the type of texts in the corpus – might they vary linguistically from one another in some systematic way? The notes provided to us by the analysts with each file suggested a categorization scheme for the corpus texts, as did the texts themselves. The analysts would often give labels to a text such as 'statement' or 'poem' that would lead one to assume that, if used as a way of categorizing the texts, some register variation may be visible. Similarly, the texts themselves very often self-labelled, for example as a poem or a magazine (article). Working from these notes, and based on a supporting close reading of the texts, we allocated each text to one of the categories given in Table 7. Texts were allocated to a category based on either their form (e.g. book, magazine, poem) or dominant function. Note that the Book and Magazine categories are ones which are functionally mixed.

Table 7: A functional classification of the corpus texts

Category	Description	Number of texts in this category
Argument	A disputation in which a proposition is argued through, leading to a conclusion.	83
Book	Lengthy prose in the form of a book. This may be a composite of a number of other categories.	53
Forum	An exchange in an online forum.	3

Interview	A Q&A Interview – these appear to be transcriptions of real interviews as opposed to the use of a Q&A structure as a rhetorical device in an Argument.	11
Lecture	Transcribed lecture/sermon.	10
Magazine	Prose in the form of a magazine. This may be a composite of a number of other categories.	34
Poem	A text composed almost exclusively of a poem or poetry.	2
Statement	A statement of what the author claims to be i.) fact or ii.) an authoritative point of view. Such statements can cover narrative prose claiming to be a true account of a particular set of events.	79

The classification in Table 7 is, of course, preliminary. However, given that such features were largely derived from the work of the analysts who were familiar with the texts and that a categorization of the texts using these categories is possible, using the framework seems warranted. Our aim in undertaking the classification was to see whether such an approach to the texts was fruitful. If it was, then this could call for a closer exploration of the categories and, in particular, for a disaggregation of the texts in the Book and Magazine categories into functional units. Table 8 gives the results of the MDA of these categories.

Table 8. An MDA of the texts, catagorized by function

Dimension	Results
1. Involved v. Informational	$F = 1.7$; $p = 0.1116541$, $r^2 = 4.2\%$
2. Narrative v. Non-Narrative	$F = 5.3$; $p = 1.168323E\text{-}5$, $r^2 = 12.1\%$
3. Elaborated v. Situation-Dependent	$F = 0.7$; $p = 0.7117613$, $r^2 = 1.7\%$
4. Overt Expression of Argumentation	$F = 1.9$; $p = 0.07288708$, $r^2 = 4.7\%$
5. Abstract v. Non-Abstract	$F = 3$; $p = 0.004809807$, $r^2 = 7.3\%$
6. Online Informational Elaboration	$F = 3.1$; $p = 0.003701417$, $r^2 = 7.5\%$

At first glance, these results seem much more promising – the last three dimensions seem to have significant results which explain between 4.7% and 7.5% of the variation observed. However, as Figure 2 shows, the poetry (Figures 2, 3 and 4) and forum discussions (Figures 3 and 4) seem to account for these results. Given the very low counts of poems and forum discussions in the corpus, it would be hard to conclude that these are useful results – they are certainly not very helpful in terms of characterizing the dataset overall, given that there are only five texts in the corpus belonging to these categories.

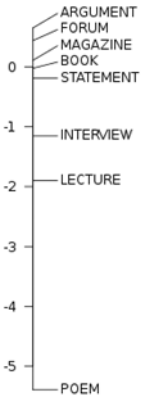


Figure 2: Dimension 4, text types only

Figure 3: Dimension 5, text types only

Figure 4: Dimension 6, text types only

Might it be that by combining the mode of production categorization with the text category we may be able to gain some insight into variation in the texts that has eluded our analysis so far? The results of the MDA this produces are given in Table 9.

Table 9: An MDA of text types categorized by mode of production

Dimension	Results
1. Involved v. Informational	F = 2.1; p = 0.0295375, r^2 = 6.7%
2. Narrative v. Non-Narrative	F = 4.5; p = 1.444902E-5, r^2 = 13.3%
3. Elaborated v. Situation-Dependent	F = 0.8; p = 0.6315774, r^2 = 2.6%
4. Overt Expression of Argumentation	F = 1.5; p = 0.1631546, r^2 = 4.7%
5. Abstract v. Non-Abstract	F = 2.6; p = 0.0075517, r^2 = 8%
6. Online Informational Elaboration	F = 3.2; p = 0.001162657, r^2 = 9.7%

Once again, Abstract v. Non-Abstract and Online Informational Elaboration seem to give results worthy of investigation, this time with higher r^2 values than observed previously. However, as Figures 5 and 6 show, data sparsity once again accounts for the results. Poetry accounts for the outlier in both figures, while Arguments produced in speech represent another outlier. However, the corpus has only two texts in this category.

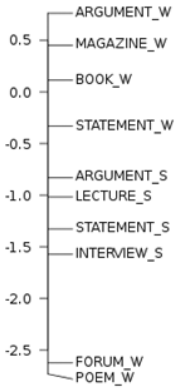


Figure 5: Dimension 5, text category plus mode combined

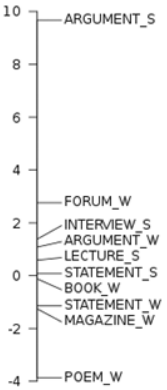


Figure 6: Dimension 6, text category plus mode combined

Of course, the possibility exists that Biber's (1988) dimensions are not the ones that would best characterise the differences in this dataset. Might it be that, within the dataset we are observing, the values identified by Biber configure in a novel way meaning that we need to build dimensions of variation, afresh, from the bottom up? To explore this, we started with a scree plot of Eigen values of principal factors (for details, see Brezina (2018)) as a way of determining the number of clusters we might search for (see Figure 7). We then calculated 7 factors on the basis of this analysis.

Once again, however, the factors do not reveal a substantial difference between the speech and writing, as Table 10 shows. None of the results are statistically significant, even at the 95% level, and the values of r^2 are very low indeed.

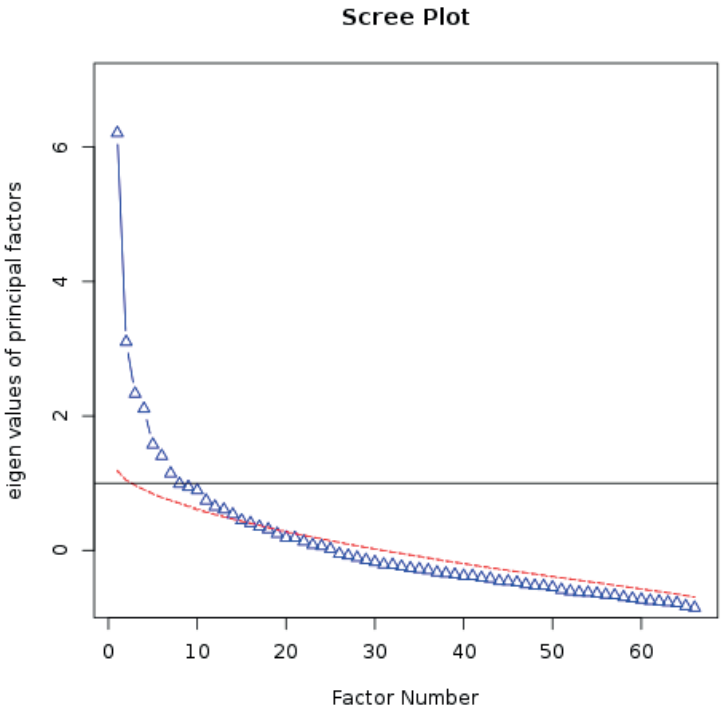


Figure 7: A plot of Eigen values of principal factors comparing the spoken and written sections of the corpus.

Table 10: A seven factor MDA

Dimension	Results
1	F (df = 1, 273) = 1.07; p = 0.3008447, r^2 = 0.39%
2	F (df = 1, 273) = 1.45; p = 0.2289638, r^2 = 0.53%
3	F (df = 1, 273) = 0.77; p = 0.3805662, r^2 = 0.28%
4	F (df = 1, 273) = 5.34; p = 0.02154485, r^2 = 1.92%
5	F (df = 1, 273) = 2.38; p = 0.1240777, r^2 = 0.86%
6	F (df = 1, 273) = 0.04; p = 0.8472702, r^2 = 0.01%
7	F (df = 1, 273) = 0.67; p = 0.4136613, r^2 = 0.24%

5 Conclusions

Our MDA strongly suggests the following about our corpus of violent Jihadist texts. Firstly, at least as measured by the features underlying Biber's original study, the texts in our corpus are generally quite homogenous linguistically. When factors are combined and the texts are looked at from different analytical perspectives, they tend to remain stubbornly grouped. Secondly, there are limits on how far the perspectives on the texts can be pushed. In principle, given the categorizations that we have for the texts, we could combine mode of production (two categories) with analysts' ratings (three categories) and text types (eight categories) to produce 48 categories to which we could assign our texts in order to explore variation. However, as has already been shown, the text collection does not support this analysis. Even when we combine mode and analyst rating, we produce an empty category – there is no moderate spoken material in the dataset. Even where there are members of a category, these may be few and they may lead to results which are unhelpful and do not characterise the data in general as we saw with Forum, Spoken Arguments, Discussion and Poems. Finally, however, we should be mindful of the useful, though admittedly weak, signals that the analyses have produced: i.) the nature of the texts in the corpus is generally narrative; ii.) the type of narrative may vary by analysts' category; iii.) narrative v. non-narrative may be worth exploring on the basis of the results presented in Tables 4, 6 and 8; and iv.) abstract v. non-abstract may be worth investigating based on the results presented in tables 4, 6, 8 and 9. These all represent potentially interesting avenues for future study.

The results of the MDA investigations give us confidence that our corpus represents a linguistically coherent body of texts in the sense that the texts seem to be similar. Not only were they based on a similar topic and produced and read by people with similar goals and who, in all likelihood, shared a similar worldview, they are also linguistically similar to the extent that we can say that the corpus is composed of what appears to be a single register of what we might call Islamic discourse, which arches across the categories of moderate, fringe and extreme. On the face of it, this lack of variation could be viewed as something of a non-result. However, the insights into the homogeneous

nature of these texts have implications for researchers interested in studying such texts in the future. For such researchers, this finding suggests that significant linguistic variation identified within sets of extremist texts similar to our own is unlikely to result from differences at the level of register. Indeed, our findings provide a warrant for regarding such texts as a single register for analytical purposes. If variation is to be sought, it is more likely to reside in factors other than those studied here, for example the time period or variety of English in which the texts were composed.

On a methodological note, our findings also suggest the utility of MDA for studying similarity across texts. By and large, MDA has tended to be applied in previous research to identify variation within and across datasets, with emergent differences often forming the focus of the study and constituting headline findings. However, the ability of MDA to reliably reveal similarities, as highlighted by this study, points to an application of the method in studies seeking to study overlaps in, *inter alia*, certain ideologies, values, and worldviews that are propagated by the language used in and across sets of texts. Indeed, what binds the homogeneous set of texts in our data is just this: they index a register which seems to be related to Islam. Such an approach could be applied in corpus-assisted discourse studies (see: Baker et al. 2008) as an entry point for grouping texts according to their ideological stances and for down-sampling texts and linguistic features for more qualitative analysis of the linguistic manifestations of those ideologies.

A limitation of our findings, suggested above, pertains to the issue of representativeness. As noted, the corpus analysed in this chapter was not assembled to provide a general assessment of writing about Islam. Our texts were associated with a subset of Muslims who engaged in terrorist acts. A full assessment of the representativeness of a corpus depends on a prior full definition of the 'population' that the sample comprising the corpus is designed to represent (Biber 1993). The types of texts that we have worked with here thus present a challenge on this front, as the wider 'population' of texts about Islam is both ill-defined and unknown. This makes a full assessment of the representativeness of this corpus in terms of general texts about Islam impossible. As such, we cannot be sure of the extent to which the texts of which we were availed represent the full set of texts about Islam which exist 'out there' in

the world. That notwithstanding, the general homogeneity of the texts that we have analysed could provide some evidence that this is a relatively restricted linguistic register, and provides a starting point for future research of such texts. This point is important in two respects. Firstly, it is useful to advise people where not to look for difference – approaching these texts to look at register differences, based upon the features outlined originally by Biber, is not productive. The differences between the texts is not rooted in register. However, this brings us to a second way in which these results are useful – they indicate the importance of the register used to those who produced the texts in our corpus. This point becomes more important when we consider other work which we have undertaken where we have found more differences between the texts. For example, in Brookes and McEnery (2020) our argument is that many of the changes that are observed are attempts at challenging the doxa of Islam – they are part of an ideological struggle for the meaning of the religion. An important feature of that struggle is what is *not* fought over – the register in which the arguments are presented. This constitutes an important form of symbolic capital (Bourdieu, 1989) which is indispensable to all of the authors in the corpus. In some ways, therefore, we might argue that the inflexibility of the register used in the corpus examined in this chapter is a signal of the importance that is attached to this register as a form of symbolic capital that allows the arguments presented by writers in the fringe and extreme categories to be accepted. In this sense, the findings presented here are critical as they reveal the importance of adherence to register even by those wishing otherwise to subvert and revise belief.

References

Anon (2013): Restricted Report. Home Office, UK.

Baker, P./Vessey, R. (2018): A corpus-driven comparison of English and French Islamist extremist texts. *International Journal of Corpus Linguistics*, 23(3), 255–278.

Baker, P./Gabrielatos, C./KhosraviNik, M./Krzyżanowski/M., McEnery, T./Wodak, R. (2008): A useful methodological synergy? Combining critical discourse

analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.

Baker, P./Gabrielatos, C./McEnery, T. (2013): *Discourse Analysis and Media Attitudes: The representation of Islam in the British Press*. Cambridge: Cambridge University Press.

Baker, P./Vessey, R./McEnery, T. (2021): *The Language of Violent Jihad*. Cambridge: Cambridge University Press.

Berber Sardinha, T. (2018): Dimensions of variation across Internet registers. *International Journal of Corpus Linguistics*, 23(2), 125–157.

Bértoli-Dutra, P. (2014): 'Multi-Dimensional analysis of pop songs'. In: T. Berber Sardinha/M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber*. Amsterdam: John Benjamins, pp. 149–176.

Biber, D. (1984): A model of textual relations within the written and spoken modes. Ph.D. Dissertation, University of Southern California.

Biber, D. (1986): Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 62, 384–414.

Biber, D. (1988): *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1993): Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.

Biber, D. (2006): *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. (2019): 'Multi-dimensional analysis: a historical synopsis'. In: T. Berber Sardinha and M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury, pp. 11–26.

Biber, D./Conrad, S. (2009): *Register, Genre and Style*. Cambridge: Cambridge University Press.

Biber, D./Conrad, S./Reppen, R. (1996): Corpus-based investigations of language use. *Annual Review of Applied Linguistics*, 16, 115–136.

Bourdieu, P. (1989): Social space and symbolic power. *Sociological Theory*, 71(1), 14–25

Brezina, V. (2018): *Statistics in Corpus Linguistics: A practical guide*. Cambridge: Cambridge University Press.

Brindle, A. (2017): *The Language of Hate: A Corpus Linguistic Analysis of White Supremacist Language*. London and New York: Routledge.

Brookes, G./McEnery, T. (2020): Correlation, collocation and cohesion: A corpus-based critical analysis of violent jihadist discourse. *Discourse & Society*, 31(4), 351–373.

Clarke, I./Grieve, J. (2019): Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. PLOS ONE, Online. Available here: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0222062.

Conoscenti, M. (2016): 'ISIS' Dabiqcommunicative strategies, NATO and Europe. Who is learning from whom?' In M. Ceretta and B. Curli (Eds.), *Discourses and Counter discourses on Europe: From the Enlightenment to the EU*. London and New York: Routledge, pp. 215–244.

Conrad, S./Biber, D. (2001): 'Multidimensional methodology and the dimensions of register variation in English'. In: S. Conrad and D. Biber, (Eds.), *Variation in English: Multidimensional studies*. Pearson Education: Harlow, pp. 13–42.

Droogan, J./Peattie, S. (2018): Reading jihad: Mapping the shifting themes of Inspire magazine. *Terrorism and Political Violence*, 30(4), 684–717.

Esposito, J. L. (2014): *The Oxford Dictionary of Islam*. Oxford: Oxford University Press.

Friginal, E. (2008): *The Language of Outsourced Call Centers: A corpus-based study of cross-cultural interaction*. Amsterdam: John Benjamins.

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

HM Government (2013): Tackling extremism in the UK. London: Cabinet Office. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/263181/ETF_FINAL.pdf (last accessed August 2018).

HM Government (2015): *Counter-Extremism Strategy*. London: Cabinet Office. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/atachmnt_data/file/470088/51859_Cm9148_Accessible.pdf.

Holbrook, D. (2015): Designing and applying an extremist media index. *Perspectives on Terrorism*, 9(5). Online. URL: http://www.terrorismanalysts.com/pt/index.php/pot/article/view/461/html.

Ingram, H. J. (2017): An Analysis of Inspire and Dabiq: Lessons from AQAP and Islamic State's Propaganda War. *Studies in Conflict & Terrorism*, 40(5), 357–375.

Kramer, M. (2003): Coming to terms, Fundamentalists or Islamists? *Middle East Quarterly*, Spring 2003, 65–77.

Leech, G. (1992): 'Corpora and theories of linguistic performance'. In: J. Svartvik (ed.), *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin: Walter de Gruyter, pp. 105–122.

McEnery, T./Hardie, A. (2012): *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Nini, A. (2019). 'The Multi-Dimensional Analysis Tagger'. In: T. Berber Sardinha/ M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*. London and New York: Bloomsbury, pp. 67–94.

Prentice, S./Taylor, P./Rayson, P./Hoskins, A./O'Loughlin, B. (2011): Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the Gaza conflict. *Information Systems Frontiers*, 13(1), 61–73.

Prentice, S./Taylor, P./Rayson, R./Giebels, E. (2012): Differentiating act from ideology: Evidence from messages for and against violent extremism. *Negotiation and Conflict Management Research*, 5(3), 289–306.

Quaglio, P. (2009): *Television Dialogue: The sitcom Friends vs. natural conversation*. Amsterdam: John Benjamins.

Rayson, P. (2008): From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519–549.

Richardson, J. E. (2017): *British Fascism: A Discourse-Historic Analysis*. Hannover: ibidem Verlag.

Simpson, P. (1993): *Language, Ideology and Point of View*. London and New York: Routledge.

Staples, S. (2015): *The discourse of nurse-patient interactions: Contrasting the communicative styles of U.S. and international nurses*. Amsterdam: John Benjamins.

Wignell, P./Tan, S./O'Halloran, K. L. (2017): Under the shade of AK47s: a multimodal approach to violent extremist recruitment strategies for foreign fighters. *Critical Studies on Terrorism*, 10(3), 429–452.

Zaeef, A.S. (2010): *My Life with the Taliban*. New York: Columbia University Press.

Gerhard Rampl und Claudia Posch

Methodisch geht es korpuswärts! Zur Produktivität des Suffixes *-wärts*

1 Einleitung

Eine Tätigkeit wie das Bergsteigen ist, wie Bubenhofer/Rothenhäusler treffend bemerken, geprägt von bestimmten Abfolgen und Handlungsmustern z.B. der physikalischen Notwendigkeit unten zu starten, sich nach oben zu bewegen und dann wieder nach unten zu kommen: „Die Anatomie einer Bergtour scheint prototypisch also ein Unten-Hoch-Gipfel-Runter-Unten zu sein, das einer topographisch erzwungenen Sequenz folgt“ (2018: 39). Man kann also auch erwarten, dass sich in Texten, die sich mit dieser Tätigkeit befassen, sprachliche Abfolgen und Muster wiederfinden, die eine solche Bergtouren-Anatomie widerspiegeln. Dieser Gedanke bildet den Ausgangspunkt für unsere Überlegungen in diesem Beitrag. Wir gehen zunächst davon aus, dass in alpinbezogenen Texten Raum- oder Direktionaladverbien eine größere Rolle spielen als in anderen Textsorten. Dies bestätigt sich in der unterschiedlichen relativen Häufigkeit, wie in 4 näher beschrieben wird.

Um den Untersuchungsgegenstand weiter einzuschränken, führten wir zwei Keyword-Analysen im thematisch eng fokussierten Korpus Alpenwort - Korpus der Zeitschrift des österreichischen Alpenvereins durch, das von uns kontinuierlich aufgebaut und versioniert publiziert wird. Alpenwort enthält 19,9 Millionen Wörter und umfasst derzeit alle Bände der als Jahrbuch erscheinenden Zeitschrift von 1870–2010. Die Daten wurden umfangreich korrigiert und vollständig mit dem STTS-Tagset annotiert (für eine genaue Beschreibung der Korpus-Genese siehe Posch/Rampl 2020) und sind über die Plattform CQPweb für Analysen frei zugänglich (Rampl/Posch 2019). Das Korpus ist so aufgebaut, dass sich über den gesamten Zeitraum der Publikation der Zeitschrift diachrone Analysen anstellen lassen. Enthalten sind unterschiedlichste Textsorten, von wissenschaftlichen Artikeln, die eher in den

frühen Publikationen eine größere Rolle spielen, über zahlreiche Besteigungs- und Expeditionsberichte bis zu literarischen Texten und kleineren Textsorten. Allen gemein ist das Interesse an den Themen „Berg“ und „Bergsteigen“.

Dabei zeigte sich die interessante Entdeckung, dass mit dem Suffix *-wärts* gebildete Direktionaladverbien nicht gleichmäßig im Korpus verteilt sind. Aus dieser Entdeckung entstand die leitende Untersuchungsfrage des vorliegenden Beitrags, nämlich wie sich dieses Suffix *-wärts* diachron entwickelt und ob sich diese Entwicklung in verschiedenen Korpora und Textsorten gleich niederschlägt.

2 Keyness der Direktionaladverbien in Alpenwort

Um die eingangs gestellte Frage, welche Raum- und Direktionaladverbien in alpinbezogenen Texten wichtig sind, beantworten zu können, führten wir Keyword-Analysen durch. Dadurch konnten wir zu einer für die Untersuchung einerseits handhabbaren und andererseits auch relevanten Teilmenge der Raum- und Direktionaladverbien gelangen. Gabrielatos folgend betrachten wir *keyness* und die Keyword-Methode als „Weg in Texte hinein (2018: 227)“ , also als eine Technik, um für linguistische Fragestellungen interessante Items überhaupt erst zu finden. In einem ersten Schritt wurden hierfür auf zwei unterschiedliche Arten *candidate key items* (CKI) berechnet, um für die weiterführenden Forschungsfragen eine Auswahl treffen zu können.

Keyitems 1: Für die erste Berechnung wurden Subkorpora der Jahrzehnte 1890 (Untersuchungskorpus, 153 Texte 1.903.643 Wörter) und 1990 (Referenzkorpus, 336 Texte, 1.687.216 Wörter) verwendet. Beide Subkorpora haben eine ähnliche Größe und sind sich auch inhaltlich ähnlich. Folgende Einstellungen in CQPweb wurden ausgewählt:

- Statistik: LogLikelihood,
- Signifikanz Cut-off: 0.01% (LL threshold = 15.14),
- Minimalfrequenz in beiden Subkorpora: 3.

Bei der manuellen Sichtung der CKI-Liste weckten schließlich einige zusammengesetzte Direktionaladverbien mit dem bisher selten untersuchten Suffix *-wärts* (6 von 46 Direktionaladverbien) unser Interesse.

Rang	Wort	Statistik
73	aufwärts	287.29
111	abwärts	20.03
399	nordwärts	72.00
478	ostwärts	61.46
715	rückwärts	40.31
753	vorwärts	38.30
988	südwärts	28.04
1022	seitwärts	26.93
1090	westwärts	25.17

Tabelle 1: CKIs aus der Keyword-Liste

LogLikelihood (LL) gilt wohl als das traditionelle Keywordmaß und findet in zahlreichen Studien Verwendung, obwohl es für die Errechnung von Keywords inzwischen nicht mehr uneingeschränkt empfohlen wird (für eine Kritik siehe u.a. Gabrielatos 2018: 234; Pojanapunya/Watson 2016: 162; Brezina/Meyerhoff 2014: 23). Wir können aus der obigen Berechnung vorerst lediglich schließen, dass es wohl einen Unterschied in der Häufigkeit der in Tabelle 1 gelisteten Direktionaladverbien mit *-wärts* zwischen den Dekaden 1890 und 1990 gibt, der statistisch signifikant erscheint ($p<0.0001$). Das heißt im Untersuchungskorpus kommen diese Items also signifikant häufiger vor als im Referenzkorpus. Ein bekanntes Problem mit LL ist, dass auch kleine Unterschiede statistische Signifikanz erlangen können und so generell eher zu viele Keyitems berechnet werden. Dies spielt hier jedoch eine untergeordnete Rolle, weil wir die CKIs nutzen, um unser Untersuchungsinteresse eingrenzen zu können.

Um nunmehr zu überprüfen, ob sich ebenfalls ein messbarer Unterschied von Direktionaladverbien mit *-wärts* über den Verlauf der gesamten Zeitspanne zeigt, wurde folgende Strategie verfolgt (vgl. auch Posch in Vorbereitung):

Keyitems 2: Die einzelnen Dekaden wurden jeweils im Vergleich zum gesamten restlichen Korpus mittels eines Effektstärkenmaßes verglichen. Wir wählten dafür ebenfalls die Standardeinstellungen von CQPweb:

- Statistik: LogRatio mit LogLikelihood Filter (angepasste LL Schwelle = 37.52, mit Šidák Korrektur für Mehrfachvergleiche),
- Signifikanz Cut-off: 0,01%,
- Minimalfrequenz in beiden Subkorpora: 3.

Die resultierenden CKIs kommen somit signifikant bis höchst signifikant häufiger in den jeweiligen Untersuchungskorpora (Dekaden) im Vergleich zum restlichen Korpus vor. Auch negative Keyitems wurden diesmal berechnet (i.e. signifikant seltenere Items).

In dieser Berechnung ist das Referenzkorpus jeweils erheblich größer als das Untersuchungskorpus (Brezina 2018: 81), aber aufgrund der thematischen Enge des Korpus sind Referenz-/Untersuchungskorpus inhaltlich sehr ähnlich. Mit dem Ergebnis dieser Berechnung lässt sich nunmehr ein chronologischer Vergleich der CKIs pro Dekade darstellen. Bei der Sichtung zeigt sich, dass mehrere Zusammensetzungen mit *-wärts* in mehreren Dekaden eine signifikante Keyness aufweisen. Die LR-Werte bewegen sich zwar auf einem eher niedrigen Niveau (von +2,83 bis -2,03), wohl weil die Bildungen insgesamt nicht sehr häufig sind, aber die gefundenen Zusammensetzungen mit *-wärts* sind bis zu ca. 8 Mal häufiger bzw. bis zu ca. 4 Mal seltener (bei *negative keyness*) im Untersuchungskorpus anzutreffen als im Referenzkorpus (siehe Abbildung 1).

Es zeigt sich die interessante Entwicklung, dass Direktionaladverbien mit *-wärts* in den anfänglichen Jahrzehnten als positive Keywords auffallen und im Laufe des 20. Jahrhunderts im Korpus Alpenwort deutlich an Bedeutung verlieren. Wir erweitern also im nächsten Schritt den Fokus generell auf das Suffix *-wärts* und möchten dessen Status und Entwicklung nun näher beleuchten.

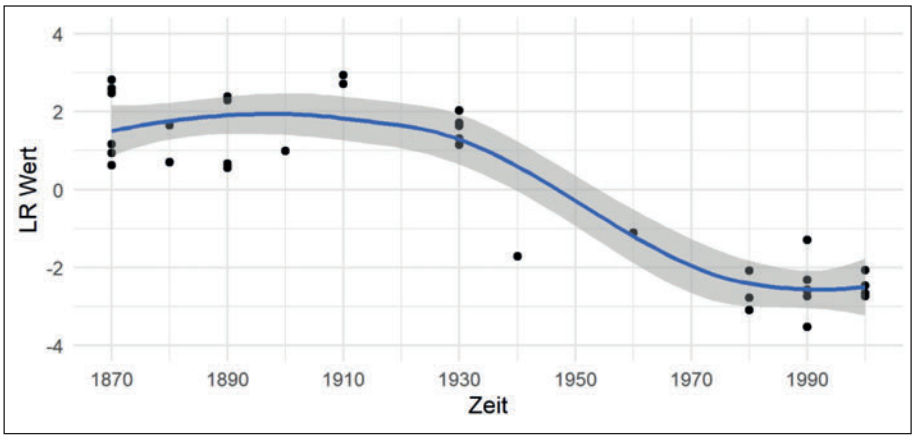


Abbildung 1: Entwicklung der Keyness von -„wärts“ diachron (stat_smooth=loess)

3 Das Adverbialsuffix *-wärts*

Aus synchroner Sicht wird das Ableitungsmorphem *-wärts* als Mittel zur Bildung von Lokal- bzw. Direktionaladverbien der Richtung betrachtet, wobei eine übertragene Verwendungsweise ebenfalls möglich ist. Ein zentraler Punkt dieser Untersuchung wird sein, ob man *-wärts* wirklich noch als produktiv betrachten kann. Bevor wir näher auf diese Frage eingehen, wollen wir einen kurzen Überblick über den derzeitigen Forschungsstand zu Entwicklung und Status (Suffix vs. Suffixoid, Produktivität) geben.

Historisch betrachtet lässt sich *-wärts* auf ahd. *-wert* bzw mhd. *-wert* (Kluge 2012: s.v. *-wärts*) zurückführen. Es ist verwandt mit dem heutigen Hilfszeitwort *werden*, wobei sich der semantische Kern schon im gemeinsamen indoeuropäischen Stamm **wert-* ‘wenden’ also ‘sich in (die betreffende) Richtung wendend’ (ibd.) finden lässt. Bereits im Grimmschen Wörterbuch wird konstatiert, dass *wärts* nur noch „in fester verbindung mit adverbien und nominibus“ (Grimm, s.v. *wärts*, adv.) vorkomme, früher aber selbständig gewesen sei. Eine detaillierte Betrachtung des Suffixes bietet Graën in seiner Dissertation zu Raumadverbien (Graën 2004: 147–151). Er zeigt auf, wie das ursprünglich eigenständige Direktionaladverb *werd*, *ward* mit der Bedeutung ‘die Richtung

habend' *-wärts* im Zeitraum von 1100 bis 1350 zum produktiven Suffix zur Bildung von Richtungsadverbien wird. Er unterscheidet dabei drei Stufen:

1) Reihenbildung in Komposition mit anderen Adverbien des Kernbestandes *(abwerts, herwert, zuowert* etc.), 2) Entwicklung zum Suffix nach bestimmten Präpositionalen Wendungen (z.B. (19) *vrouwe Lachtasis daz ist mir bekant,/ di trecket mit irre hochvart / den vadem hin zu tale wart* BRUN 10449) und 3) in der Position nach vorangehenden *gegen-* oder *zuo-*Adverbialen (z.B. (21) *zv der burg wert* HERB 1375).

In 2 und 3 gibt es eine zweifache direktionale Kennzeichnung, wobei nach Wegfall der Präposition *wert* als einziges richtungsbestimmendes Merkmal bleibt und somit der Weg zum Suffix geebnet ist (wobei noch die Hinzunahme des adverb-typischen *-s* zu konstatieren ist). Graën erwähnt insgesamt 395 Belege, die Verwendung von *-wert* als Suffix setzt nach ihm in der zweiten Hälfte des 12. Jahrhunderts ein.

In der aktuellen Literatur zur deutschen Wortbildung wird *-wärts* als Suffix (u.a. bei Fleischer/Barz 2012: 369; Elsen 2014: 238; Lohde 2006: 293) bzw. Suffixoid (Kluge 2012: s.v. *-wärts*; Ganslmayer 2012: 378) bezeichnet. Henzen reiht *-wärts* in das Kapitel „Zur Ableitung übergetreten sind Adverbien mit einem Nomen als zweitem Glied, wo dieses wie ein Suffix analogisch weiterwirkt" (Henzen 1965: 232) ein. Wir wollen an dieser Stelle nicht weiter auf die Affix/Affixoid Diskussion eingehen und verweisen auf dahingehende Untersuchungen (vgl. Nübling et al. 2017: 94–97; Elsen 2014: 145–148; Donalies 2002: 25f). Dass einige Autor*innen *-wärts* als Suffixoid bezeichnen, spiegelt aber wider, dass eine vollständige Grammatikalisierung, obwohl es heute kein eigenständiges *wärts*, *werts*, *warts* o.ä. mehr gibt, nicht stattgefunden hat. So trägt *-wärts*, ähnlich wie *-weise*, *-weg*, *-dings* oder *-maßen* (vgl. Simmler 1998), einerseits immer noch eine klar definierbare Bedeutung, andererseits kann die Reihenbildung durchaus kritisch gesehen werden (siehe folgende Kapitel). Explizit als produktiv wird das Suffix von Barz (2016: 773), Elsen (2014: 238), Dargiewicz (2012: 64), Lohde (2006: 294) und Graën (2004: 147) beschrieben, oft wird auf die Produktivität nicht näher eingegangen. Aus dem Kontext ist aber bei den meisten Autor*innen zu verstehen, dass das Suffix grundsätzlich als produktiv angesehen wird (z.B. Nübling 2016: 581; Weinrich

2005: 568; Engel 2004: 418; Schulz/Griesbach 1972: 212). Heinle behauptet für die Entwicklung vom Frühneuhochdeutschen bis ins 20. Jahrhundert eine relative Konstanz der Beleglage von *-wärts* mit stärkster Produktivität in Verbindung mit Lokaladverbien z.B. *bergab-wärts* (Heinle 1985: 1915). Motsch bezeichnet dieses Wortbildungsmuster als „schwach aktiv" (2004: 235f.).

Zur derzeitigen Verwendungsweise des Suffixes gibt es kaum eingehendere Untersuchungen, insbesondere nicht solche, die sich auf reale Sprachdaten berufen. Es stellt sich nun die Frage, wie produktiv das Suffix *-wärts* eigentlich (noch) ist? Wir versuchen diese Frage zu beantworten, indem wir die Produktivitätsmaße TTR und HTR in den Korpora Alpenwort, Text&Berg, DWDS, COSMASIIweb, Schweizer Textkorpus vergleichen. Gleichzeitig arbeiten wir ganz konkret praktische Probleme und Unterschiede in der Handhabung dieser sehr verschiedenen Korpora heraus, um zu zeigen, wo Schwierigkeiten bei der Vergleichbarkeit liegen können.

3.1 Produktivitätsanalyse mit TTR und HTR

Die Frage, wie morphologische Produktivität mit quantitativen Methoden zuverlässig gemessen bzw. analysiert werden kann, ist durchaus umstritten (vgl. Aftabi et al., 2021: 1–3; Schneider-Wiejowski, 2012: 43–60). Eine Möglichkeit, Produktivität zu erfassen, ist die Analyse der Type-Frequenz. Baayen (2009: 901) bezeichnet das Verhältnis von Types zu Gesamttokens als *realized productivity* und sagt, dass sich die Produktivität auf „past achievements" beziehe. Stefanowitsch indessen definiert die Type-Token-Ratio (TTR) als Verhältnis von gefundenen Types zu gefundenen Tokens und beschreibt dieses folgendermaßen:

> Likewise, observing the type frequency (i.e. the TTR) of an affix under different conditions provides information about the relationship between these conditions and the affix itself, albeit one that is mediated by the lexicon: it tells us how important the suffix in question is for the subparts of the lexicon that are relevant under those conditions (Stefanowitsch 2020: 316).

Eine weitere, von Baayen (1992: 115) definierte Größe hat sich in der korpuslinguistisch orientierten Produktivitätsanalyse als gängige Vergleichs- und Untersuchungsgröße etabliert. Es handelt sich um die *expanding productivity* (vgl. auch Baayen 2009: 902) bzw. Hapax-Token-Ratio (HTR) (Stefanowitsch 2020: 318). Dieser Wert bezeichnet das Verhältnis von Hapax Legomena zu Gesamttokenanzahl und gibt damit Aufschluss über die aktuelle Produktivität eines Suffixes.

Mit diesen Größen lassen sich sowohl die Produktivität in einem Korpus zu einer bestimmten Zeit feststellen, als auch, bei Vorliegen einer zeitlich gestaffelten Textsammlung, diachrone Produktivitätsentwicklungen nachvollziehen. Dies wurde z.B. von Schneider-Wiejowski für verschiedene Suffixe im Schweizerdeutschen gemacht (Schneider-Wiejowski 2013).

4 Der *-wärts*-Komplex in verschiedenen Korpora

Für Produktivitätsanalysen besteht Konsens, dass die Angabe der relativen Tokenfrequenz allein keine getreue Darstellung der Produktivität darstellt (vgl. Stefanowitsch 2020: 315). Dennoch ist die Betrachtung dieser Größe interessant, da sie eine Aussage über die generelle Häufigkeit der Verwendung eines Suffixes – sei es nun produktiv oder nicht – zulässt. Vergleicht man die relativen Häufigkeiten der verschiedenen Korpora, fällt auf, dass die alpinbezogenen Texte eine deutlich höhere maximale Tokenfrequenz (ca. 800 fpmw) gegenüber den anderen Korpora aufweisen (ca. 200 – 250). Es bestätigt sich hier also die eingangs getätigte Annahme, dass in Texten mit Alpinbezug direktionale Adverbien und damit auch solche mit Suffix *-wärts* häufiger verwendet werden. Außerdem zeigt sich in allen Korpora, in welchen eine zeitliche Staffelung vorliegt, dass nach einem Höchststand Anfang/Mitte des 20. Jahrhunderts die absolute und relative Tokenfrequenz bis zum Ende des 20. Jahrhunderts drastisch abnimmt. Die Größe dieser Abnahme gibt dann schon einen ersten Hinweis auf die Entwicklung der Produktivität.

4.1 Alpenwort und Text&Berg

Um dem aufgespürten Phänomen *-wärts* nachzugehen bietet sich ein näherer Blick auf das schon eingeführte Korpus Alpenwort (siehe S. 119) an sowie ein Vergleich mit dem sehr ähnlichen „Schwesterkorpus" Text&Berg (Text+Berg-Korpus R151v01). Die Ähnlichkeit besteht sowohl inhaltlich als auch in der Zeitspanne der beiden Korpora: Text&Berg beinhaltet alle Jahrbücher des Schweizer Alpenclubs bwz. die Monatszeitschrift *Alpen* 1864–2015. (Göhring/Volk 2011). Verwendet wurden daraus alle deutschsprachigen Texte, was einer Größe von ca. 23 Mio. laufender Wortformen entspricht (Bubenhofer/Rothenhäusler 2018: 41). Wie auch Alpenwort ist Text&Berg vollständig korrigiert und POS-getagged vorhanden, beide Korpora sind über CQPweb abfragbar und lassen sich deshalb ausgezeichnet vergleichen.

Wie aus Abbildung 2 ersichtlich ist, steigen TTR und HTR sowohl im Alpenwort als auch im Text&Berg Korpus im Lauf der Zeit deutlich an, was auf eine Zunahme der Produktivität hinzuweisen scheint. Zusätzlich gibt es im Alpenwort Korpus zwei ausgeprägte Spitzen, die erklärungswürdig sind. Wie Stefanowitsch (2021: 320) zeigt, ist weder der Zusammenhang zwischen Korpusgröße und Types bzw. Hapax, noch der zwischen Treffermenge und Types bzw. Hapax linear. In beiden Fällen gilt die Regel, dass die Anzahl der Types bzw. Hapax bei geringer Tokenmenge bzw. Treffermenge zuerst stark ansteigt, bevor es zu einer gewissen Plateaubildung kommt. Das heißt, dass für den erwähnten Anstieg von TTR und HTR primär die Abnahme der Treffermenge ausschlaggebend ist, und nicht eine wirkliche Produktivitätssteigerung, bei der mit der Zunahme der TTR beziehungsweise HTR auch eine Zunahme der Treffermenge zu erwarten wäre. Die Ausreißer im Alpenwort Korpus haben zwei unterschiedliche Ursachen. Die Spitze im Jahrzehnt 1940 kann sehr gut mit der unterschiedlichen Teilkorpusgröße erklärt werden. Während in allen anderen Jahrzehnten die Tokenzahl zwischen knapp einer Million und gut zwei Millionen Tokens schwankt, ist aufgrund der durch den Zweiten Weltkrieg verursachten Publikationslücke das Teilkorpus nur ca. 300.000 Tokens groß. Dies führt aufgrund des oben beschriebenen nicht linearen Zusammenhangs zur Spitze. Wie auch Stefanowitsch (2020: 321) analysiert: „Therefore, comparing TTRs derived from samples of different sizes will always make the

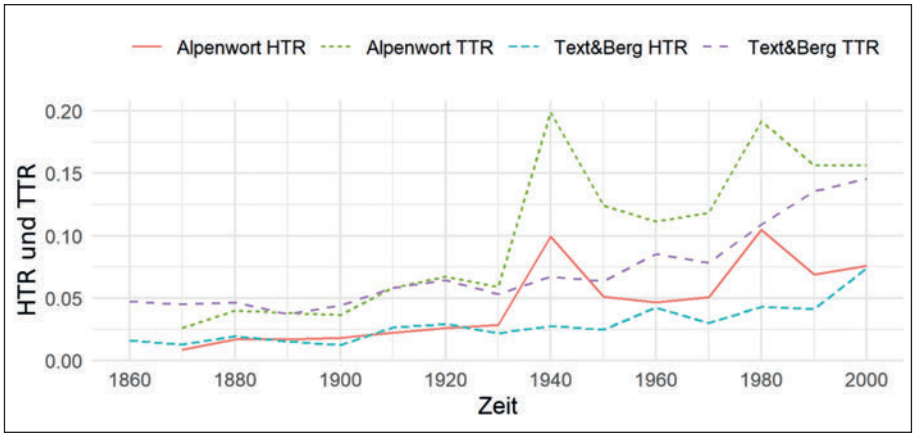
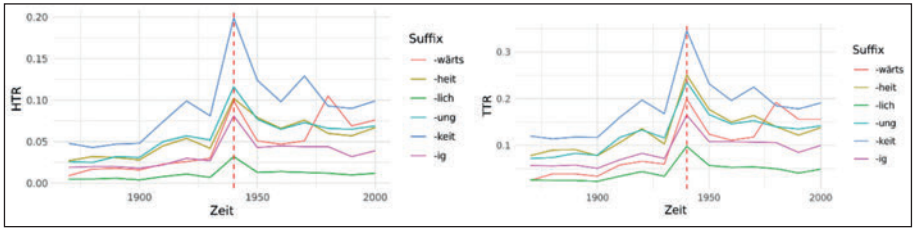


Abbildung 2: Entwicklung von TTR und HTR im Untersuchtungszeitraum (Intervall 10 Jahre)

smaller sample look more productive." Dass es sich um einen Zusammenhang mit der Korpusgröße handelt, erkennt man, wenn man die Produktivität anderer Suffixe analysiert. Hier ergibt sich ebenfalls für das Jahrzehnt 1940 ein auffälliger Produktivitätsgipfel (vgl. Abbildungen 3 und 4).

Das Teilkorpus 1980 scheint in der Tat ein Ausreißer zu sein, wobei vor allem das Jahr 1988 auffällt, das 38% der Hapax liefert (allerdings von verschiedenen Autor*innen). Insofern haben wir hier die Situation, dass bei einem Phänomen geringer Frequenz bereits kleine Unterschiede relativ große Auswirkungen haben.



Abbildungen 3 und 4: Entwicklung von TTR und HTR für verschiedene Suffixe

4.2 DWDS

Das Digitale Wörterbuch der Deutschen Sprache (DWDS, siehe Barbaresi/Geyken 2020) bietet eine ausgezeichnete Grundlage für korpuslinguistische Fragestellungen: einerseits lassen sich über die online-Werkzeuge, die auf https://www.dwds.de/ zur Verfügung gestellt werden, bereits aussagekräftige Abfragen und Visualisierungen tätigen, andererseits bietet die Seite benutzungsfreundliche Exportschnittstellen. Über diese Exportmöglichkeiten lassen sich Abfrageergebnisse lokal weiterverarbeiten, man ist also nicht durch die Vorgaben von DWDS-online beschränkt. Abgerundet werden diese Möglichkeiten durch eine ausgezeichnete Dokumentation mit zahlreichen Abfragebeispielen. Die Referenzkorpora, die wir für unsere Analyse verwendet haben, decken einen Zeitbereich von ca. 1600–2010 ab.

Um die Daten vergleichbar zu verarbeiten und darzustellen, wurden die im Folgenden dargestellten Verarbeitungsschritte durchgeführt. Die Abfrage über das Online-Interface ist einfach und wird über eine umfassende Suchsyntax gut unterstützt. Es gibt allerdings einige Eigenheiten, die insbesondere bei Phänomenen mit geringer Okkurrenz eine Rolle spielen können und deshalb hier thematisiert werden. Erstens: Wirkt sich die fehlende verbindliche Orthografie in den ersten beiden Jahrhunderten auf das Ergebnis aus und wenn ja wie? Kann urheberrechtlich geschütztes Material ebenfalls eine Auswirkung haben? Wie auf S. 130 dargelegt, werden in der KWIC-Ansicht urheberrechtlich geschützte Inhalte nämlich nicht dargestellt, was vor allem für die jüngeren Texte deutliche Auswirkungen zeigt.

1) Abfrage und Analysevorbereitung: Bei allen drei zur Verfügung stehenden Referenzkorpora (dtak, kern, korpus21) wurden Lemmata, die auf *-wärts* enden, abgefragt (in der Abfragesyntax `$l=*wärts`). Die Abfrage über das Lemma ist in diesem Fall wichtig, weil ältere Sprachstufen eine oft stark abweichende Orthographie aufweisen (*-wertz*, *-waerts*, *-werths* etc.). Damit erhält man insgesamt 30.448 exportierbare Funde (im Gegensatz zu lediglich 26.791 Funden mit der Abfrage über die Wortform). Der Export im .csv-Format besteht aus Genre, Titel, Kontext und vor allem Jahr, so dass sich die Anzahl der Funde als Zeitfunktion abbilden lässt. Die Funde wurden dabei ebenfalls in Jahrzehnte zusammengefasst (jeweils Jahre 0–9). Die unvollstän-

digen Jahrzehnte 1590–1599 und 2010–2019 wurden entfernt, um statistische Ausreißer wegen extrem kleiner Korpusgrößen zu verhindern. Abbildung 5 zeigt die absoluten Frequenzen des Lemmas *-wärts* und es wird sichtbar, dass bis ca. 1780 ein substantieller, bis ins 19. Jahrhundert ein geringer Unterschied in der Treffermenge zwischen Lemma-Suche (blau liniert) und Wortform-Suche (rot punktiert) besteht. Die Abfrage per Lemma ist also für die DWDS-Korpora auf jeden Fall zu bevorzugen.

Nicht mit im Export sind Textstellen, die aus urheberrechtlichen Gründen nicht freigegeben wurden. So wird man zum Beispiel bei der Abfrage im kern-Korpus darauf hingewiesen, dass nur 281 von 571 Treffern, also nicht einmal die Hälfte, dargestellt werden können. Auf der Suche nach Abhilfe für diese unbefriedigende Situation wurde folgende Lösung gefunden: Die urheberrechtlich beschränkte Darstellung gilt nur für Ansichten, in denen Kontext abgerufen wird. Abfragen, die auf Wortebene basieren, werden korrekt dargestellt. Somit lässt sich mit der Abfrage `COUNT($l=*wärts)#BY[$w,$l,date/1]`[1] eine Liste erstellen, in der sowohl die Wortform als auch das dazugehörige Lemma für jeweils ein Jahr als Summe vorhanden ist (siehe Tabelle 2).

Summe	Wortform	Lemma	Jahr
1	Abwerts	abwärts	1666
4	Abwerts	abwärts	1688
1	Anderwärts	anderwärts	1682
1	Auffwerts	aufwärts	1639
2	Auffwerts	aufwärts	1659
4	Auffwerts	aufwärts	1688
1	Auffwärts	aufwärts	1657
1	Aufwerts	aufwärts	1681
2	Aufwerts	aufwärts	1682

Tabelle 2: Wortform und Lemma pro Jahr

Mit dieser Abfrage erweitert sich das Ergebnis auf 32.443 Funde. Dies ist insofern bedeutend, als diese nicht gleichmäßig verteilt, sondern primär im 20.

1 vgl. https://www.dwds.de/d/korpussuche#count

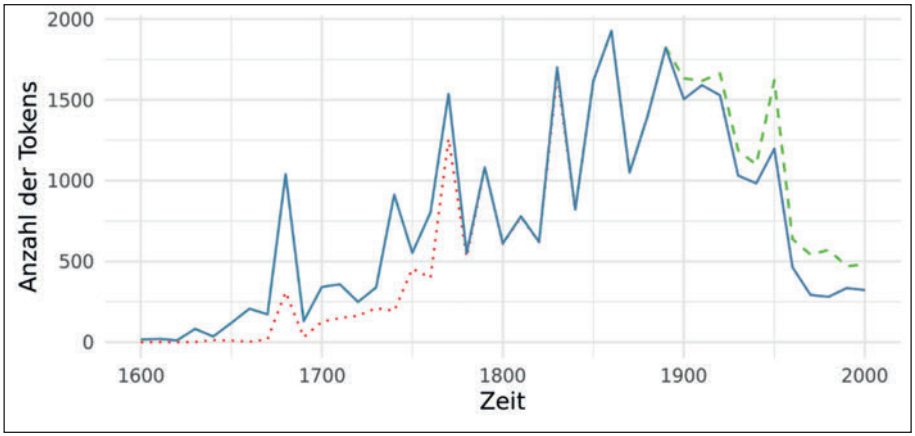


Abbildung 5: Absolute Anzahl der Tokens mit Suffix -wärts

Jahrhundert angesiedelt sind. In Abbildung 5 sind die zusätzlichen Funde als grün strichlierte Linie eingezeichnet. Für die weitere Verarbeitung wurden die Daten zu Jahrzehnten (wieder Jahre 0–9) akkumuliert.

2) Für weitere Berechnungen lässt sich die Gesamtkorpusgröße in der DWDS-Oberfläche einfach abfragen: Mit `COUNT( * ) #BY[date/1]` erhält man eine Liste mit der absoluten Tokenanzahl pro Jahr. Diese Liste wurde wieder in Jahrzehnte zusammengefasst. Abbildung 6 zeigt die Korpusgröße als Zeitfunktion und zeigt ebenfalls, dass die Größen der jeweiligen Subkorpora sich deutlich unterscheiden (sie variieren von knapp über 100.000 bis über 15 Millionen Tokens), wobei die Korpusgröße im 20. Jahrhundert jeweils über 10 Millionen Tokens liegt und gegen 1600 hin abnimmt.

3) Zur eigentlichen Produktivitätsberechnung muss jetzt noch die Anzahl der Types sowie der Hapax Legomena berechnet werden. Aufgrund der bereits angesprochenen orthographischen Probleme wurden für diese Berechnung ebenfalls die Lemmata verwendet. Dies hat außerdem den Vorteil, dass durch die letzte Orthographiereform verursachte Probleme (*flussabwärts* vs. *flußabwärts*) ausgeglichen werden, da bekannte Lemmata in der neuen Rechtschreibung zusammengefasst werden (beides *flussabwärts*). Abbildung 7 zeigt die Anzahl der Hapax berechnet anhand der Wortformen (blau strichliert) und anhand der Lemmata (rot liniert).

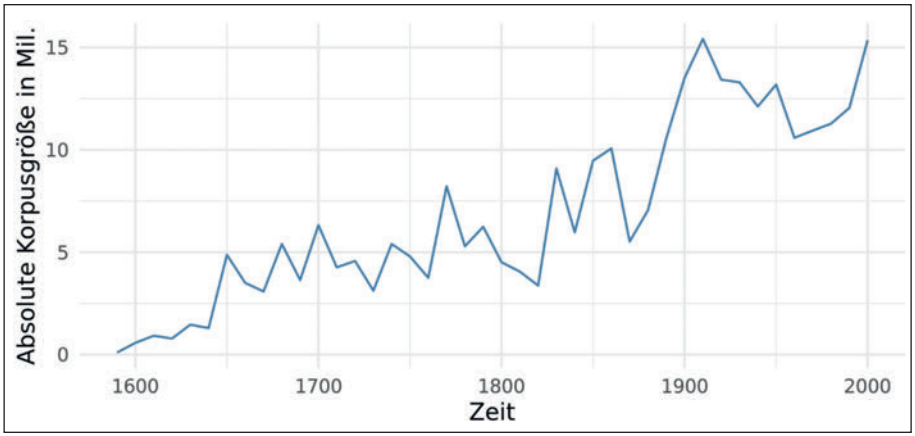


Abbildung 6: Absolute Korpusgröße im Untersuchungszeitraum (zusammengefasst in Dekaden)

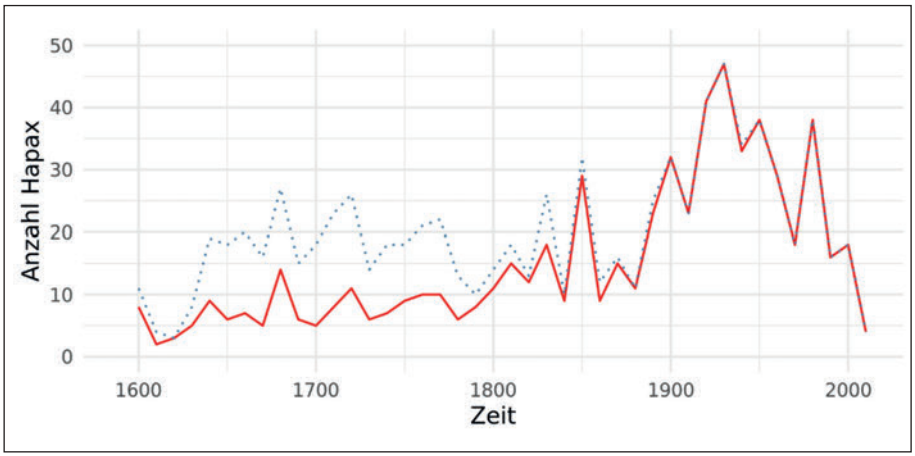


Abbildung 7: Hapax Legomena im Untersuchungszeitraum anhand Wortform bzw. Lemma

Auch für die Berechnung der Types hat die Verwendung von Lemmata große Auswirkungen im Zeitraum vor 1900. Abbildung 8 zeigt die Anzahl der Types berechnet anhand der Wortformen (grün strichliert) und anhand der Lemmata (blau liniert).

Aufbauend auf diesen Daten können nun die relative Tokenfrequenz, TTR und HTR berechnet werden. Betrachtet man die relative Tokenfrequenz

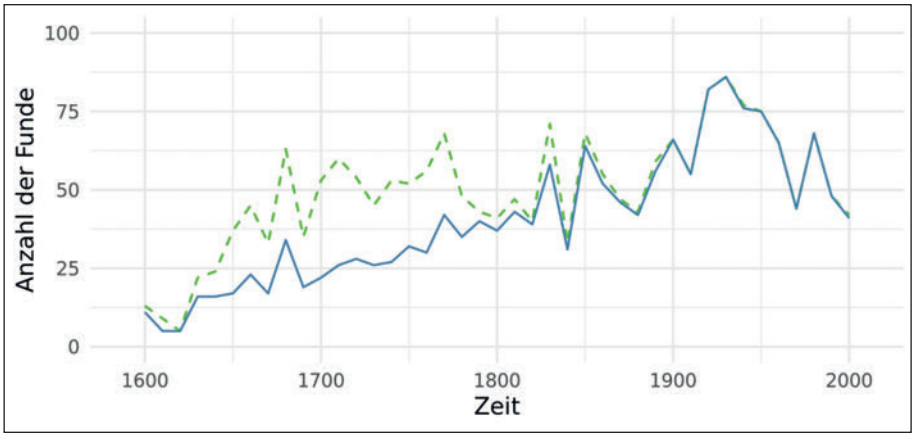


Abbildung 8: Types im Untersuchungszeitraum anhand Wortform bzw. Lemma

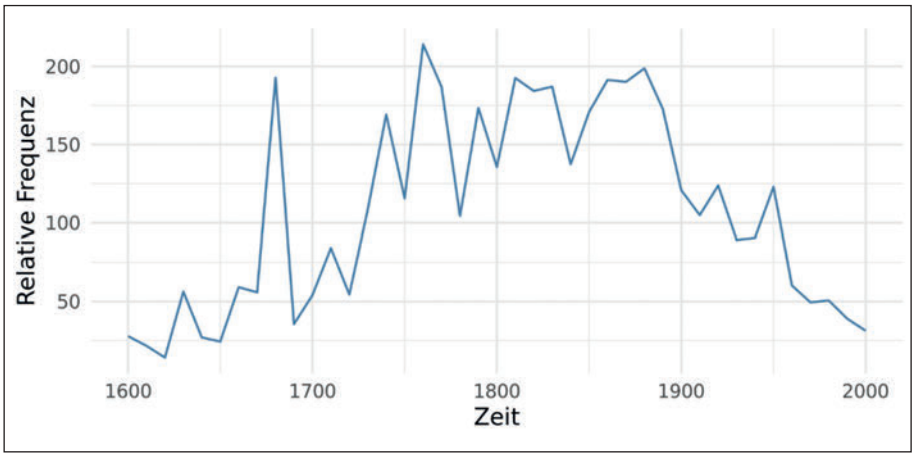


Abbildung 9: Relative Frequenz im Untersuchungszeitraum

(Abbildung 9), so lässt sich ein Ansteigen der Verwendung von Adverbien auf *-wärts* vom 17. bis zum 18. Jahrhundert erkennen. Die Spitzen zeigen an, wie stark sich einzelne Autor*innen auswirken können. So stammen z.B. von den 1041 Funden im Jahrzehnt 1680 902 Funde von nur drei Autoren (416 Johann Christoph Pinter von der Au, 348 Wolf Helmhard von Hohberg, 138 Daniel Casper von Lohenstein). Die Präferenz einzelner Personen für

diese Adverbien gepaart mit der relativ geringen Korpusgröße führt bei einem Randphänomen, wie es die Ableitung mit *-wärts* darstellt, bereits zu gröberen Verzerrungen. Auch die unterschiedliche Zusammenstellung des Korpus (vgl. https://www.deutschestextarchiv.de/doku/textauswahl) könnte einen Einfluss auf die Ausreißer haben, wenngleich sich dies nicht so direkt feststellen lässt wie bei den Autor*innen. Ab dem Ende des 19. Jahrhunderts beobachten wir ein relativ konstantes Nachlassen und da wir hier ein deutlich größeres Korpus besitzen, können wir auch mit größerer Sicherheit vom allgemeinen Nachlassen der Verwendung dieser Adverbien sprechen.

Betrachten wir TTR und HTR (Abbildungen 10 und 11) als Maße für die Produktivität, so ergibt sich bei beiden ein ähnliches Bild: In der ersten Phase, bis in die Mitte des 17. Jahrhunderts ist das Korpus so klein, dass keine sinnvollen statistischen Aussagen getroffen werden können. Erst bei einer Korpusgröße von ca. 3 Millionen Tokens pendeln sich beide Werte auf einem stabilen, sehr niedrigen Niveau ein.

Am interessantesten ist hierbei die Beobachtung, dass auch in der Zeitspanne, in der die Verwendung von Adverbien auf *-wärts* absolut steigt, keine Zunahme von TTR oder HTR zu verzeichnen ist. Mit anderen Worten, die Produktivität steigt nicht an.

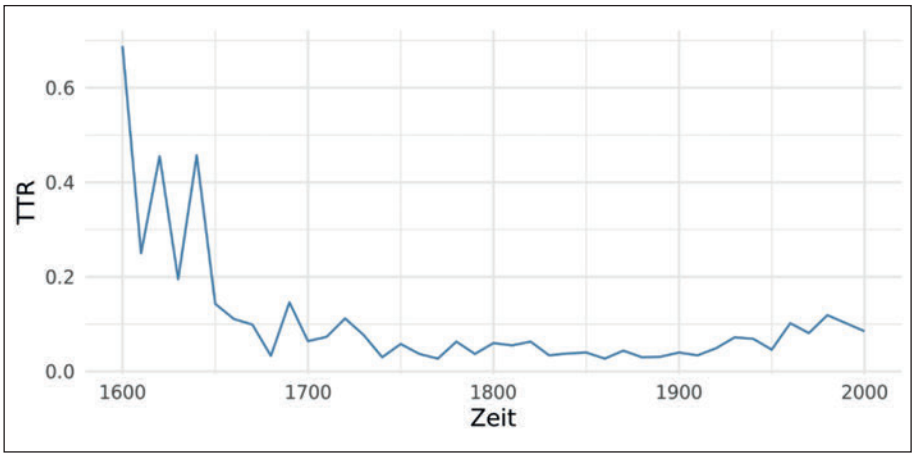


Abbildung 10: TTR im Untersuchungszeitraum

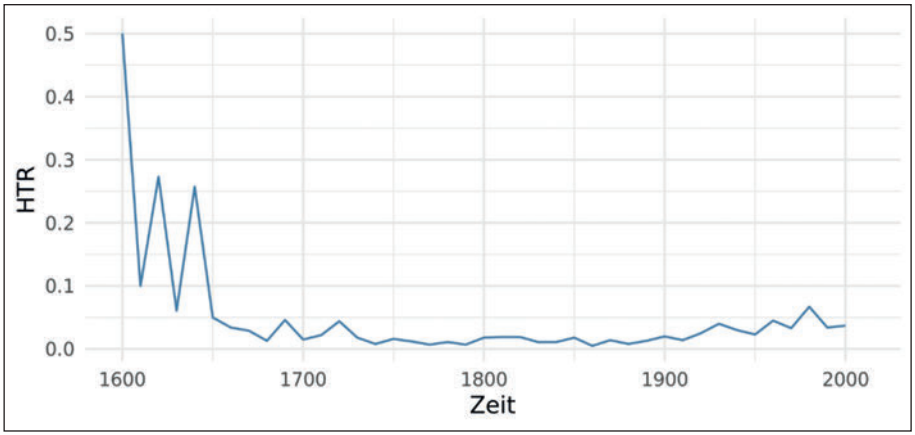


Abbildung 11: HTR im Untersuchungszeitraum

Zusammenfassend lässt sich sagen, dass das DWDS für eine diachrone Untersuchung eine ausgezeichnete Grundlage bildet. Beachtet man die Einschränkungen, die sich aus Orthographie und Urheberrecht ergeben, so kann man dennoch weitgehend automatisch statistische Aussagen über den dargebotenen Zeitraum erstellen. Für seltene Phänomene ist die Größe bzw. Kleinheit des Korpus insbesondere in der ersten Hälfte des 16. Jahrhunderts zu beachten. Weiters kann die unterschiedliche Zusammensetzung des Korpus problematisch für manche Fragestellungen sein.

4.3 DeReKo

Im Gegensatz zu den beiden thematisch fokussierten hochspezialisierten Korpora Alpenwort und Text&Berg bietet die COSMASIIweb-Schnittstelle des Instituts für Deutsche Sprache in Mannheim Zugang zu 573 unterschiedlichen Korpora. Aktuell besteht dort das verfügbare Korpus DeReKo (Deutsches Referenzkorpus, siehe Kupietz et al. 2018) aus 50,6 Mrd. (Stand 2.2.2021) laufenden Wortformen (lt. Webseite). Die enthaltenen Korpora werden laufend erweitert und gelten somit als Referenzkorpus des gegenwärtigen Deutschen. Das DeReKo ist ausdrücklich nicht als ein ausgewogenes Korpus konzipiert (Kupietz et al. 2009: 1849), was die Vergleichbarkeit der Größen TTR und

HTR mit dem DWDS oder dem Schweizer Textkorpus praktisch unmöglich macht.[2] Das DeReKo hat keinen Anspruch auf große zeitliche Tiefe und es kann hauptsächlich synchron gearbeitet werden. Im gesamten Korpus sind sehr unterschiedliche Teilkorpora mit ebenfalls unterschiedlichen Zeitbreiten vorhanden, so dass diese nicht methodisch sauber diachron betrachtet werden können. Die interessanteste Untersuchungsmöglichkeit im Bezug auf *-wärts*-Suffigierung wäre die lexikalische Varianz der entstehenden Wörter. Diese Untersuchung würde den hier gesetzten Schwerpunkt einer quantitativen Untersuchung sprengen. Eine diesbezügliche Studie, wieder unter Einbezug der hier verwendeten Korpora, ist in Vorbereitung.

TTR und HTR lassen sich im DeReKo zwar errechnen, die Werte sind allerdings nicht mit den anderen Ergebnissen vergleichbar, da, wie vorher bereits erwähnt, die Korpusgröße und die Trefferanzahl nicht linear sind. Zu beachten ist, dass bei den Hapax eine große Menge falscher Einträge zu finden sind, die aus OCR-Fehlern etc. stammen. Im hier vorliegenden Fall betrug die Anzahl der falschen Hapax immerhin 18.7%. Die errechnete TTR von 0.0043 und die HTR von 0.0022 kann jedoch für sich genommen wenig Auskunft über die Produktivität des Suffixes geben. Deshalb ist die Verwendung des DeReKo für Analysen dieser Art nur begrenzt tauglich.

4.4 Schweizer Textkorpus

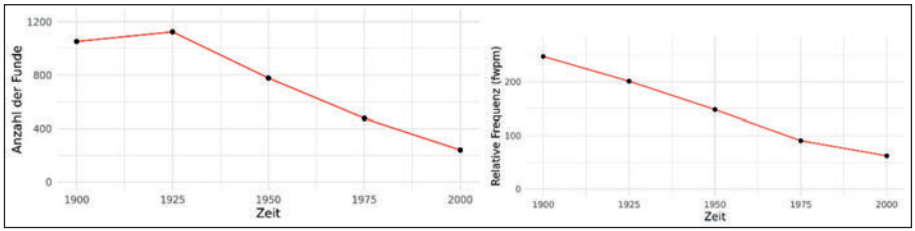
Das Schweizer Textkorpus (Bickel et al. 2009) orientiert sich in seinem Aufbau am BNC. Es bietet ein Korpus, das Gebrauchstexte, Sachtexte, Belletristik und journalistische Prosa im Bezug auf Korpusteilgröße und zeitlicher Distribution ausgewogen zur Verfügung stellt. Das Korpus deckt derzeit einen Zeitbereich von 1900–2018 ab und ist, bezogen auf das 21. Jahrhundert, weiterhin im Aufbau begriffen. Der Vorteil für den hier untersuchten Zeitbereich 1900–2000 ist also, dass das Korpus sowohl im Hinblick auf Textsorten als auch auf Größe sehr homogen ist. Damit können, zumindest auf das Deutsche in der Schweiz bezogen, vergleichbare Aussagen zur Produktivität ge-

2 Dies könnte sich mit der neuen Abfrageoberfläche KORAP ändern, für diese Studie wurde diese Möglichkeit jedoch nicht näher untersucht.

macht werden. Das Korpus ist mit Hilfe der DDC Query Language bequem zu benutzen, ein kleiner Nachteil ist jedoch, dass die Exportmöglichkeiten beschränkt sind (max. 500 Treffer können exportiert werden). Damit wird die Datensammlung zur Errechnung von nicht direkt im Interface abfragbaren Größen wie TTR und HTR etwas umständlicher. Die absolute Korpusgröße wird in der Dokumentation für einen Zeitraum von jeweils 25 Jahren für die einzelnen Textsorten sowie für das Gesamtkorpus angegeben. Damit lässt sich die relative Frequenz für diese Zeitbereiche prinzipiell einfach errechnen.

Die für diese Untersuchung verwendete Abfrage zum Erhalt aller Wörter auf *-wärts* ist `*wärts #SEPARATE_HITS`. Der Zusatz `#SEPARATE_HITS` ist nötig, weil ohne diesen mit der DDC Query Language nur jeweils das erste Wort eines Satzes, auf das die Suchbedingung zutrifft, gefunden bzw. angezeigt wird. Der Unterschied zwischen der Suche mit `*wärts` gegenüber `*wärts #SEPARATE_HITS` beträgt immerhin 3.382 vs. 3.622 Treffer.[3]

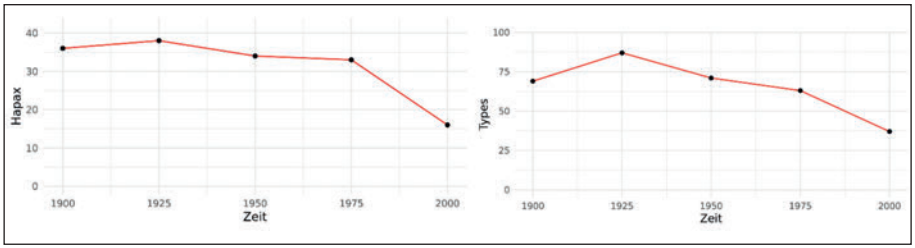
Wie aus den Abbildungen 12 und 13 zu sehen ist, fällt sowohl die absolute Zahl der Treffer als auch die relative Frequenz, ähnlich wie im Alpenwort Korpus, nach einem Hoch Anfang des 20. Jahrhunderts stetig ab. Die relative Frequenz ist mit anfänglich 200 fpmw im Bereich des DWDS.



Abbildungen 12 und 13: Anzahl der Funde und relative Frequenz im Untersuchungszeitraum

Sowohl die Anzahl der Types als auch die Anzahl der Hapax nimmt in der zweiten Hälfte des Untersuchungszeitraums ab (Abbildungen 14 und 15). Die Produktivitätsindizes TTR und HTR (Abbildung 16) nehmen jedoch im selben Zeitraum zu. Da die Teilkorpusgröße im Schweizer Textkorpus annähernd gleich bleibt, scheint dies also auf eine Produktivitätszunahme hinzu-

3 Die Untersuchung von Schneider-Wiejowski (2013) wäre dahingehend anzupassen.



Abbildungen 14 und 15: Anzahl der Hapax und Types im Untersuchungszeitraum

weisen. Dies ist, wie oben bereits erwähnt, aber insofern trügerisch, als der Zusammenhang zwischen Types bzw. Hapax und gefundenen Tokens nicht linear ist. Wie bei Stefanowitsch (2020: 320, Figure 9.1) zu sehen ist, steigt die Anzahl der Types/Hapax im Verhältnis zur Tokenmenge zuerst steil an, bevor es zu einer Plateaubildung kommt. Das heißt, dass auch bei annähernd gleicher Korpusgröße TTR und HTR nicht linear zur Treffermenge verlaufen. Je kleiner die Treffermenge, desto größer sind TTR und HTR. Der Anstieg in der zweiten Hälfte des 20. Jahrhunderts ist also stärker auf die Abnahme der Verwendung insgesamt zurückzuführen, als dass es eine wirkliche Steigerung der Produktivität gäbe. Diese Entwicklung ist also ähnlich wie in oben beschriebenen thematisch fokussierten Korpora.

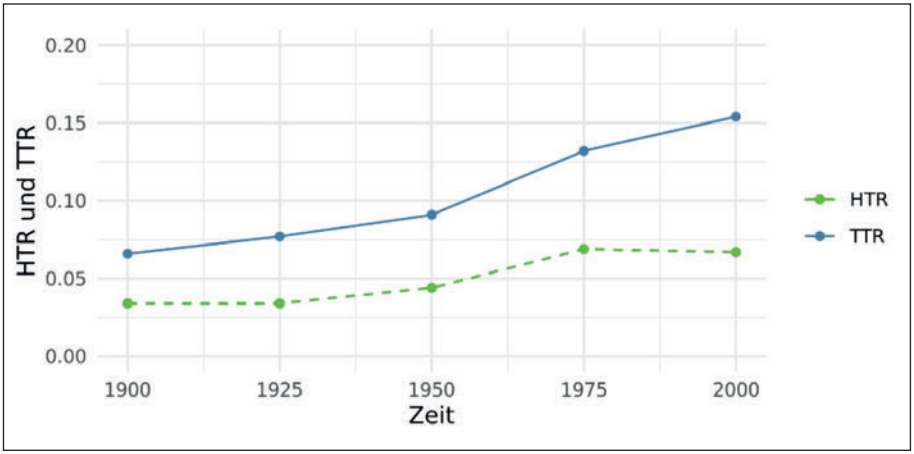


Abbildung 16: HTR und TTR im Untersuchungszeitraum

5 Zusammenfassung

Einige bei Keyword-Vergleichen im Korpus Alpenwort auffällige Bildungen mit dem Suffix *-wärts* weckten unser Interesse dieses spannende Derivationsmorphem näher zu beleuchten. Mit unserem Beitrag gehen wir einerseits mittels TTR- und HTR Berechnungen der Frage nach, ob das Suffix *-wärts* heute noch als produktiv gewertet werden kann. Andererseits wollen wir aber auch interessierten Forscher*innen darlegen, wie und ob sich für diese Fragestellung stark unterscheidende Arten von Korpora vergleichen lassen bzw. welche Fallstricke hierbei auftreten können.

Die eingangs erwähnte Hypothese, dass Direktionaladverbien in raumbezogenen Texten, wie sie in den Alpenwort und im Text&Berg Korpora zu finden sind, häufiger gebraucht werden als in anderen Texten, konnte mittels Frequenzanalyse belegt werden. In Anlehnung an die anfangs zitierten Worte von Bubenhofer/Rothenhäusler (2018) sind diese wohl Teil einer bestimmten, nennen wir es Bergtext-Anatomie. Für die Untersuchung von Direktionaladverbien bietet dies den Vorteil, in diesen Korpora eine größere Grundmenge des Untersuchungsgegenstands zu haben. Allgemeiner ausgedrückt: Thematische Korpora bieten Vorteile, weil das Auftreten eines bestimmten linguistischen Phänomens in ihnen gezielter untersucht werden kann als in ausgewogenen oder breit angelegten Referenzkorpora. Dies gilt insbesondere dann, wenn es sich um eher seltene Phänomene handelt.

Die Frage nach der Produktivität des Suffixes *-wärts* selbst muss facettiert beantwortet werden. So scheinen die Indizes TTR und HTR in allen diachron auswertbaren Korpora (Alpenwort, Text&Berg, Schweizer Textkorpus, DWDS) auf einen Anstieg der Produktivität in der zweiten Hälfte des 20. Jahrhunderts hinzuweisen. Gegen diese Interpretation spricht die Tatsache, dass sowohl die absolute als auch relative Tokenfrequenz im selben Zeitraum drastisch abnehmen und die vermeintliche Zunahme praktisch ausschließlich auf das nichtlineare Verhältnis von HTR und TTR zu gefundenen Tokens zurückzuführen ist. Eine geringe Anzahl von gefundenen Tokens führt dazu, dass geringfügige Änderungen, z.B. die auf einen Autor zurückzuführende Präferenz dieser Direktionaladverbien, zu unverhältnismäßigen Produktivitätsspitzen führen. Dies ist insbesondere in den frühen Jahrzehnten im DWDS

und anhand der Spitze im Jahrzehnt 1980 im Alpenwort Korpus gut zu beobachten. Auch die absolute Korpusgröße kann eine Rolle spielen, wie die Spitze in der Dekade 1940 im Alpenwort Korpus zeigt.

Die ernüchternde Feststellung ist, dass sich weder HTR noch TTR als Messgröße für ein seltenes Phänomen wie das hier untersuchte zu eignen scheinen. Zwei Lösungen scheinen möglich zu sein: Die erste Lösung besteht in der Vergrößerung des Korpus, so dass man sich mit HTR und TTR in einem Bereich befindet, in dem eine Abnahme der Tokenfrequenz weniger starke Auswirkungen auf die Maßzahl selbst hat (dies entspricht einer Verschiebung auf der Kurve TTR/Token bzw. HTR/Token auf einen Bereich, in dem die Kurve abflacht). Die zweite Lösung wäre, den TTR und HTR Wert abhängig von der absoluten bzw. relativen Tokenfrequenz zu korrigieren. Eine dahingehende Diskussion würde allerdings den Rahmen dieser Untersuchung sprengen. Unabhängig von diesen Problemen kann aber festgestellt werden, dass der Rückgang der Gebrauchsfrequenz des Suffixes *-wärts* im 20. Jh. über verschiedene Korpora hinweg zu beobachten ist. Dieser Rückgang scheint textsortenunabhängig zu sein. Daher wäre die Einschätzung zur Produktivität folgendermaßen zu korrigieren: De facto werden Wörter mit dem Suffix wesentlich weniger gebraucht als vor hundert Jahren und das Suffix ist derzeit nicht mehr produktiv. Eine weitere interessante Frage, die an dieser Stelle aber nicht beantwortet werden kann, ist, ob in den jüngeren Texten alternative Formulierungen anstelle der direktionalen Adverbien mit *-wärts* verwendet werden und welche diese sind.

Die vielleicht wichtigste methodische Erfahrung zur Analyse der Suffixproduktivität ist, dass die absolute Korpusgröße sowie die absolute Tokenhäufigkeit nicht zu unterschätzende Auswirkungen auf die Größen TTR und HTR haben und ein Vergleich verschiedener Korpora bzw. verschiedener Teilkorpora deshalb nur mit größter Vorsicht gemacht werden sollte. So ist beim DWDS, das eine hervorragende Quelle für diachrone Untersuchungen darstellt, aufgrund der großen Unterschiede der Größen der Teilkorpora ohne Angabe dieser schwer abzuschätzen, wie stark die jeweiligen Produktivitätsmaße von dieser Größe abhängig sind. Es ist deshalb angebracht, zur Produktivität auch die Teilkorpusgrößen anzugeben, um eine Einschätzung der Ergebnisse vornehmen zu können. Ausgewogene Korpora wie das Schweizer

Textkorpus bieten hier den eindeutigen Vorteil, dass durch die Ähnlichkeit der Teilkorpusgrößen eine dahingehende Beeinflussung ausgeschlossen werden kann. Es besteht, trotz der Verwendung eingeführter statistischer Methoden, eine gewisse Schwierigkeit verschiedene Korpora zu vergleichen.

Ein weiteres, nicht zu unterschätzendes Problem ist, dass die Korpusabfragewerkzeuge über unterschiedliche Suchsyntax verfügen. Hier obliegt es derzeit noch den einzelnen Wissenschaftler*innen, sich gewissenhaft in die einzelnen Abfragesprachen einzuarbeiten. Eine Lösung des Problems ist die Bündelung verschiedener Korpora unter einer gemeinsamen Abfrageoberfläche (z.B. ist Text&Berg auch im DWDS verfügbar). Damit ließe sich zumindest teilweise ein weiteres Problem, nämlich die Verwendung unterschiedlicher Begrifflichkeiten, lösen. Beispielsweise wird in der Dokumentation die absolute Größe des Schweizer Textkorpus in Textwörtern angegeben, welche Zeichensetzung ausblendet. Die meisten anderen Korpora haben jedoch eine Tokendefinition, die Interpunktion mit einbezieht. Dies dürfte Benutzer*innen nicht so ohne weiteres klar sein und kann dazu führen, dass Größen wie die relative Tokenfrequenz unterschiedlich berechnet werden und damit nicht direkt vergleichbar sind. Natürlich wäre es unrealistisch zu erwarten, dass alle Korpora identische Angaben und (Abfrage-)Möglichkeiten liefern. Nichtsdestotrotz wären Hinweise auf Vergleichbarkeit in den Korpusdokumentationen nützlich. Eine Übersichtsstudie über genau diese Anwendungsunterschiede stellt ein Desiderat dar.

Ziel dieses Beitrags ist unsere Leser*innen über Möglichkeiten datengeleitet ein Forschungsinteresse zu entwickeln zu unterrichten. Weiters wollten wir mit unseren quantitativen Analysen darlegen, dass das zentrale Forschungsinteresse, das Suffix *-wärts*, in deutschen Texten sinkend produktiv ist. Es ist uns besonders wichtig aufzuzeigen, dass gerade quantitative Analysen extrem vorsichtig betrieben werden müssen und wir empfehlen insbesondere ein Augenmerk auf Vergleichsgrundlagen zu legen, besonders wenn verschiedene (Arten) von Korpora und verschiedene Plattformen verwendet werden. Nicht zuletzt hoffen wir, dass die Diskussion der Schwierigkeiten sowie Ergebnisse für andere Korpusanalysen sinnvoll sind sowie zur Methodenentwicklung beitragen können.

Webseiten Korpora (Stand 04/2021)

Alpenwort CQPweb Edition. http://sprawi-cqpweb.uibk.ac.at/CQPweb/
DWDS – Digitales Wörterbuch der deutschen Sprache. https://www.dwds.de/
DeReKo – Deutsches Referenzkorpus. https://cosmas2.ids-mannheim.de/cosmas2-web/
Schweizer Text Korpus. https://www.chtk.ch/index.php/de/
Text&Berg digital. http://textberg.ch/site/de/willkommen/

Bibliographie

Aftabi, S. Z./Ahangar, A. A./Mishmast Nehi, H. (2021): Derivational Suffix Productivity in Persian: A Fuzzy Analysis. *Journal of Quantitative Linguistics*, 1–25. https://doi.org/10.1080/09296174.2021.1887575.

Baayen, H. (1992): A quantitative approach to morphological productivity. In: G. Booij/J.van Marle (Hg.), *Yearbook of morphology 1991*. Dordrecht: Kluwer, 109–149.

Baayen, R. H. (2009): Corpus linguistics in morphology: Morphological productivity. In: A. Lüdeling (Hg.), *HSK: Vol. 29,2. Corpus linguistics 2: An international handbook*. Berlin: De Gruyter, 899–919.

Barbaresi, A./Geyken, A. (2020): Die Webkorpora im DWDS - Strategien des Korpusaufbaus und Nutzungsmöglichkeiten. In: K. Marx/H. Lobin/A. Schmidt (Hg.), *Deutsch in Sozialen Medien: Interaktiv - multimodal – vielfältig*. Berlin, Boston: de Gruyter, 345–348.

Barz, I. (2016): Die Wortbildung. In: A. Wöllstein (Hg.), *Duden in 12 Bänden: Band 4: Die Grammatik (9th Ausgabe)*, Mannheim: Bibliographisches Institut, 644–774.

Bickel, H./Gasser, M./ Häcki Buhofer, A./Hofer, L./Schön, C. (2009): Schweizer Text Korpus – Theoretische Grundlagen, Korpusdesign und Abfragemöglichkeiten. *Linguistik online* 39(3), 5–31, URL: http://dx.doi.org/10.13092/lo.39.474.

Brezina, V. (2018): *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.

Brezina, V./Meyerhoff, M. (2014): Significant or random? *International Journal of Corpus Linguistics*, 19(1), 1–28. https://doi.org/10.1075/ijcl.19.1.01bre.

Bubenhofer, N./Rothenhäusler, K. (2018): „Die Aussicht ist grandios!“ – Korpuslinguistische Analyse narrativer Muster in Bergtourenberichten. In: N. Eller-Wildfeuer/ P. Rössler/A. Wildfeuer (Hg.), *Alpindeutsch. Einfluss und Verwendung des Deutschen im alpinen Raum*. Regensburg: edition vulpes, 39–60.

Bubenhofer, N./Volk, M./Leuenberger, F./Wüest, D. (Hg.) (2015): *Text+Berg-Korpus (Release 151v01): Digitale Edition des Jahrbuch des SAC 1864–1923, Echo des Alpes 1872–1924, Die Alpen, Les Alpes, Le Alpi 1925–2014, The Alpine Journal 1969–2008*. Universität Zürich: Institut für Computerlinguistik.

Dargiewicz, A. (2012): Die Sprache lebt und verändert sich. Zu neuesten Tendenzen in der deutschen Wortbildung. *Acta Neophilologica*, 14(1), 61–76.

Donalies, E. (2005): *Die Wortbildung des Deutschen: Ein Überblick (2., überarbeitete Auflage)*. Tübingen: Narr.

Elsen, H. (2014): *Grundzüge der Morphologie des Deutschen (2. aktualisierte Ausgabe)*. Berlin: de Gruyter.

Engel, U. (2004): *Deutsche Grammatik (Neubearbeitung)*: München: Iudicium.

Fleischer, W./Barz, I. (2012): *Wortbildung der deutschen Gegenwartssprache (4., völlig neu bearbeitete Auflage)*. Tübingen: Niemeyer.

Gabrielatos, C. (2018): Keyness Analysis: Nature, metrics and techniques. In: C. Taylor/A. Marchi (Hg.), *Corpus Approaches to Discourse: A Critical Review* Milton: Taylor and Francis, 226–258.

Ganslmayer, C. (2012): *Adjektivderivation in der Urkundensprache des 13. Jahrhunderts: Eine historisch-synchrone Untersuchung anhand der ältesten deutschsprachigen Originalurkunden*. Berlin: De Gruyter.

Göhring, A./Volk, M. (2011): The Text+Berg corpus: an alpine french-german parallel resource. In: V. P. Mathieu Lafourcade (Hg.), *Actes des conférences TALN 2011 et Recital 2011*. Montpellier. https://doi.org/10.5167/uzh-48404.

Graën, S. (2004): *Die Raumadverbien des Mittelhochdeutschen (1050–1350):: Wörterbuch und Untersuchungen*. Göttingen: Georg-August-Universität Göttingen. http://hdl.handle.net/11858/00-1735-0000-0006-AEE4-4.

Grimm, J (o. J.): *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm*. Wörterbuchnetz des Trier Center for Digital Humanities (Hg.), Version 01/21, https://www.woerterbuchnetz.de/DWB.

Heinle, E.-M. (1985): Wortbildung des Neuhochdeutschen bis zur Mitte des 20. Jahrhunderts. In: W. Besch/O. Reichmann/S. Sonderegger (Hg.), *HSK: Vol. 2.2.*

Sprachgeschichte: Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung, Berlin: De Gruyter, 1911–1917.

Henzen, W. (1965): *Deutsche Wortbildung.* Halle (Saale): Niemeyer.

Kluge, F./Seebold, E. (2012): *Etymologisches Wörterbuch der deutschen Sprache: EBookPlus (25., aktualisierte und erweiterte Auflage)*. Berlin: de Gruyter. https://doi.org/10.1515/9783110223651.

Kupietz, M./Belica/C., Keibel/H./Witt, A. (2010): The German Reference Corpus DEREKO: A Primordial Sample for Linguistic Research. In: N. Calzolari /K. Choukri/B. Maegaard/J.Mariani/J.Odijk/S. Stelios Piperidis/M. Rosner/D. Tapias. (Hg.), *Seventh international conference on Language Resources and Evaluation (LREC)*, 1848–1854, URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Kupietz, M./Lüngen, H./Kamocki, P./Witt, A. (2018): The German Reference Corpus DeReKo: New Developments – New Opportunities. In: N. Calzolari/K. Choukri/C. Cieri/T. Declerck/S. Goggi/K. Hasida/H. Isahara/B. Maegaard/J. Mariani/H. Mazo/A. Moreno/J. Odijk/S. Piperidis/T. Tokunaga (Hg.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA), 4353–4360.

Lohde, M. (2006): *Wortbildung des modernen Deutschen: Ein Lehr- und Übungsbuch*. Tübingen: Narr.

Motsch, W. (2004): *Deutsche Wortbildung in Grundzügen*. Berlin: de Gruyter, 235f.

Nübling, D. (2016): Die nicht flektierbaren Wortarten. In: A. Wöllstein (Hg.), *Duden in 12 Bänden: Band 4: Die Grammatik (9. Auflage)*. Mannheim: Bibliographisches Institut, 579–643.

Nübling, D./Dammel, A./Duke, J./Szczepaniak, R. (2017): *Historische Sprachwissenschaft des Deutschen: Eine Einführung in die Prinzipien des Sprachwandels*. Tübingen: Narr.

Pojanapunya, P./Watson Todd, R. (2016): Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), 133–167. https://doi.org/10.1515/cllt-2015-0030

Posch, C./Rampl, G. (2020): Lima or Cima? Structure recognition and OCR in building the corpus of the Austrian Alpine Club Journal. *International Journal of Corpus Linguistics,* 25(4), 489–503.

Posch, C./Rampl, G. (2017): *Projekt Alpenwort – Korpus der Zeitschrift des Deutschen und Österreichischen Alpenvereins (1869–1998)*. Innsbruck: Institut für Sprachwissenschaft, Universität Innsbruck, URL: http://alpenwort.at.

Posch, C. (2022): *Digital Linguistics. Integrating Digital Humanities, Corpus Linguistics and Critical Discourse Studies*. Habilitationsschrift. Manuskript in Vorbereitung.

Rampl, G./Posch, C. (2019): *Alpenwort – Korpus der Zeitschrift des Deutschen und Österreichischen Alpenvereins (1869 – 1998) CQPweb Edition*. Innsbruck: Institut für Sprachwissenschaft, Universität Innsbruck, http://sprawi-cqpweb.uibk.ac.at/CQPweb/.

Schneider-Wiejowski, K. (2012): *Produktivität in der deutschen Derivationsmorphologie*. Dissertation. Bielefeld: Universität Bielefeld. https://d-nb.info/1021023833/34.

Schneider-Wiejowski, K. (2013): Sprachwandel anhand von Produktivitätsverschiebungen in der schweizerdeutschen Derivationsmorphologie. *Linguistik Online* 38(2), 79–90.

Schulz, D./Griesbach, H. (1972): *Grammatik der deutschen Sprache.* München: Max Hueber.

Simmler, F. (1998): *Morphologie des Deutschen: Flexions- und Wortbildungsmorphologie*. Berlin: Weidler.

Stefanowitsch, A. (2020): *Corpus Linguistics: A Guide to the Methodology // Corpus linguistics: A guide to the methodology.* Berlin: Language Science Press.

Wöllstein, A. (Hg.) (2016): *Der Duden: in zwölf Bänden; das Standardwerk zur deutschen Sprache ; Band 4. Die Grammatik,* 9., vollständig überarbeitete und aktualisierte Auflage). Berlin: Dudenverlag.

Weinrich, H./Thurmair, M. (2007): *Textgrammatik der deutschen Sprache (4., rev. Aufl.)*. Hildesheim: Olms.

Elisabeth Gruber-Tokić, Gerald Hiebel, Gerhard Rampl, Claudia Posch

Digital echo of mountains: Content indexing of alpine texts

1 Abstract

This paper is organized in two sections, whereas the first section gives a short overview of the interdisciplinary project *Semantics for Mountaineering History* (SEMOHI). The research project is currently working towards a methodology to extract information related to places, people and alpine activities (e.g. first ascent) from the Alpenwort corpus (Posch/Rampl 2017) and to represent this information using semantic web standards. The Alpenwort corpus is a text corpus that consists of the digitized ZAV volumes from 1868/69–1998 and is searchable by word type (Posch/Rampl 2017).

The second section deals with the methodology and concentrates on specific challenges of two major work packages and the approaches to solve them: One is the semantic annotation and semantic representation of places, people and alpine activities in TEI and CIDOC CRM (in RDF) and the other one deals with Named Entity Recognition (NER) and Named Entity Linking (NEL) for alpine discourse.

2 SEMOHI – Semantics of Mountaineering History

Mountaineering has become an activity corresponding to an image of conquest, human achievement and heroism. Mountains epitomize a seemingly unsurmountable hazardous barrier and therefore have been pulling people towards them for centuries. But not only the activity of mountaineering itself, also the accounts of these achievements have been fascinating people ever since. Mountains and the discussion about them as well as mountaineer-

ing have gained constantly rising significance and attention all over Europe. These discourses about mountaineering are essential to the activity itself (Rak 2007: 111). Mountaineering does not exist at all without the discourses around it – even less so without descriptions of places and people it evolves around: Scientific as well as non-scientific reports on ascents, summits and expeditions to the most secluded mountain ranges in the world were published in the Austrian Alpine Club Journal (*Zeitschrift des Deutschen und Österreichischen Alpenvereins* – ZAV) since the 1860s.

As a result, the ZAV is a unique text source for the German-speaking countries. The ZAV issues from 1868/69–1998 were transformed into a linguistically annotated alpine heritage corpus by the project *Alpenwort* (2014–2017) at the University of Innsbruck. The interdisciplinary follow-up project SEMOHI started in March 2017 and focuses its scientific research on the semantic annotation and representation of places, people and alpine activities detected in the Alpenwort corpus. SEMOHI's main objectives may be summed up as follows:

a) Development of a workflow for Named Entity Recognition and Named Entity Linking
b) Identification of place and person names
c) Identification of alpine activities (e.g. first ascents)
d) Semantic annotation and representation of alpine activities, locations and persons in machine-readable form
e) Open Access / online dissemination

In order to keep this paper's content clear and concise, the main objectives a) and d) are treated in detail.

3 Methodology and its implementation

The "Methodology" section concentrates on the methods and strategies chosen for implementation and deals with project-specific challenges.

The first subsection focuses on the implementation of semantic annotation of alpine activities and their semantic representation using CIDOC CRM. By means of appropriate text examples, it is shown which text related tasks have to be overcome in semantic annotation.

The second subsection outlines the methodology for the implementation of Named Entity Recognition / Named Entity Linking and also deals with special challenges that arise due to the partial historicity of the texts and the orthography.

3.1 Semantic annotation and representation of alpine activities

Semantic annotation can be used to define the meaning of words or phrases in a text. A word which is used to indicate a place or a person receives semantic annotation as a place or as a person. Events may be depicted by one or more phrases or sentences. The entities and the events that a text refers to can be extracted (e.g. persons, who ascended a mountain) and transformed to a semantic representation. For this depiction we need a formal definition of the entities and their possible relations. The ontology CIDOC CRM is perfectly suited for this purpose because it includes the necessary classes and relations to denote texts as well as activities of people at the places that are described. Figure 1 illustrates the main classes and relations used by CIDOC CRM. On top the figure shows the CIDOC CRM classes with their identifiers (Exx), as used in the definition of the CIDOC CRM. The rectangles in the bottom line show the corresponding designations that we used in our article for reasons of understandability. A fundamental distinction is made between:

- *E2 Temporal Entities*: periods, events or activities that have a temporal extent (*E52 Time Span*) – Events
- *E18 Physical Things*: Things that are constituted by material
- *E28 Conceptual Objects*: Things created by the human mind – Conceptual Objects
- *E39 Actors*: people participating in events – Person
- *E53 Places*: locations on earth or on other physical things – Locaction

All of these may have *E41 Appellations* like place names or person names and they are of a certain *E55 Type*, e.g. a place may be a mountain or a town.

The figure shows also the fundamental relations defined in the CIDOC CRM between the classes. e.g. *Temporal Entities* happen at *Places* and affect/involve *Physical Things* like a first ascent event happens at a mountain involves a specific mountaineering equipment like a rope.

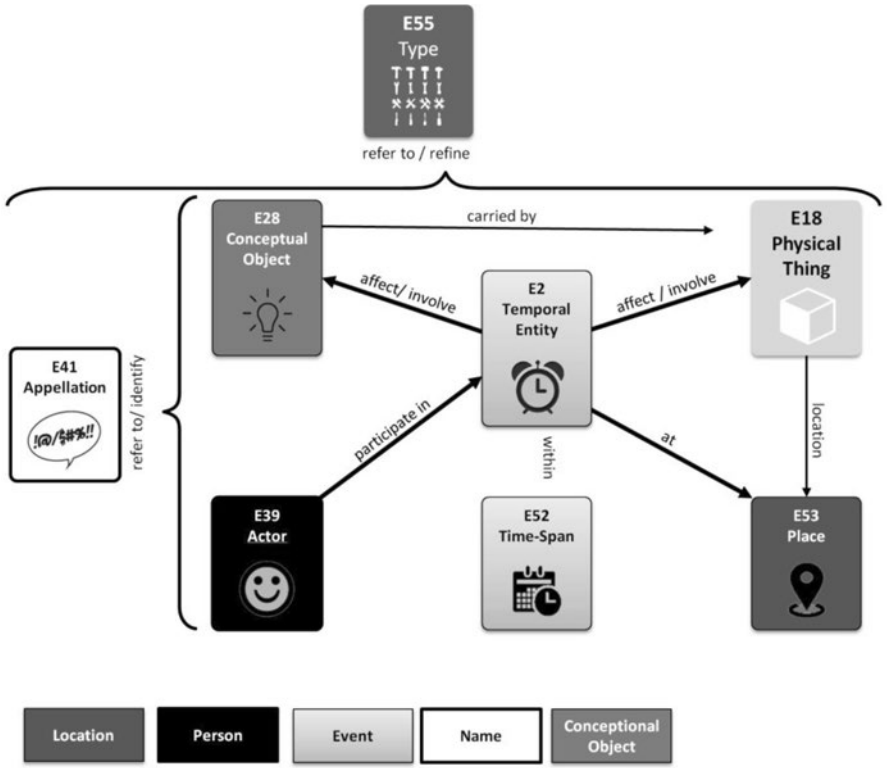


Fig. 1: Main classes and relations of CIDOC CRM

Fig. 2 shows the semantic representation of a text regarding the first ascent of a mountain when the classes of CIDOC CRM are applied: A text is composed (class: event) by an author (class: person). The text must be published to be available to the public and interested readers (class: event). The text itself (class: conceptional object) has a subject, in this case a first ascent of a mountain (class: event). This event took place at a certain date (class: event having a time-span), a specific place (class: location) and different participants (class: person), which both have names (class: name). As a result, it is possible to depict the core information of a text and demonstrate the network and relations between alpine activities, persons, places and the text itself.

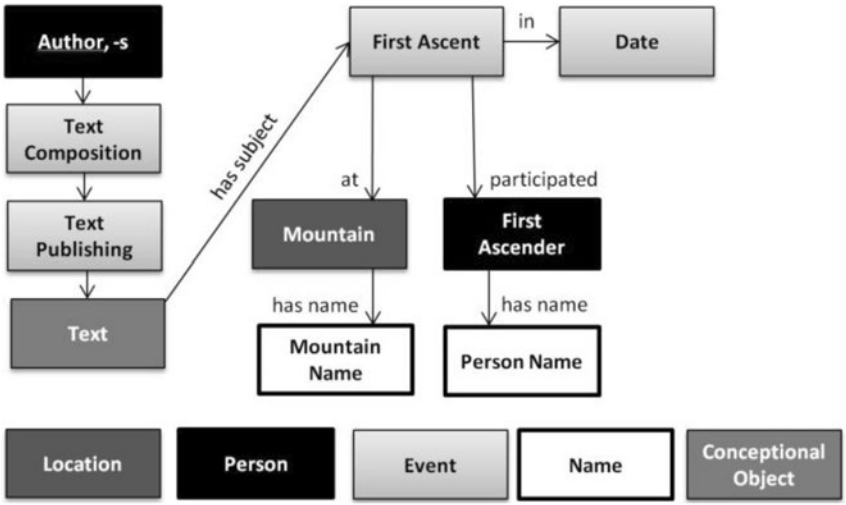


Fig. 2: Semantic representation of a network of alpine activities, persons, places and texts

As already mentioned, the identification and annotation of alpine activities in the texts constitute a central issue of this project. The questions first arising are: What is an event? What is needed to annotate it? – In the Text Encoding Initiative handbook the *event*-tag is defined as "data relating to any kind of significant event associated with a person, place, or organization" (TEI 2018). Furthermore, in order to fully designate an event also a date may be of major interest, although a date for the event itself is mandatory.

Nevertheless, during the process of semantic annotation of first ascent events two challenges occurred: 1. a text passage about a first ascent event includes two different dates and/or two different places. 2. a person name is missing within a text passage because the person was already mentioned before. The following text passage was chosen from the article *Bergsteigen in den östlichen Zillertaler Alpen* (av_1980_105_10 in Posch/Rampl 2017). It serves to illustrate the mentioned challenges of semantic annotation of first ascent events and the resulting approach to overcome them:

> 1856 stieg ein namentlich nicht bekannter Bauer aus der Prettau auf die Reichenspitze. Die erste touristische Ersteigung durfte 1866 Paul Grohmann, der große

> Dolomitenerschließer, für sich verbuchen. Ein Jahr zuvor, am 24. Juli 1865, gelang ihm die erste Ersteigung des Hochfeilers, der höchsten Erhebung in den Zillertaler Alpen. […] Ludwig Purtscheller stand 1893 auf dem Kuchelmooskopf, 3215 m.
>
> In 1856 a farmer from the Prettau, whose name is unknown, climbed to the top of the Reichenspitze. Paul Grohmann, the great Dolomite explorer, made the first tourist ascent in 1866. One year earlier, on 24 July 1865, he had succeeded in climbing the Hochfeiler, the highest peak in the Zillertal Alps. […] In 1893 Ludwig Purtscheller stood on the Kuchelmooskopf, 3215 m.

Let's start with a first ascent event that is perfectly suited for semantic annotation because all referred to entities are unique and therefore unambiguously assignable.

> Ludwig Purtscheller stand 1893 auf dem Kuchelmooskopf, 3215 m.
>
> **<event><persName>**Ludwig Purtscheller**</persName>** stand **<date when="1893">**1893**</date>** auf dem **<placeName>**Kuchelmooskopf**</placeName>**, **<measure>**3215 m**</measure> </event>.**

Within this text passage the first ascent of the *Kuchelmooskopf* (Tyrol, Austria) by *Ludwig Purtscheller* in 1893 is described. The following figure 3 shows the semantic representation using the CIDOC CRM classes.

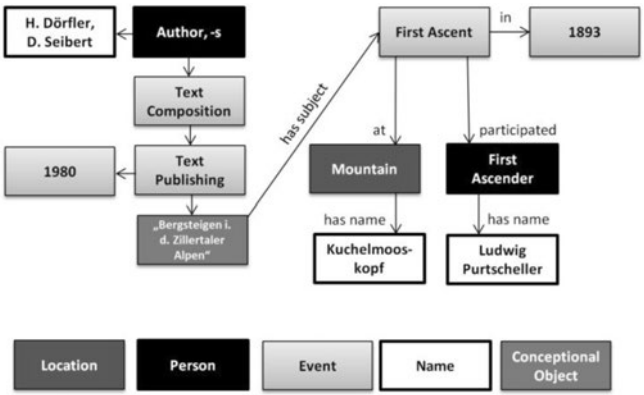


Fig. 3: Semantic representation of the first ascent event of the Kuchelmooskopf in 1893

As already mentioned, the same text passage includes an example which poses two problems: a) two different dates, b) two different places:

> 1856 stieg ein namentlich nicht bekannter Bauer aus der Prettau auf die Reichenspitze. Die erste touristische Ersteigung durfte 1866 Paul Grohmann, der große Dolomitenerschließer, für sich verbuchen.
>
> **<event> <date when="1856">**1856**</date>** stieg ein namentlich nicht bekannter Bauer aus der **<placeName>**Prettau**</placeName>** auf die **<placeName>**Reichenspitze**</placeName>**. Die erste touristische Ersteigung durfte **<date when="1866">**1866**</date> <persName>**Paul Grohmann**</persName></event>**, der große Dolomitenerschließer, für sich verbuchen.

The resulting semantic representation would be ambiguous, because two different dates refer to the same event. In addition, it is not obvious which place name refers to the first ascent event. A further problem addresses the missing person name *ein namentlich nicht bekannter Bauer* „a farmer unknown by name". As a consequence, the semantic annotation of an event and the related data must be modified and specified.

Although the text sample always deals with the same type of alpine activity, it is necessary to distinguish beween different alpine activities and other events. Therefore, special event types were defined (e.g. event type=„first_ascent", event type=„first_tour_ascent" etc.). Second, the annotated place names receive further information regarding the type of place it is.

> **<event type=„first_ascent"><date when="1856">**1856**</date>** stieg ein namentlich nicht bekannter Bauer aus der **<placeName type=„place">**Prettau**</placeName>** auf die **<placeName type=„mountain">**Reichenspitze**</placeName></event>.**
> **<event type=„first_tour_ascent">** Die erste touristische Ersteigung durfte **<date when="1866">**1866**</date> <persName>**Paul Grohmann**</persName>**der große Dolomitenerschließer, für sich verbuchen **</event>**.

This modification permits the following semantic representation: Due to the two different types of events it is possible to assign the dates and the places correctly (see fig. 4).

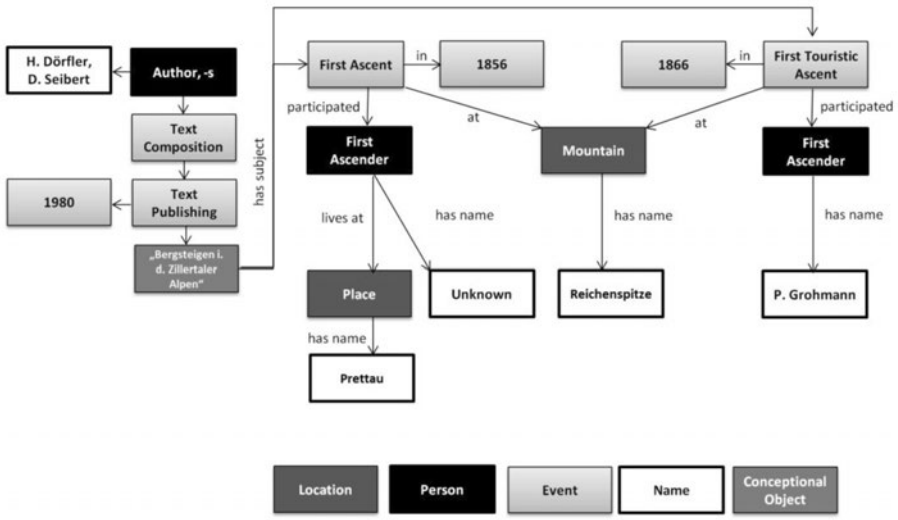


Fig. 4: Challenges regarding the semantic annotation of events: two dates, two places.

Another issue concerns co-reference resolution: Within the following *event*-tag, the person name is missing because it was already mentioned in the previous paragraph. The author anaphorically refers to the first ascender of the *Hochfeiler* (Tyrol, Austria) with the personal pronoun *ihm* 'him'.

> Ein Jahr zuvor, am 24. Juli 1865, gelang ihm die erste Ersteigung des Hochfeilers, der höchsten Erhebung in den Zillertaler Alpen. […] Ludwig Purtscheller stand 1893 auf dem Kuchelmooskopf, 3215 m.

> **<event =„first_ascent“>**Ein Jahr zuvor, am **<date when="1865-07-24“>**24. Juli 1865**</date>**, gelang ihm die erste Ersteigung des **<placeName type=„mountain“>**Hochfeilers**</placeName></event>**, der höchsten Erhebung in den **<placeName type=„mountain range“>**Zillertaler Alpen**</placeName>**.

This results in the following misleading semantic representation, because it is impossible to assign the correct name to the first ascender (fig. 5):

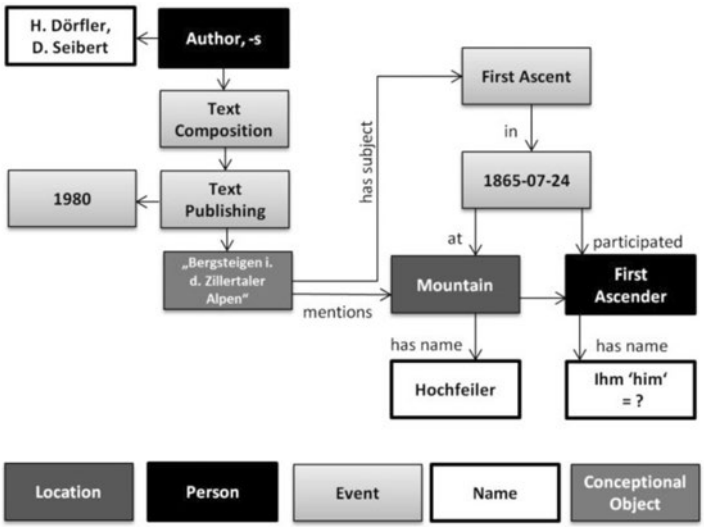


Fig. 5: Challenges regarding the semantic annotation of events: missing person name

In order to solve the problem of co-reference and also prevent further obstacles regarding place names ID's were established. Due to the fact that several texts are analysed together or merged, each local ID of a text must receive a global ID. The chosen way to do this is concatenate a text identifier with the place identifier. Every annotated place name and person name receives an ID. Thus, the semantic annotation of an event changes as follows:

> 1856 stieg ein namentlich nicht bekannter Bauer aus der Prettau auf die Reichenspitze. Die erste touristische Ersteigung durfte 1866 Paul Grohmann, der große Dolomitenerschließer, für sich verbuchen. Ein Jahr zuvor, am 24. Juli 1865, gelang ihm die erste Ersteigung des Hochfeilers, der höchsten Erhebung in den Zillertaler Alpen.

> **<event type=„first_ascent"><date when="1856">**1856**</date>** stieg ein namentlich nicht bekannter Bauer aus der **<placeName type=place**

xml:id=place_1>Prettau**</placeName>** auf die **<placeName type= mountain xml:id=place_2>**Reichenspitze**</placeName></event>**.
<event type=„first_tour_ascent“> Die erste touristische Ersteigung **<place type=„mountain“ ref=“#place_2“>** durfte **<date when=“1866“>**1866**</date> <persName xml:id=“person_1“>**Paul Grohmann**</persName>**der große Dolomitenerschließer, für sich verbuchen **</event>.**
<event type=„first_ascent“>Ein Jahr zuvor, am **<date when=“1865-07-24“>**24. Juli 1865**</date>**, gelang **<person ref=“#person_1“>**ihm**</person>** die erste Ersteigung des **<placeName type=„mountain“ xml:id=“place_3“>**Hochfeilers**</placeName>** der höchsten Erhebung in den **<placeName type=„mountain range“ xml:id=“place_4“>**Zillertaler Alpen**</placeName></event>.**

According to this modification of the semantic annotation it is possible to represent the semantic network of first ascent events using CIDOC CRM classes, as shown in fig. 6: The representation should be read from the left, starting with the authors, their names, the text composition and its publication date. Farther to the right the subject of the text itself is indicated. Without reading the whole paragraph it is possible to obtain an overview of its content.

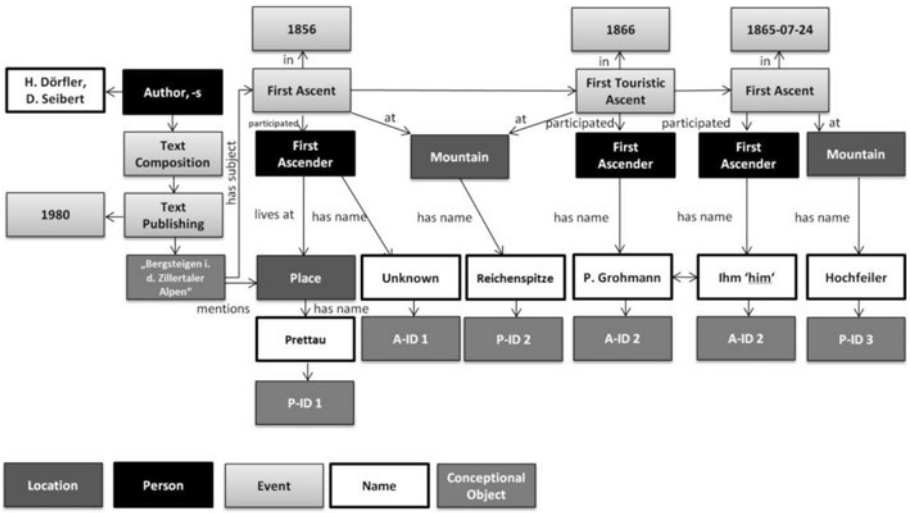


Fig. 6: First ascent events using IDs for person names and place names

3.2 Named Entity Recognition and Named Entity Linking

Although the articles in the Alpenwort corpus differ in content, they use a specific element for orientation and identification in alpine regions: names or named entities (NE). Place names are not only subject to research in scientific disciplines such as onomastics, but are also of central interest for the topics already mentioned, such as alpinism, mountain paths or marketing and tourism.

The main approach to achieve automated Named Entity Recognition (NER) (Riedl/Padó 2018; Benikova et al. 2015) and Named Entity Linking (NEL) is matching the entries of a gazetteer of place names (Berman et al. 2016) with emphasis on alpine regions with the named entities identified within the Alpenwort corpus. To increase the prospect of a positive outcome of the automated place name recognition within the Alpenwort corpus, the creation of an extended gazetteer for the alpine area is critical. For this reason already existing gazetteers and digital sources with different spatial dimension and density of place names were incorporated to one single gazetteer. This gazetteer now contains data from the national mapping institutions of Austria and Switzerland, the project *Flurnamendokumentation Tirol* ('Field Name Survey in the Tyrol'), maps from the Austrian Alpine Club (*Alpenvereinskarten*) as well as internet sources like geonames, wikidata and OpenStreetMap (OSM).

The process of place name recognition and linking to gazetteer entries is divided into several steps. First, it is essential to perform the recognition of place names and their linking to gazetteer entries manually to create a gold standard that may be used for evaluation. Seven different ZAV articles were selected: 1873, 1965, 1980, 1981, 1990, 1994 and 1997. Most chosen texts originate from the recent past of the journal due to the fact that the spelling of place names within older articles is fundamentally different from the current spelling (e.g. *Thal* vs. *Tal* "valley")[1]. The second step concerns automated named entity recognition and automated linking. Finally, the results of the automated NER/NEL are compared with the manual results of the gold standard articles and the process is evaluated and improved.

1 The problem of spelling variation and differing orthography within the articles has to be addressed elsewhere. A detailed examination would exceed the scope of this paper.

In order to achieve automated place name recognition for German texts different challenges arise: The first challenge concerns the entries within the gazetteer. The created gazetteer consists of 5.9 million place names and a lot of them synonymous with appellatives (e.g. *Boden* 'ground, floor', *Hütte* 'hut', *Tal* 'valley', *Weg* 'path'). Respectively, only the analysis of the local context provides an exact key whether the word found within a text is an appellative or a place name. For example, the German word *Boden* may either refer to the appellative 'ground, floor, bottom' or designate the name of a mountain village in Reutte (Tyrol, Austria).

Another problem regards the declension of place names in German (NOM: der *Kuchelmooskopf*, GEN: des *Kuchelmooskopfes* etc.). As a result, declined place names cannot be string-matched to the entries of the gazetteer as the writing is different. Furthermore, the declined version of a place name could exist randomly and name a completely different object or place: in Austria, there is a mountain called *Gabler* (GEN: *Gablers*; *den Gipfel des Gablers*) on the one hand. In Germany occurs a place called *Gablers* (NOM: *Gablers*, GEN: *Gablers*). Thus, the declined place name *Gablers* would be correctly identified as named entity but linked to the wrong place. To solve this problem, all possible German declensions of a gazetteer name (consisting of one and two words) are simulated and matched with the word(s) in the text.

Further difficult tasks concern the dropping of words (e.g. *Geiger* instead of *Großer Geiger*, *Venediger* instead of *Großvenediger*) and alternative spellings which are not included in the gazetteer (*-spitz* instead of *-spitze*), dialectal variation, outdated spellings or spelling mistakes.

Regarding word dropping, the names of the gazetteer are modified dropping the part of the name that is less significant. The significance of a word as a name inside an article is calculated through the ratio of the number of appearances in one text in relation to the appearance of the word in the whole corpus. For example, normal German words (*und* 'and', *sie* 'she', *auch* 'too, as well, also') achieve significance scores close to zero while specific names reach values between 50 and 15,000. The already mentioned mountain name *Gabler* achieves 54.71 significance score or *Kuchelmooskopf* accomplishes 7,773.65 significance score. This significance score is used in various ways throughout the NER/NEL workflow to identify names and exclude false positives.

Another challenge is depicted by the disambiguation of place names. After the identification of a place name within a text it is necessary to link it to the correct gazetteer entry. This process becomes very complicated, if there are synonymous entries within the gazetteer: the place name *Roßkopf* refers to 50 different gazetteer entries which are located either in Austria or in Germany. This is almost impossible for human named entity linkers let alone machines. As a consequence, this problem is approached with the method of first differentiating between a global and a local context in the text. The method was chosen because most articles deal with a very specific geographic area (core region). This consideration is based on the fact that most titles in the table of contents already include a specific place name. Within those core regions, place names of local significance are used as well.

Local significance is characterized by the fact that the name makes the location only identifiable within the local context. The main reason is that the place name will occur several times in the gazetteer like in the example of the Roßkopf. Specific place types are excluded from searches on a global scale like the names of buildings or microtoponyms as they produce many false positives in a global context. Place names without local significance have to be either of global significance or have to be accompanied by place names with global significance to establish an according context.

4 Summary

The main objective of this article is the representation of the interdisciplinary project SEMOHI and the discussion of text-specific challenges regarding semantic annotation and semantic representation of alpine texts. Standards like TEI and CIDOC CRM are used for semantic annotation and representation are explained. In addition, the scientific approach to perform automated Named Entity Recognition and Named Entity Linking in the context of this alpine text corpus is outlined.

Future work will have to improve the quality of the automated NER and NEL taking into account machine learning methodologies. For the semantic representation we will aim for an RDF representation that can be accessed using a SPARQL Endpoint (W3C 2008).

Bibliography

Alexandria Digital Library Gazetteer (2004): Santa Barbara CA: Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Copyright: UC Regents. http://legacy.alexandria.ucsb.edu/gazetteer/, 04.07.2019

Benikova, D./Yimam, S. M./Biemann, C. (2015): GermaNER: Free Open German Named Entity Recognition Tool. In: *Proceedings of GSCL*. Essen, Germany, pp. 31–38.

Berman, M. Lex/Mostern, R./Southall, H. (2016): *Placing Names.Enriching and Integrating Gazetteers*. Bloomington, Indianapolis: Indiana University Press.

CIDOC CRM (2018): *Definition of the CIDOC Conceptual Reference Model*, http://www.cidoc-crm.org, 24.01.2019

Doerr, M. (2003): The CIDOC CRM an Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24(3), 75–92.

Flurnamen Tirol (2012): *Flurnamendokumentation im Bundesland Tirol*. http://onomastik.at/content/flurnamendokumentation-im-bundesland-tirol, 04.07.2019

Hiebel, G./Doerr, M./Eide, Ø. (2017): CRMgeo: A Spatiotemporal Extension of CIDOC-CRM. *International Journal on Digital Libraries Special Issue* 18, 271.

Hiebel, G./Rampl, G./Posch, C./Gruber, E./Zangerle, Eva (2017): Bergnamen – Bergwelten. Toponymie im (Kon-)Text. In: *Mainzer Namentagung 2017: Toponyme - eine Standortbestimmung*. published online /Internetpublikation, p. 10.

Iso 19112 (2003): *ISO 19112 Spatial referencing by geographic identifiers*. https://www.iso.org/standard/26017.html, 04.07.2019

Knoblock, C./Szekely, P./Ambite, J. L./Goel, A./Gupta, S./Lerman, K./Muslea, M./Taheriyan, M./Mallick, P. (2012): Semi-automatically Mapping Structured Sources into the Semantic Web. In: *Proceedings of the 9th International Conference*

on The Semantic Web: Research and Applications. Berlin/Heidelberg, Springer, 375–390.

Lampe, K.-H./Krause, S./ Doerr, M. (2010): *Definition des CIDOC Conceptual Reference Model. Version 5.0.1 autorisiert durch die CIDOC CRM Special Interest Group (SIG)*. ICOM Deutschland, Beiträge zur Museologie, Band 1. http://www.cidoc-crm.org/sites/default/files/cidoccrm_end.pdf, 18.11.2020

Posch, C./Rampl, G. (Hg.) (2017): *Alpenwort – Korpus der Zeitschrift des Deutschen und Österreichischen Alpenvereins (1869 – 1998)*. Abteilung Sprachwissenschaft am Institut für Sprachen und Literaturen, Universität Innsbruck, URL: http://alpenwort.at DOI: http://doi.org/10.5281/zenodo.1243678, 18.11.2020

Posch, C./Rampl, G. (2016): Alpenwort. Korpus der Zeitschrift des Deutschen und Österreichischen Alpenvereins. In: *VERBAL-NEWSLETTER - Zeitschrift des Verbandes für Angewandte Linguistik* XVII/1/2016, 7.

Rak, J. (2007). Social Climbing on Annapurna: High-altitude Mountaineering Narratives. *ESC*, 33(1-2), 109–146.

Rampl, G./Posch, C. (Hg.) (2020): *Alpenwort – Korpus der Zeitschrift des Deutschen und Österreichischen Alpenvereins (1869 – 2010)*. Abteilung Sprachwissenschaft am Institut für Sprachen und Literaturen, Universität Innsbruck, URL: http://sprawi-cqpweb.uibk.ac.at/CQPweb/, 18.11.2020.

Riedl, M./Padó, S. (2018): A Named Entity Recognition Shootout for German, In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Shaw, R. (2016): Gazetteers Enriched: A Conceptual Basis for Linking Gazetteers with Other Kinds of Information. In: M. L. Berman/R. Mostern/H. Southall (Hg.): *Placing Names. Enriching and Integrating Gazetteers*. Bloomington, Indianapolis: Indiana University Press, 51–66.

TEI (2018): Text Encoding Initiative, http://www.tei-c.org/, 04.07.2019.

W3C (2008): SPARQL Query Language for RDF, https://www.w3.org/TR/rdf-sparql-query/,18.11.2020.

Karoline Irschara[A], Claudia Posch[B], Birgit Waldner[C], Anna-Lena Huber[D], Bernhard Glodny[E], Leonhard Gruber[F], Stephanie Mangesius[G*]

Building the MedCorpInn corpus: Issues and goals

1 Introduction

In this paper we give a comprehensive literature review as well as an overview of issues and procedures concerning the ongoing project *MedCorpInn – Retrospective Intersectional Corpuslinguistic Analysis of Radiology Reports of Innsbruck Medical University*, which was initiated in 2019 as a collaborative interdisciplinary project between the University of Innsbruck and the Medical University of Innsbruck. The project is situated within the following fields of research: Digital Linguistics (i.e. corpus building as well as CADS with a focus on language and gender) and Gender Medicine. In this contribution we first give a brief overview of the literature on health care and text corpora and then summarize the process of building the text corpus MedCorpInn. Last but not least a preliminary discussion of CADS and Gender Medicine research questions follows. As a premise we assume medical reports as discursive, linguistic events which are influenced by and influencing social factors and thus are a data basis which can be studied with a CADS-approach as well as a Gender Medicine lens. As discourse medical reports are susceptible for systemic biases, which may be produced and reproduced on the surface and/or on the structural level of language. Such biases could result in real life discrimination the research of which is a present desideratum in Gender Medicine.

* A, B, C Institut für Sprachwissenschaft, Universität Innsbruck; C, E, F, G Universitätsklinik für Radiologie; D Universitätsklinik für Augenheilkunde, Medizinische Universität Innsbruck.

2 Healthcare and text corpora – an overview of the literature

Research on healthcare communication in linguistics has been established as a qualitative method since the 1980s, when especially doctor-patient communication was studied (Waitzkin 1984; Collins, Peters and Watt 2011). Since the end of the 1990s, several other communicative practices have been researched as well, e.g. interactions of nurses (Kelly 1998; Crawford, Brown and Nolan 1998), physiotherapists (Parry 2004), pharmacists (Pilnick 1998) and psychotherapists or psychiatrists (Lewis 1995; Morris 1995; Buchholz 1998; Koerfer and Martens-Schmidt 2000). Alongside research on particular features of medical terminology (Johnson 1999; Xu et al. 2009; Jiang et al. 2011), multimodal approaches to medical communication have been established, including Corpus Linguistics, (Critical) Discourse Analysis and Metaphor Analysis (Crawford, Brown and Harvey 2014: 76; Demjén 2020).

However, there is almost no substantial CADS research focusing on clinical texts, let alone clinical texts in German and there are no large corpora containing such texts (e.g., most of the NLP corpora contain about 3,000–4,000 documents). The only such project is the Stockholm Electronic Patient Record Corpus (SEPR Corpus) which contains over one million patient records from over 2.000 clinics. In one study parts of this corpus were compared to the Swedish corpus PAROLE in order to describe differences and similarities such as domain specific compounds, abbreviations, narratives etc. (Dalianis et al. 2006). The bilingual (French-Spanish) EMCOR corpus includes medical research papers, case reports, patient information and review articles and has mainly been used to create terminology resources to assist in the translation of medical texts (Vila and Trigo 2012). Only isolated studies were using corpora applying conversation and discourse analysis on medical language, such as the ClInt (Clinical Interview) Corpus for Bilingual Spanish-Catalan (Vila Rigat et al. 2010) or the NHS (National Health Service) Direct Corpus (Adolphs et al. 2004) for English. However, there are no comparable resources and studies for German.

Those studies using German medical text data focus on developing NLP applications for information extraction and usually are not analyzing the data itself. What is more, existing clinical corpora are usually only accessible to

scientific staff for the duration of a project and data is not available afterwards or for other researchers for obvious privacy reasons (Hellrich et al. 2015). The first small German-language medical corpus FRAMED (FReiburg Annotated MEDicine text corpus) was completed in 2004 by Hahn et al. (2004) and entails 300 medical documents of different types. The text types include among others discharge summaries, surgery, pathology and histology reports as well as non-clinical medical expert texts. While the FRAMED corpus is not publicly available, the tools developed for sentence splitting, tokenization and POS tagging of German clinical texts trained on the corpus are shared (Hellrich et al. 2015).

In 2014, Krieger et al. had put together a corpus of 544 clinical documents from various medical domains, such as echocardiography, EEG, lung function, chest X-rays, e.g. to perform information extraction via a hybrid parsing and relation extraction strategy (Krieger et al. 2014). Fette et al. (2012) integrated about 200,000 clinical reports of the domains echocardiography, lung function, chest X-ray and bicycle stress test into a clinical data warehouse after performing tokenization and POS-tagging of all texts from a particular domain and training a labeling algorithm for automated information extraction. Recently, Hahn et al. (2018) presented 3,000 PA, a German-language clinical document corpus composed of approximately 3,000 EPRs from three different clinical sites. To circumvent IPRs and privacy constraints, Lohr et al. (2019) created JSYNCC, which is the first and only publicly available corpus of German clinical language, consisting of 867 fictitious medical documents written by experts.

Studies using medical data for information retrieval foci are rapidly evolving. Toepfer et al. (2015) used 520 German transthoracic echocardiography reports for terminology construction and development of a generic ontology based on an information extraction algorithm. They also performed terminology reordering and mapping on a reference guideline for German echocardiography reports. Richter-Pechansky et al. (2018) used a collection of 180,000 notes from cardiology to generate a prototype for German medical text de-identification. In 2016, 450 surgery reports were assembled to build language models in view of metadata from two German medical thesauri (Lohr and Herms 2016). In the same year, a semantic annotation scheme (with tags

for e.g. body_part, tissue, body_fluids, etc.) was developed for 1,725 medical documents from the nephrology domain, comprising discharge summaries and clinical notes. For the annotation a semi-automated annotation using additional resources such as UMLS was used. (Roller et al. 2016) Roller et al. (2018) also used 626 clinical notes from the nephrology domain for the detection of named entities and relations using Conditional Random Fields (CRF), Support Vector Machines (SVM) and Convolutional Neural Networks (CNN). Also in 2018, Kara et al. created a small gold standard corpus of 55 nephrological text documents composed of discharge summaries and clinical notes as well, including POS and dependency annotations. To further evaluate parsing accuracy, students familiar with the nephrology data created a collection of fictitious clinical notes and discharge summaries. They also created fictitious notes in the subdomains of surgery, cardiac rehabilitation, discharge, internal medicine and relocation so avoiding legal constraints of records containing personal health information of real patients (Kara et al. 2018). Kreuzthaler et al. used 1,696 de-identified German-language clinical in and outpatient discharge letters from the dermatology department of an Austrian university hospital for the detection of sentence boundaries and abbreviations using SVM with linear kernels (Kreuzthaler and Schulz 2015). In 2016, they used the same corpus to focus on automated abbreviation detection, merging statistical and dictionary-based disambiguation strategies (Kreuzthaler et al. 2016). Becker et al. (2019) used 2,513 German clinical colorectal cancer notes from electronic health records abstracted by a human to create a NLP pipeline able to identify specific guideline-based patient information and subsequent annotation with UMLS for retrospective evaluation of the therapy recommendation. Bretschneider et al. (2013) used 2,713 de-identified reports of radiology examinations of lymphoma patients from the University Hospital Erlangen to build a reference corpus for the development of a method able to automatically detect pathological findings by classifying the sentences in the radiology reports as either pathological or non-pathological. As an extension of their work, they annotated the sentences classified as pathological with the corresponding RadLex IDs and linked the pathological finding with the corresponding position in the radiologic image. As the German RadLex version contains only about 6,300 terms, a process to extend the lexicon with

vocabulary and pathological classification was developed (Bretschneider et al. 2013). A knowledge-based extraction of measurement-entity relation using two datasets, one consisting of 2,584 radiology reports of lymphoma patients and 6,007 German radiology reports with computed tomography (CT) imaging modality was presented by Oberkampf et al. (2014). Krebs et al. (2017) assembled a corpus of 3,000 chest X-ray reports for the implementation of a semi-automatic terminology generation algorithm. Usually, also NLP tasks in radiology consist in information extraction tasks such as the detection of uniform recommendations according to medical guidelines, the mentioning of certain contents as indicators of quality or the extraction of epidemiological data (Jungmann et al. 2019).

The short overview of the literature on language corpora and healthcare showed that such corpora still are not very common in linguistic healthcare research. The corpus MedCorpInn wants to be a step towards closure of this gap. In the following section we therefore will make transparent some of our basic decisions in corpus building and data processing. We will also address ethical aspects and outline further steps to be taken.

3 Designing and building the MedCorpInn corpus

The first major goal of the project is to build a large annotated corpus of radiology reports to provide a solid data fundament for the intended research in CADS and Gender Medicine. Radiology reports were selected because they are an essential element of everyday communicative practices between radiologists and referring doctors and provide useful insights into internal clinical discourse and the close cooperation between the two universities made it possible to obtain this kind of data. A small pilot study was completed in preceding projects, resulting in the linguistically annotated corpus KARBUN which consists of 100,000 radiology reports written in German. The workflow developed in the pilot study was used as a best practice model for the extended corpus of medical reports in MedCorpInn.

3.1 Data and metadata

The corpus MedCorpInn consists of 5,002,933 written reports in German from the Departments of Radiology and Neuroradiology at the University Hospital of Innsbruck (2,540,022 female patients; 2,440,474 male patients). It contains reports from every day of the years 2007–2019, which were exported from the clinical system in a plain text format with structured metadata and a basically unstructured text for the reports. The reports serve as a legal record which documents and interprets different imaging procedures, such as ultrasound, computed tomography, magnetic resonance imaging, positron emission tomography, angiography, X-ray, fluoroscopy etc. There are more than 300 types of examination by which the corpus can be divided. They indicate the mode of examinations as well as the anatomic object or region that is being examined (e.g. whole-body CT, shoulder CT, knee CT etc.). Often, the reports would suggest further steps to be taken and refer patients to additional examinations. The data is enriched by an extensive amount of metadata, comprising 39 different categories. They indicate demographic and personal information (e.g. the patient's age, gender, occupational status, type of insurance, provenience;) as well as information regarding different medical procedures (e.g. admission type, medical indication, time frame etc.). Some of these metadata categories are set up as selection criteria which makes it possible to subsequently filter queries according to these criteria. Such selection criteria are for example gender, age group, type of insurance, and the provenience of the patients as well as gender of the doctors and the mode of examination.

3.2 Anonymization and data cleaning

To ensure patient and doctor anonymity, names and IDs were already removed before extracting the data from the clinical information system. Several additional anonymization processes were conducted with which sensitive mentions in the data were replaced, removed or manipulated (Šuster, Tulkens and Daelemans 2017). Some categories such as age or occupation are transformed into broader categories or groups. Furthermore, data pre-processing revealed that some doctors' and patients' names occasionally appear in the report texts. Anonymization was carried out by using RegEx

replacements, since physicians' names are typically preceded by academic grades and/or positions within the clinic. There are rarely patient names in the text and when there are the names are preceded by "Herr" ('Mr.') or "Frau" ('Ms.'). Further corpus linguistic statistics were used in preprocessing to highlight remaining potential names.

The report texts in the free text field are of varying length, many rather short. Sometimes they contain various kinds of subheadings which do not constitute part of the text, but rather are similar to structural elements. The use of these structural type elements is very inconsistent, as at times there are no subheadings or the same subheading is used more than once. This creates problems with tokenization, sentence boundary disambiguation and later also in linguistic statistics. Therefore, the subheadings were unified as more general categories and set up as paragraph types. For example, the terms *Zuweisungsmodus* ('referral mode'), *Zuweisungstext* ('referral text') and *Zuweisungsgrund* ('reason for referral') are consolidated in the term *Zuweisung* ('referral'), since they all point to the reason why a patient was referred to radiological examination. The individual reports also include patient IDs, which were pseudonymized: Each patient was assigned a new ID which allows patient follow ups to be possible for further research.

3.3 Linguistic annotation

Another project goal is POS-tagging the report texts. Standard POS-taggers have trouble with the recognition of medical terms since they are usually designed for data work with standard German texts (often media texts). Particularly Latin terms, (pseudo-)Latin terms with parts of German morphology and medical, ad-hoc and non-standard abbreviations will not be identified by standard taggers. Work with the best practice data showed that standard POS-taggers are prone to interpret the abundant non-standard abbreviations in the texts as sentence endings, which might bias statistical evaluations. Also, frequently occurring ellipses and incomplete sentences showed to cause problems in the course of the tagging process. Thus, we will strive for using state-of-the-art machine learning approaches combined with the use of project-specific thesauri of medical abbreviations etc.

We are particularly interested in tagging and extracting measurement results in the texts. Measurements of lengths or diameters of tumors, injuries, organs etc. occur frequently in the corpus. This will be helpful when analyzing the number distribution of the corpus as well as the precision of the measurements (e.g. with or without decimal points, with or without intensifiers or mitigators) and if this is somehow connected to social categories in the metadata. Additionally, the distribution of numbers within the corpus should be in concordance with the Newcomb-Benford law (Newcomb 1881; Benford 1938). This law describes the distribution of digits or the frequency of their occurrence in many naturally occurring collections of numbers. If this distribution is not confirmed, this could point to certain singularities in the data which then can be further studied.

3.4 Ethics aspects

The data management of the current project is bound to comply with data confidentiality (§6 DSG, current version) as defined in the approval by the ethical review committee of the Medical University of Innsbruck. The project also adheres to the advanced FAIR data management principles for health data (FAIR-Health), as proposed by Holub et al. (2018). The authors focus on several challenges of sensitive data usage in medical research, especially on reproducibility and privacy protection. Privacy-enhancing technologies are applied with personal data before this data can be used for further research purposes; also, data has to be constantly controlled and checked for remaining anonymization deficiencies.

For confidentiality reasons, the full corpus cannot be made available open access. It is currently stored locally, access-protected and the use of the linguistic data is subject to individual approval. However, all tools and non-confidential data created for this project will be made available for reuse.

4 Research with MedCorpInn

4.1 The study of linguistic bias in healthcare

Recent studies suggest that systemic bias and discrimination may occur on a linguistic surface level and can be detected in linguistic communication (Isaac et al. 2011; Menz and Lalouschek 2006; Trix and Psenka 2016). It is important to note that biases potentially leading to such discriminations must not be perceived as something a person actively and deliberately does for base motives. Biases are rather unintentional, unconscious presumptions, which often also underlie socio-political structures that cannot be changed by the individual and lead to discriminatory practices: "Much of the gender discrimination that appears to take place is almost unconscious, reflecting the norms of the society in which both the health worker and the patient are biased" (Govender and Penn-Kekana 2008:99). In the project MedCorpInn we conceive of language as a powerful means through which such implicit biases and hence discrimination can be inflicted and reproduced. Regarding healthcare contexts, it has been shown that bias introduced by the treating clinician contributes to healthcare inequalities, and that the language used to refer to patients often reflects this: A randomized vignette study in 2018 revealed that stigmatizing language used to describe patients in medical records can influence subsequent physicians-in-training regarding their attitudes towards patients (Goddu et al. 2018). The effect of lexical choices towards physician's attitudes has also been examined by Kelly and Westerhoff (2010), who found that "referring to an individual as 'a substance abuser' vs. 'having a substance use disorder' evokes different judgments about behavioral self-regulation, social threat, and treatment vs. punishment" (202) and thus reinforces negative and stigmatizing stereotypes. Thus, discursive strategies applied can elicit systematically different judgements, even among highly trained mental health professionals.

The impact of social factors in doctor-patient interaction has been researched extensively and biases have been found to be pervasive: For example, some studies indicate that patients with fewer economic possibilities are less involved in conversations than patients with more (Willems et al. 2005); ethnic minority patients are less likely to be recommended certain treatments

than non-ethnic minority patients (Ibrahim et al. 2003; Rucker-Whitaker et al. 2003). It was also found that BPOC receive shorter visits than white patients and that the interactions are less patient-centered (Cooper et al 2012; Peck and Denney 2012). Racial and socio-economic biases also considerably contribute to patient dissatisfaction with their visits (Johnson et al. 2004; Street et al. 2010). Physicians may tend to change their institutional procedures with non-native speakers (on the topic of intercultural communication in the health context see for example Paternotte et al. 2016), e.g. non-medical issues and bureaucratic negotiations are discussed more intensively, while medical topics are neglected (Valero-Garcés 2002). Furthermore, it has been shown that physicians might use a more directive communication style when speaking with non-native patients which impacts efforts of shared decision making. However, such communication strategies might also conflict with patients' expectations depending on their social and/or cultural background (Roberts 2006). Bührig and Meyer (2015) point out that not only diverging linguistic and cultural backgrounds, but also systemic factors such as time pressure are underlying difficulties when communicating with non-native speakers.

Also linguistic gender bias has been investigated both with regards to the doctor's and the patient's gender (West 1984; Borges 1986; Wodak et al. 1990; Maynard 1991; Pauwels 1995). Although this subject of investigation entails a long tradition of research, gender biases appear to be pervasive. Through a CDA-study of doctor-patient consultations, Hedegaard et al. (2014) found that male patients were constructed as competent, while female patients were characterized as fragile through gender stereotypical communication by health professionals; this might have an impact on how patients' statements are conceived. Thus, stereotyped assumptions and expectations are consolidated and reproduced. This study also found that male patients were more likely to describe their issues with performance-oriented statements, while female patients made more use of emotional-oriented statements which were both reinforced by gender-stereotypical questions (Heedegard et al. 2004:14).

It was reported that gender correlates with different linguistic representations of pain, e.g. women tend to downplay their pain and to describe themselves as being able to bear the pain, while men are more likely to overrate their pain and view themselves as mastering it (Menz 2010: 9). This is also

relevant with regards to gender specific differences in support-seeking behavior: Empirical research shows a reluctance of men of different ages and ethnic/cultural backgrounds to seek help from health professionals (Addis and Mahalik 2003), which might be due to traditional gender norms and discourses, e.g. help-seeking is seen as inherently non-masculine or feminine, and therefore does not conform to the stereotypical notion of men being in control of their pain on their own (Charteris-Black: 163).

4.2 Corpus-assisted gender linguistic research with MedCorpInn

As an example of how language use patterns might interact with social factors we would like to mention a case study from the preceding project *KARBUN*. Thus, the texts of our best practice corpus (100,068 texts, 7.8 million tokens) were split into two subcorpora depending on whether the reports were written on female or on male patients. Key items were calculated comparing subcorpus_female with subcorpus_male in a local installation of CQPweb (Hardie 2012). For the identification of keywords, LogRatio statistic was used with significant cut-off 0.01% (LL threshold was adjusted to 37, minimum frequency set to 3). The interesting form *Zystchen*, a diminutive of *Zyste* ('cyst') was found on rank 78 of this list: 433 times in reports for female patients (fpmw of 110.07) while in reports for male patients this word form only occurred 31 times (7.82 fpmw): Hence, linguistic questions related to gender medicinal issues arise from this finding: Are such small cysts reported more often with female patients because women are found to have more cysts in general? Do women have more organs prone to cyst formation such as ovaries or glandular breast tissue? Are diseases with cystic changes to organs more common in women? Do male patients simply have fewer cysts or are they generally larger? Which words are used to describe small cysts in men? Is the word *Zystchen* really linked with gender or is there also another link to be found in the metadata, e.g. a certain type of examination such as mammography or is the word only used by a certain department etc.?

To further investigate the key item *Zystchen*, we looked at the context in which the basic form *Zyste* ('cyst') and the diminutive form *Zystchen* ('tiny cyst') are used. We specifically looked at adjectives, since they provide infor-

mation on how nouns they refer to are modified and characterized. Almost all of the manually filtered adjectival collocates for the diminutive *Zystchen* (‘tiny cyst’) could also be identified as collocates of *Zyste* (‘cyst’) as the following table shows (LogRatio, L3-R3, C5-NC5; LogLikelihood filter applied).

Top 25 adjectival collocates of the terms *Zyste* (‘cyst’) and *Zystchen* (‘tiny cyst’)

Zyste			*Zystchen*		
Frequency	LogRatio value	Collocate	Frequency	LogRatio value	Collocate
20	10.154	schmutzige	18	10.091	mikrozystisch
45	8.376	parapelvinen	33	8.984	unkomplizierte
33	7.947	unkomplizierte	88	7.937	disseminierten
51	7.569	komplizierte	221	7.214	retroareoläre
49	7.524	arachnoidale	324	7.105	einzelner
16	7.431	medulläre	1109	7.095	vereinzelte
16	7.431	proteinreiche	178	7.06	blander
23	7.376	eingedickten	3594	6.918	Einzelne
138	7.376	eingeblutete	144	6.672	disseminierte
402	6.444	soliden	487	6.494	solide
44	5.906	viele	274	6.414	winziges
141	5.498	septierte	1694	6.024	blande
88	5.435	disseminierten	226	6.003	mehrerer
99	5.415	randsklerosierte	1725	5.958	winzige
610	5.318	größte	377	5.942	parapelvine
271	5.138	kortikalen	806	5.844	einzelnen
121	5.096	blanden	514	5.814	zahlreiche
208	4.776	haltende	1767	5.798	kleines
1993	4.716	messende	300	5.587	winzigen
160	4.5	zahlreichen	9343	4.968	kleinere
1725	4.5	Winzige	9343	4.968	kleine
94343	4.226	kleine	678	4.662	vereinzelt
4259	4.218	entsprechend	2898	4.565	kleiner
169	4.036	kleineren	610	4.55	größte
146	4.024	größten	1911	4.163	großes
581	3.901	größere	1669	3.771	subchon
767	3.899	Kleinere	3234	3.724	mehrere
969	3.879	Multiple	5193	3.219	kleinen
248	3.662	kleinste	19204	2.944	unauffällige
300	3.554	winzigen	10081	2.342	geringe

Among the top 25 adjectival collocates, many word forms were found to be the same for *Zyste* ('cyst') as well as for the diminutive *Zystchen* ('tiny cyst'). These top 25 adjectival collocates may be categorized semantically as follows:

1) Location of cysts
2) Amount of cysts
3) Stage of disease
4) Size of cyst

Descriptions regarding the size of cysts are frequent: Adjectives such as *winzigste* ('tiniest'), *winzig* ('tiny') *klein* ('small'), *kleinere* ('smaller'), *groß* ('big', *größer* ('bigger') and *größte* ('biggest') occur frequently together with both the regular and the diminutive form of 'cyst'. This is interesting because it points to the lexical option to use the more regular form as well: radiologists use small cyst or tiny cyst when linguistically minimizing 'cyst'; and sometimes they minimize small cyst twice by adding further diminutive adjectives to the diminutive form. Furthermore, the top collocate *mikrozystisch* ('microcystic') exclusively appears together with the diminutive form of cyst. As stated above, such (double) diminutives almost exclusively occur with female patients, most of them in mammographic screenings.

Does this indicate an unconscious bias, for example that women are perceived as more emotionally fragile and thus clinicians feel they have to filter information or tone it down in some way? The pilot study showed that the use of these forms is almost entirely restricted to mammographic screenings. Are there any medical reasons that can account for this? For example, are cysts in general smaller in (female) breast tissue? These complex and multifactorial questions cannot be answered by looking at the corpus but rather require the eye and knowledge of a medical professional. They are directly connected to more general questions on biases in healthcare. The next section provides a general overview on biases in healthcare and discusses the recently growing body of literature from a gender medicine viewpoint.

5 Bias in healthcare – a review of the literature

Initially, the project germinated after observations of different waiting times for some examinations between women and men, showing disadvantages of women, and the causes remaining elusive. Unconscious biases leading to discrimination because of age, gender, nationality, class or status have been found to play a significant role in healthcare. Williams et al. state that a person is hindered from exploiting their "health potential" (2014: 32) by such biases. Biases leading to systemic discrimination[2] thus play a significant role in every process and every situation in all of healthcare and the following short review can only touch on very few selected aspects. Severe underrepresentation of women is a factor in medicine (Abdellatif et al. 2019), for example particularly in the surgical disciplines (Wu et al. 2019), in radiology (Qamar et al 2020), in cardiology (Shahid 2019), and in academic medicine (Rosso et al. 2019). As healthcare personnel, women are systematically underrepresented in clinical research (Gupta et al. 2019). Gender biases have been proven for different levels and occupational fields as well as in study results and academic and scientific work – this means that they are not isolated events but rather systemic and may affect individuals on many levels. There are certain questions which are only ever asked of women for example on childcare (Weber et al. 2019). Also authorships in medical publications (Bernardi et al. 2018) as well as in review activities (Steinberg et al. 2018) reflect such biases. Women are less frequently represented on editorial boards and are less likely to be invited to speak at conferences, to receive grants (Aldrich et al. 2019) or to be invited to comment in journals (Thomas et al. 2019). When grants are approved, start-up packages are lower for women than for men (Sege et al. 2015). Perhaps this is one reason why, while as many women as men start careers in medicine (Bates et al. 2016), far fewer women than men become full professor or department chair (Lautenberger et al. 2014). For ophthalmology, for example, it was shown that

1 The terms structural, systemic and institutional discrimination are often used synonymously (Feagin 2006) and broadly refer to historical, cultural and institutional policies, norms and practices that maintain social inequalities by privileging certain groups and disadvantaging others (Bonilla-Silva 1997; Jones 2000). All these terms focus on societal power structures rather than on individual, intentional actions of discrimination (Priest and Williams 2018).

once women have begun their medical training, they have less access to a wet lab for practice, they have fewer procedures per week, and are more likely than men not to have performed a certain number of training procedures (Gibson et al. 2005). Furthermore, financial and administrative skills are more likely to be lacking, and there is a delay in opening a practice (Lira et al. 2013). Female ophthalmologists perform fewer operations than male ophthalmologists (French et al. 2016), and receive less money overall, as well as less money for comparable single cases (Buys et al. 2019).

Gender inequality also extends into the private lives of medical workers: the relationships of female ophthalmologists are less stable than those of male ophthalmologists and they have fewer children (Deva and Danesh-Meyer 2008). The same pattern can also be found in other areas, for example for female plastic surgeons it was found that they are less likely to be married than male plastic surgeons (Furnas et al. 2018); plastic surgeons are also discouraged from having children during their residency (Eskenazi and Weston 1995) and pregnancy is a widespread reason for discrimination in this field (Bucknor et al. 2018).

Also in radiology, the field the corpus MedCorpInn is situated in, women face bias and discrimination. Studies found that they are underrepresented in research in general (Vernuccio et al. 2019) and are less often recipients of leadership awards (Martin et al. 2019). Fewer of them are in leadership positions (Qamar et al. 2020) and as well as in various radiology fellowships (West and Nguyen 2017). Women are, 10 years after joining a department, less advanced in their career than the men who had joined at the same time (Dial et al. 1989).

If the perspective is changed to patients the picture is not much different: also as patients it is more likely for women to encounter and be affected by biases.

One particularly interesting aspect in different studies are waiting times: Women have to wait longer than men, for example, for angiography after a heart attack (Meyer et al. 2019), for their treatment of complex rhythm disturbances by catheter ablation (Carnlöf 2017) or for their cataract operation (Smirthwaite et al. 2014), and therefore have a worse outcome in some cases (Meyer et al. 2019; Carnlöf 2017).

Not only do women have to wait longer for treatment, they may not get it at all (Sagy et al. 2018) or have worse outcomes. The social category gender impacts pain management, for example in chronic pain or neck pain as well (Samulowitz et al. 2018; Racine et al. 2012; Hamberg et al. 2002). Hoffman and Tarzian (2001) report that women are more likely to seek treatment for chronic pain than men, but that they are also more likely to be exposed to undertreatment.

The category gender also impacts the diagnosis and treatment of coronary heart disease (Regitz-Zagrosek et al. 2004; Daugherty et al. 2017). In cardiology, women have to expect worse outcomes as well as differences in access to top-level medicine and state-of-the-art medication (Hochleitner 2013). After a heart attack, women are less likely to receive adequate initial treatment such as thrombolysis (Trappolini et al. 2002), and have a higher one-year mortality rate than men, which cannot be explained by a greater extent of myocardial damage (Kosmidou et al. 2017).

Visible and directly researchable discrimination, e.g. in the form of fewer operations, less pay, or less advanced careers after the same period, is only one aspect of discrimination against women. Other forms of discrimination and biases are invisible and difficult to detect, such as the lack of invitations to female colleagues to give lectures or to take up positions in professional societies, exclusion during special operations.

Traces of both, visible and invisible discrimination, with regard to medical examinations or applications or measures could be reflected in the language used, for example in medical reports as are included in the MedCorpInn corpus. Furthermore, the large body of data allows for specific gender medicinal research questions, some of which will be more closely described in the following section.

5.1 The corpus data from a Gender Medicine viewpoint

The data of MedCorpInn is unique and important for the field of Gender Medicine for several reasons: It is collected at one of the largest radiology departments in Central Europe and includes neuroradiology. It represents a de-identified version of principally sensitive data, hence patient follow ups are

possible and can be applied for different research questions. Several issues relevant for Gender Medicine in different clinical areas members of the project team are specialized in can be addressed:

1) As the corpus can be divided according to male/female patients (2,540,022 female patients; 2,440,474 male patients), it is possible to look for different ratios, for example of specific types of examinations on any given day of the week, which would indicate whether one of the two groups is receiving fewer examinations than the other. A first look at the best practice data suggests there might be a difference in the number of examinations women/men receive. For example, if we look at examinations that do not include frequent sex-specific exams (like mammography, prostate exams) it appears that women are examined less frequently overall. The data can be traced from any given initial examination and investigate if, e.g., women are less likely than men to get a certain type of follow-up treatment/examination.

2) The well-tagged information on measurements and sizes may help answer the question if the sizes of tumors or structures in women are linguistically less precisely specified than in men. Can the texts reveal that such measurements are more likely to be estimated rather than exactly measured for female patients?

3) The extensive metadata will enable the team to investigate if during their training in radiology, women are enabled to do the same examinations as men, in the same quantity, after the same time?

Such questions can only be answered with the text corpus because it is impossible, of course, to find and compare several million measurement results or assigned follow-up examinations etc. manually. Furthermore, the large amount of compiled data also increases the statistical significance of the subsequent analyses. Because of the well documented and prepared metadata highly specific subcorpora can be queried, e.g. all computed tomographies of the liver in the data can be identified and both the number and size of the tumors contained in these findings can be extracted.

Furthermore, gender-specific differences concerning the probability for patients to gain access to intensive care units may be studied: Previous research found that most often, women are referred to such units only if they are more seriously ill than men. This may also depend on the gender of the referring doctor, since women refer female patients to intensive care later than they do male patients (Sagy et al. 2018).

6 Conclusion and outlook

In this paper, we have outlined some general insights into our approaches and reasons for building a large, linguistically tagged corpus of radiology reports. We wanted to present some general issues with corpus building and with implementing a text processing pipeline, e.g. data cleaning, anonymization, ethics aspects and tagging. Furthermore, we explored existing literature to mark out the fields of research for which this corpus can be useful. Definitely the interdisciplinary scientific collaboration and use of data within the fields of linguistics, (gender) medicine and informatics/Natural Language Processing will be a result by itself. By working together, we want to be able to describe linguistic peculiarities of the data (e.g. text structure, use of medical terminology, syntactic features etc.) and of salient patterns of language use. We also will be able to work with statistical outputs which provide information about salient linguistic patterns (if any) along the lines of social/economic categories in the data. In this way new methods of detecting structural/intersectional biases on the linguistic surface of large datasets shall be generated. We would also want the project to have a connection to medical practice and hope to arrive at ideas for guidelines to eliminate structural biases which are potentially harmful to patients. This ultimately contributes to a "recipient-tailored health communication" as suggested by Brown et al. (2006: 79).

Bibliography

Abdellatif, W./Ding, J./Jalal, S./Chopra, S./Butler, J./Ali, I. T./Shah, S./Khosa, F. (2019): Leadership Gender Disparity Within Research-Intensive Medical Schools: A Transcontinental Thematic Analysis. *Journal of Continuing Education in the Health Professions*, 243–50.

Addis, M. E./Mahalik, J. R. (2003): Men, masculinity, and the contexts of help seeking. *American Psychologist*, *58*(1), 5–14.

Adolphs, S./Brown, B./Carter, R./Crawford, P./Sahota, O. (2004): Applying corpus linguistics in a health care context. *Journal of Applied Linguistics*, 1(1), 9–28.

Aldrich, M. C./Cust, Anne E./Raynes-Greenow, C. (2019): Gender equity in epidemiology: a policy brief. *Annals of Epidemiology*, 35, 1–3.

Baker, S./Silins, I./Guo, Y./Ali, I./ Högberg, J./Stenius, U./Korhonen, A. (2016): Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3), 432–440.

Bates, C./Gordon, L./Travis, E./Chatterjee, A./Chaudron, L./Fivush, B./Gulati, M./Jagsi, R./Sharma, P./Gillis, M./Ganetzky, R./Grover, A./Lautenberger, D./Moses, A. (2016): Striving for Gender Equity in Academic Medicine Careers: A Call to Action. *Academic Medicine*, 91(8), 1050–1052.

Becker, M./Kasper, S./Böckmann, B./Jöckel, K. H./Virchow, I. (2019): Natural language processing of German clinical colorectal cancer notes for guideline-based treatment evaluation. *International Journal of Medical Informatics*, 127, 141–146.

Benford F. (1938): The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 78(4), 551–572.

Bernardi, K./Lyons, N. B. /Huang, L./Holihan, J. L/Olavarria, O./Martin, A/Milton, Alexis/Loor, Michele/Zhang, Feibi/Tyson, Jon/Ko, Tien/Liang, Mike (2018): Gender Disparity in Authorship of Peer Reviewed Medical Publications. *The American Journal of the Medical Sciences*, 360(5), 511–516.

Bonilla-Silva, E. (1997): Rethinking racism: Toward a structural interpretation. *American Sociological Review*, 62(3), 465–480.

Borges, S. (1986): A feminist critique of scientific ideology: an analysis of two doctor-patient encounters. In S. Fisher and A. Todd (Hg.), *Discourse and Institutional Authority: Medicine, Education and Law*, 26–48.

Bretschneider, C./Zillner, S./Hammon, M. (2013): Grammar-based lexicon extension for aligning German radiology text and images. In: INCOMA Ltd. Shoumen (Hg.) *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP2013*, 105–112. https://www.aclweb.org/anthology/R13-1014/

Brown, B./Crawford, P./Carter, R. (2006): *Evidence-based health communication*. Open University Press.

Bucknor, A./Kamali, P./Phillips, N./Mathijssen, I./Rakhorst, H./Lin, S. J./Furnas, H. (2018): Gender Inequality for Women in Plastic Surgery: A Systematic Scoping Review. *Plast Reconstr Surg*, 141(6), 1561–1577.

Buchholz, M. B. (1998): Die Metapher im psychoanalytischen Dialog, *Psyche*, 52(6), 545–71.

Buys, Y. M./Canizares, M./Felfeli, T./ Jin, Y. (2019): Influence of Age, Sex, and Generation on Physician Payments and Clinical Activity in Ontario, Canada: An Age-Period-Cohort Analysis. *Am J Ophthalmol*, 197, 23–35.

Bührig, K./Meyer, B. (2015): 16. Ärztliche Gespräche mit MigrantInnen. In: A. Busch und T. Spranz-Fogasy (Hg.): *Handbuch Sprache in der Medizin*. Berlin, München, Boston: De Gruyter, 300–316.

Carnlöf, C./Iwarzon, M./Jensen-Urstad, M./Gadler, F./Insulander, P. (2017): Women with PSVT are often misdiagnosed, referred later than men, and have more symptoms after ablation. *Scandinavian Cardiovascular Journal*, 51(6) 299–307.

Charteris-Black, J./Seale, C. (2010): *Gender and the language of illness.* London: Palgrave Macmillan.

Collins, S./Peters, S./Watt, I. (2010): Medical communication. In: H. Hamilton/W. S. Chou (Hg.). *The Routledge Handbook of Applied Linguistics*: Routledge.

Cooper, L. A./Roter, D. L./Carson, K. A./Beach, M. C./Sabin, J. A./Greenwald, A. G./Inui, T. S. (2012): The Associations of Clinicians' Implicit Attitudes About Race With Medical Visit Communication and Patient Ratings of Interpersonal Care. *American Journal of Public Health*, 102(5), 979–987.

Crawford, P./Brown, B./Nolan, P. (1998): Communicating care: *The language of nursing. Cheltenham*. Gloucester: Stanley Thornes.

Crawford, P./Brown, B./Harvey, K. (2014): Corpus linguistics and evidence-based health communication. In: Wen-ying S. J. Chou/H. Hamilton-Ehernberger (Hg.).

The Routledge Handbook of Language and Health Communication. London, New York: Routledge, 75–90.

Dalianis, H./Hassel, M./Velupillai, S. (2006): The Stockholm EPR Corpus – Characteristics and Some Initial Findings. In: *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research*, Kalmar, Sweden, October 14–16, 2009.

Daugherty, S. L./Blair, I. V/Havranek, E. P./Furniss, A./Dickinson, L. M./Karimkhani, E./Main, D. S./Masoudi, F. A. (2017): Implicit Gender Bias and the Use of Cardiovascular Tests Among Cardiologists. *Journal of the American Heart Association*, 6(12).

Demjén, Z. (2020): *Applying linguistics in illness and healthcare contexts*. London, New York: Bloomsbury Academic.

Deva, N. C./Danesh-Meyer, H. V. (2008): Gender differences in income. *Ophthalmology*, 115(2), 411.

Dial, T. H./Bickel, J./Lewicki, A. M. (1989): Sex differences in rank attainment among radiology and internal medicine faculty. *Acad Med*, 64(4), 198–202.

Eskenazi, L./Weston, J. (1995): The pregnant plastic surgical resident: Results of a survey of women plastic surgeons and plastic surgery residency directors. *Plastic and Reconstructive Surgery*, 95(2), 330–335.

Feagin, J. R. (2006): *Systemic racism: A theory of oppression*. New York, London: Routledge/Taylor & Francis Group.

Fette, G./Ertl, M./Wörner, A./Kluegl, P./Störk, S./Puppe, F. (2012): Information Extraction from Unstructured Electronic Health Records and Integration into a Data Warehouse. In: Goltz, U. et al. (Hg.), *INFORMATIK 2012*. Bonn: Gesellschaft für Informatik e.V., 1237–1251.

French, D. D./Margo, C. E./Campbell, R. R./Greenberg, P. B. (2016): Volume of Cataract Surgery and Surgeon Gender: The Florida Ambulatory Surgery Center Experience 2005 Through 2012. *Journal of Medical Practice Management*, 31(5), 297–302.

Furnas, H. J./Garza, Rebecca M./Li, Alexander Y./Johnson, D. J./Bajaj, A. K./Kalliainen, L. K./Weston, J. S./Song, D. H./Chung, K. C./Rohrich, R. J. (2018): Gender Differences in the Professional and Personal Lives of Plastic Surgeons. *Plastic Reconstructive Surgery*, 142(1), 252–264.

Gibson, A./Boulton, M. G./Watson, M. P./Moseley, M. J./Murray, P. I./Fielder, A. R. (2005): The first cut is the deepest: basic surgical training in ophthalmology. *Eye*, 19(12), 1264–1270.

Govender, V./Penn-Kekana, L. (2008): Gender biases and discrimination: A review of health care interpersonal interactions. *Global Public Health*, *3* (SUPPL. 1), 90–103.

Goddu, A. P./O'Conor, K. J./Lanzkron, S./Saheed, M. O./Saha, S./Peek, M. E./Haywood, C./Beach, M. C. (2018): Do Words Matter? Stigmatizing Language and the Transmission of Bias in the Medical Record. *Journal of General Internal Medicine*, 33(5), 685–691.

Gupta, G. R./Oomman, N./Grown, C./Conn, K./Hawkes, S./Shawar, Y. R./Shiffman, J./Buse, K./Mehra, R./Bah, C. A./Heise, L./Greene, M. E./Weber, A. M./Heymann, J./Hay, K./Raj, A./Henry, S./Klugman, J./Darmstadt, G. L. (2019): Gender equality and gender norms: framing the opportunities for health. *Lancet*, 393(10190), 2550–2562.

Hahn, U./Wermter, J. (2004): An Annotated German-Language Medical TextCorpus as Language Resource. In: *ELRA (Hg.), Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 473–476.

Hahn, U./Matthies, F./Lohr,C./Löffler, M. (2018): 3000PA-Towards a National Reference Corpus of German Clinical Language. *Studies in Health Technology and Informatics* 247, 26–30.

Hamberg, K./Risberg, G./Johansson, E. E./Westman, G. (2002): Gender Bias in Physicians' Management of Neck Pain: A Study of the Answers in a Swedish National Examination. *Journal of Women's Health & Gender-Based Medicine*, 11(7), 653–666.

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

Hedegaard, J./Ahl, H./Rovio-Johansson, A./Siouta, E. (2014): Gendered Communicative Construction of Patients in Consultation Settings. *Women & Health*, 54(6), 513–529.

Hellrich, J./Matthies, F./Faessler, E./Hahn, U. (2015): Sharing models and tools for processing German clinical texts. *Studies in Health Technology and Informatics* 210, 734–738.

Hochleitner, M. (2013): Genderaspekte bei kardiovaskulären Krankheiten. *Zeitschrift Für Gerontologie Und Geriatrie*, 46(6), 517–519.

Hoffman, D./Tarzian, A. (2001): The girl who cried pain: a bias against women in the treatment of pain. *Journal of Law, Medicine & Ethics* 29, 13–27.

Holub, P./Kohlmayer, F./Prasser, F./Mayrhofer, M. T./Schlünder, I./Martin, G. M./ Casati, S./Koumakis, L./Wutte, A./Kozera, Ł./Strapagiel, D./Anton, G./Zanetti, G./Sezerman, O. U./Mendy, M./Valík, D./Lavitrano, M./Dagher, G./Zatloukal, K./ van Ommen, G./Litton, J.-E. (2018): Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. *Biopreservation and Biobanking,* 16(2), 97–105.

Hunt, D./Carter, R. (2012): Seeing through The Bell Jar: Investigating Linguistic Patterns of Psychological Disorder. *Journal of Medical Humanities*, 33(1), 27–39.

Ibrahim, S. A./Whittle, J./Bean-Mayberry, B./Kelley, M. E./Good, C./Conigliaro, J. (2003): Racial/ethnic variations in physician recommendations for cardiac revascularization. *American Journal of Public Health*, 93(10), 1689–1693.

Irschara, K. (2018): *Von Zystchen und gut 3 cm. Eine korpus- und genderlinguistische Analyse radiologischer Befunde*. Institut für Sprachen und Literaturen, Sprachwissenschaft. [unpublished MA Thesis].

Isaac, C./Chertoff, J./ Lee, B./Carnes, M. (2011): Do students' and authors' genders affect evaluations? A linguistic analysis of Medical Student Performance Evaluations. *Academic Medicine : Journal of the Association of American Medical Colleges*, 86(1), 59–66.

Jiang, M./Chen, Y./Liu, M./Rosenbloom, S. T./Mani, S./Denny, J. C./Xu, H. (2011): A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5), 601–606.

Johnson, S. B. (1999): A Semantic Lexicon for Medical Language Processing. *Journal of the American Medical Informatics Association*, 6(3), 205–218.

Johnson, R. L./Roter, D./Powe, N. R./Cooper, L. A. (2004): Patient race/ethnicity and quality of patient-physician communication during medical visits. *American Journal of Public Health*, *94*(12), 2084–2090.

Jones, C. (2000): Levels of Racism: A Theoretic Framework and a Gardener's Tale. *American Journal of Public Health*. 90, 1212–1215.

Jungmann, F./Kuhn, S./Tsaur, I./Kämpgen, B. (2019): Natural language processing in radiology: Neither trivial nor impossible. *Radiologe*, 59(9), 828–832.

Kara, E./Zeen, T./Gabryszak, A./Budde, K./Schmidto, D./ Roller, R. (2018): A domain-adapted dependency parser for German clinical text. In: A. Barbaresi/H. Biber /F. Neubarth/R. Osswald (Hg.), *Proceedings of KONVENS 2018*, Wien: ÖAW, 12–17.

Kelly, R. (1998): Nurses Talking: a radical policy, ethnomethodology, for researching critical care nursing. *Nursing in Critical Care*, 3, 41–46.

Kelly, J. F./Westerhoff, C. M. (2010): Does it matter how we refer to individuals with substance-related conditions? A randomized study of two commonly used terms. *International Journal of Drug Policy*, 21(3), 202–207.

Koerfer, A./Martens-Schmid, K. (2000): Erzählen in der Psychotherapie. *Psychotherapie und Sozialwissenschaft*, 2(2), 83–86.

Kosmidou, I./Redfors, B./Selker, H. P./Thiele, H./Patel, M. R./Udelson, J. E./Ohman, E. M./Eitel, I./Granger, C. B./Maehara, A./Kirtane, A./Généreux, P./Jenkins, P. L./Ben-Yehuda, O./Mintz, G. S./Stone, G. W. (2017): Infarct size, left ventricular function, and prognosis in women compared to men after primary percutaneous coronary intervention in ST-segment elevation myocardial infarction: results from an individual patient-level pooled analysis of 10 randomized trials. *European Heart Journal*, 38(21), 1656–1663.

Krebs, J./Corovic, H./Dietrich, G./Ertl, M./Fette, G./Kaspar, M./Krug, M./Störk, S./ Puppe, F. (2017): Semi-automatic terminology generation for information extraction from German chest X-ray reports. *Studies in Health Technology and Informatics*, 243, 80–84

Krieger, H.-U./Spurk, C./Uszkoreit, H./Xu, F./Zhang, Yi/Müller, F/Tolxdorff, T. (2014): Information Extraction from German Patient Records via Hybrid Parsing and Relation Extraction Strategies. In: N. Calzolari (Hg.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik: ELRA, 2043–2048.

Kreuzthaler, M./Schulz, S. (2015): Detection of sentence boundaries and abbreviations in clinical narratives. *BMC Medical Informatics and Decision Making* 15(2):S4.

Kreuzthaler, M./Oleynik, M./Avian, A./Schulz, S. (2016): Unsupervised Abbreviation Detection in Clinical Narratives. In: The COLING 2016 Organizing Committee

(Hg.) *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, 91–98.

Lautenberger, D. M./Dandar, V. M./Raezer, C. L./Sloane, R. A. (2014): *The State of Women in Academic Medicine: The Pipeline and Pathways to Leadership 2013–2014*. Washington, DC: Association of American Medical Colleges.

Lewis, B. (1995): Psychotherapeutic discourse analysis. *American Journal of Psychotherapy*, 49(3), 371–384.

Lira, R. P./Chaves, F. R./Arieta, C. E. (2013): Initial challenges in the career of ophthalmologists. *Arquivos Brasileiros de Oftalmologia*, 76(2), 134–135.

Lohr, C./Buechel, S./Hahn, U. (2019): Sharing copies of synthetic clinical corpora without physical distribution – A case study to get around iPRS and privacy constraints featuring the German JSYNCC corpus. In: N. Calzolari et al. (Hg.). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki: ELRA.

Lohr, C./Herms, R. (2016): A corpus of German clinical reports for ICD and OPS-based language modeling. In: N. Calzolari et al. (Hg.). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki: ELRA, 20–23.

Martin, J. F./Hewett, L./Gordon, L. L./Lewis, M. C./Cluver, A./Collins, H. (2019): Do Gender Disparities Among Major Radiological Society Award Recipients Exist? *Academic Radiology*, 27(7), Epub ahead of print.

Maynard, D. W. (1991): Interaction and Asymmetry in Clinical Discourse. *American Journal of Sociology*, 97(2), 448–495.

McCarthy, M./Handford, M. (2004): "Invisible to us": a preliminary corpus-based study of spoken business English. In U. Connor/T. Upton (Hg.), *Discourse in the Professions. Perspectives from Corpus Linguistics.* Amsterdam: Benjamins, 167–201

Menz, F./Lalouschek, J. (2006): "I just can't tell you how much it hurts." Gender-relevant Differences in the Description of Chest Pain. In: F. Salager-Meyer/M. Gotti (Hg.), *Advances in medical discourse analysis. Oral and written contexts*. Bern: Peter Lang Verlag, 133–154.

Menz, F. (2010): Sprechen über Schmerzen. Univ.-Verl. Rhein-Ruhr.

Meyer, M. R./Bernheim, A. M./ Kurz, D. J./O'Sullivan, C. J./Tüller, D./Zbinden, R./ Rosemann, T./Eberli, F. R. (2019): Gender differences in patient and system delay

for primary percutaneous coronary intervention: current trends in a Swiss ST-segment elevation myocardial infarction population. European Heart Journal, 8(3), 283–290.

Morris, G. H. /Chenail, R. J. (Hg.) (1995): The talk of the clinic. Explorations in the analysis of medical and therapeutic discourse. Hillsdale, NJ: Erlbaum.

Newcomb S. (1881): Note on the Frequency of Use of the Different Digits in Natural Numbers. American Journal of Mathematics, 4(1), 39–40.

Oberkampf, H./Bretschneider, C./Zillner, S./Bauer, B./Hammon, M. (2014): Knowledge-based extraction of measurement-entity relations from german radiology reports. In: 2014 IEEE International Conference on Healthcare Informatics, 149–154.

Paternotte, E./Scheele, F./Seeleman, C. M./Bank, L./Scherpbier, A./van Dulmen, S. (2016): Intercultural doctor-patient communication in daily outpatient care: relevant communication skills. *Perspectives of Medical Education* 5(5), 268–275.

Parry, R. (2004): The interactional management of patients' physical incompetence: a conversation analytic study of physiotherapy interactions. *Sociology of Health and Illness*, 26(7), 96–1007.

Pauwels, A. (1995): *Cross-cultural communication in the health sciences: communicating with migrant patients*. Macmillan Education Australia.

Peck, B. M./Denney, M. (2012): Disparities in the conduct of the medical encounter: The effects of physician and patient race and gender. *SAGE Open*, 2(3), 1–14.

Pilnick, A: (1998): Why didn't you just say that?: Dealing with issues of asymmetry, knowledge and competence in the pharmacist/client encounter. *Sociology of Health and Illness*, 20(1), 29–51.

Priest, N./Williams, D. R. (2018): Racial Discrimination and Racial Disparities in Health. In: B. Major/J. F. Dovidio/B. G. Link (Hg.): *The Oxford handbook of stigma, discrimination, and health*. New York: Oxford University Press.

Qamar, S. R./Khurshid, K./Jalal, S./McInnes, M. D. F./Probyn, L./Finlay, K./Hague, C. J./Hibbert, R. M./Joshi, M./Rybicki, F. J./Harris, A./Nicolaou, S./Khosa, F. (2020): Gender Disparity Among Leaders of Canadian Academic Radiology Departments. *American Journal of Roentgenology*, 214(1), 3–9.

Racine, M./Tousignant-Laflamme, Y./Kloda, L. A./Dion, D./Dupuis, G./Choinière, M. (2012): A systematic literature review of 10 years of research on sex/gender and

experimental pain perception – Part 1: Are there really differences between women and men? *Pain*, 153(3), 602–618.

Regitz-Zagrosek, V./Lehmkuhl, E./Hocher, B./Goesmann, D./Lehmkuhl, H. B./Hausmann, H./Hetzer, R. (2004): Gender as a risk factor in young, not in old, women undergoing coronary artery bypass grafting. *Journal of the American College of Cardiology*, 44(12), 2413–2414.

Richter-Pechanski, P./Riezler, S./Dieterich, C. (2018): De-Identification of German Medical Admission Notes. *Studies in Health Technology and Informatics*, 253, 165–169.

Roberts, C. (2006): Continuities and discontinuities in doctor – patient consultations in a multilingual society. In: F. Salager-Meyer/M. Gotti (Hg.), *Advances in medical discourse analysis. Oral and written contexts*. Bern: Peter Lang, 177–196.

Roller, R./Uszkoreit, H./Xu, F./Seiffe, L./Mikhailov, M.7Staeck, O./Schmidt, D. (2016): A fine-grained corpus annotation schema of German nephrology records. In: A. Rumshisky/K. Roberts/S. Bethard/T. Naumann (Hg.) *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, Osaka: The COLING 2016 Organizing Committee, 69–77. Roller, R./Rethmeier, N./Thomas, P./Hübner, M./Uszkoreit, H./Staeck, O./Budde, K./Halleck, F./Schmidt, D. (2018): Detecting Named Entities and Relations in German Clinical Reports. In: G. Rehm/T. Declerck (Hg.), *Language Technologies for the Challenges of the Digital Age*. Cham: Springer International Publishing, 146–154.

Rosso, C./Leger, A./Steichen, O. (2019): Glass ceiling for women in academic medicine in France. *La Revue de Medecine Interne*, 40(2), 82–87.

Rucker-Whitaker, C. (2003): Explaining Racial Variation in Lower Extremity Amputation. *Archives of Surgery*, 138(12), 1347.

Sagy, I./Fuchs, L./Mizrakli, Y./Codish, S./Politi, L./Fink, L./Novack, V. (2018): The association between the patient and the physician genders and the likelihood of intensive care unit admission in hospital with restricted ICU bed capacity. *QJM*, 111(5), 287–294.

Samulowitz, A./Gremyr, I./Eriksson, E./Hensing, G. (2018): "Brave Men" and "Emotional Women": A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain. *Pain Research and Management*, 2018, 1–14.

Sege, R./Nykiel-Bub, L./Selk, S. (2015): Sex Differences in Institutional Support for Junior Biomedical Researchers. *JAMA*, 314(11), 1175–1177.

Shahid, M. (2019): Addressing the Underrepresentation of Women in Cardiology through Tangible Opportunities for Mentorship and Leadership. *Methodist Debakey Cardiovasc Journal*, 15(1), e1–e2.

Smirthwaite, G./Lundström, M./Albrecht, S./Swahnberg, K. (2014): Indication criteria for cataract extraction and gender differences in waiting time. *Acta Ophthalmologica*, 92(5), 432–438.

Steinberg, J. J./Skae, C:/Sampson, B. (2018): Gender gap, disparity, and inequality in peer review. *Lancet*, 93(10140), 2602–2603.

Street, R. L./Gordon, H./Haidet, P. (2010): "Physicians" communication and perceptions of patients: Is it how they look, how they talk, or is it just the doctor? *Social Science*, 65(3), 586–598.

Šuster, S./Tulkens, S./Daelemans, W. (2017): A Short Review of Ethical Challenges in Clinical Natural Language Processing. In: *First Workshop on Ethics in Natural Language Processing (EACL'17)*, preprint.

Thomas, E. G./Jayabalasingham, B./Collins, T./Geertzen, J./Bui, C./Dominici, F. (2019): Gender Disparities in Invited Commentary Authorship in 2459 Medical Journals. *JAMA Netw Open*, 2(10), e1913682.

Toepfer, M./Corovic, H./Fette, G./Klügl, P./Störk, S./Puppe, F. (2015): Fine-grained information extraction from German transthoracic echocardiography reports, *BMC Medical Informatics and Decision Making*, 15(1).

Trappolini, M. /Chillotti, F. M./Rinaldi, R./Trappolini, F./Coclite, D./Napoletano, A. M. /Matteoli, S. (2002): Sex Differences in Incidence of Mortality After Acute Myocardial Infarction. *Italian Heart Journal*, 3(7), 759–766.

Trix, F./Psenka, C. (2016): Exploring the Color of Glass. *Discourse & Society*, 14(2), 191–220.

Valero-Garcés, C. (2002): Interaction and conversational constrictions in the relationships between suppliers of services and immigrant users. *Pragmatics*, 12(4), 469–495.

Waitzkin, H. (1984): Doctor-patient communication. Clinical implications of social scientific research. *JAMA*, 252(17), 2441–2446.

Weber, A. M./Cislaghi, B./Meausoone, V./Abdalla, S./Mejía-Guevara, I./Loftus, P./Hallgren, E./Seff, I./Stark, L./Victora, C. G./Buffarini, R./Barros, A. J.D./

Domingue, B. W./Bhushan, D./Gupta, R./Nagata, J. M./Shakya, H. B./Richter, L. M./Norris, S. A./Ngo, T. D./Chae, S./Haberland, N./McCarthy, K./Cullen, M. R./ Darmstadt, G. L. (2019): Gender norms and health: insights from global survey data. *Lancet*, 393(10189), 2455–2468.

West, C. (1984): When the Doctor is a "Lady": Power, Status and Gender in Physician-Patient Encounters. *Symbolic Interaction*, 7(1), 87–106.

West, D. L./Nguyen, H. T. (2017): Ethnic and Gender Diversity in Radiology Fellowships. *Journal of Racial and Ethnic Health Disparities*, 4(3), 432–445.

Willems, S./De Maesschalck, S./Deveugele, M./Derese, A./De Maeseneer, J. (2005): Socio-economic status of the patient and doctor-patient communication: Does it make a difference? *Patient Education and Counseling*, 56(2), 139–146.

Williams, S. D./Hansen, K./Smithey, M./Burnley, J./Koplitz, M./Koyama, K./Young, J./Bakos, A. (2014): Using Social Determinants of Health to Link Health Workforce Diversity, Care Quality and Access, and Health Disparities to Achieve Health Equity in Nursing. *Public Health Reports*, 129(Suppl 2), 32–36.

Wodak, R./Menz, F./Lalouschek, J. (1990): *Alltag in der Ambulanz: Gespräche zwischen Ärzten, Schwestern und Patienten*. Tübingen: Narr Verlag.

Wu, B./Bhulani, N./Jalal, S./Ding, J./Khosa, F. (2019): Gender Disparity in Leadership Positions of General Surgical Societies in North America, Europe, and Oceania. *Cureus* 11(12), e6285.

Vernuccio, F./Arzanauskaite, M./Turk, S./Torres, E. T./Choa, J. M. D./Udare, A. S./ Haroun, D./Serra, M. M./Shelmerdine, S./Bold, B./Bae, J. S./Romero, E. E./ Vilgrain, V. (2019): Gender discrepancy in research activities during radiology residency. *Insights into Imaging*, 10, 125.

Vila Rigat, M./González Fuente, S./Martí Antonín, M. A./Llisterri Boix, J./Machuca Ayuso, M. J. (2010): ClInt: a Bilingual Spanish-Catalan Spoken Corpus of Clinical Interviews. *Procesamiento Del Lenguaje Natural*, 45, 105–111.

Vila, T. V./Trigo, E. S. (2012): EMCOR: a medical corpus for terminological purposes. *The Journal of Specialised Translation*, 18, 139–159.

Xu, H./Stetson, P. D./Friedman, C. (2009): Methods for Building Sense Inventories of Abbreviations in Clinical Notes. *Journal of the American Medical Informatics Association*, 16(1), 103–108.

Andrea Lackner

Das Österreichische Gebärdensprachkorpus im Entstehen

Einleitende Gedanken[1]

Der leitende Gedanke der Sprachwissenschaft und damit in Beziehung stehender Disziplinen ist, dass Sprachbeschreibung und Sprachtheorien der Einbindung und Berücksichtigung der gesprochenen, geschriebenen UND der gebärdensprachlichen Praxis bedürfen. Um Ausdrucksformen natürlicher Sprachen zu beschreiben und daraus folgende Sprach- und Grammatiktheorien anzuwenden oder gar erst zu entwickeln, bedarf es daher der Einbeziehung unterschiedlicher Sprachmodalitäten. Es ist in diesem Sinne wichtig, nicht nur auf die Vielfalt von vokal-auditiven Modalitäten zu achten, sondern auch visuell-gestische zu berücksichtigen.

Die Gebärdensprachforschung – insbesondere die Gebärdensprachlinguistik – erfährt in vielen Ländern einen Aufschwung. Mit zunehmender Anzahl gesetzlicher Anerkennungen nationaler Gebärdensprachen (in Österreich 2005) sind diese vermehrt Gegenstand wissenschaftlicher Forschung geworden und tragen insbesondere im internationalen Austausch dazu bei, das Verständnis von Sprache und Kommunikation voranzubringen. Als Grundlage für die Forschungsarbeit nimmt das Sammeln und Archivieren von gebärdensprachlichen Texten (Korpusaufbau) wie auch das Beschreiben und Analysieren der darin vorkommenden Gebärden und ihrer begleitenden nicht-manuel-

1 Die aktuelle Erstellung eines ÖGS-Korpus findet im Rahmen des Projekts „Der Beitrag nichtmanueller Elemente zur Syntax der ÖGS“ statt, das vom Fonds zur Förderung der wissenschaftlichen Forschung (FWF, Projektnr. P29847, signnonmanuals.aau.at) finanziert und seitens der Autorin geleitet wird. Ein herzlicher Dank gilt dem gesamten Projektteam für die tolle Zusammenarbeit in der Erstellung unseres Korpus. Für diese Veröffentlichung möchte ich im Besonderen Laura Raffer, Nikolaus Riemer-Kankkonen, Isabel Graf, Elisabeth Scharfetter und Hannelore Lackner für Feedback und Korrektur danken.

len Elemente (Kopf-, Körper- und Gesichtsbewegungen, welche die Gebärden – die manuellen Elemente – begleiten) einen zentralen Stellenwert ein.

In diesem Sinne stellt die Erstellung, Annotation und Analyse von Gebärdensprachkorpora ein Muss dar, um (1) das mögliche Repertoire an sprachlichen Mitteln – in Form von manuellen und nicht-manuellen Komponenten – aufzuzeigen, (2) das Ausdrücken von sprachlichen Konzepten in Form von sequenziellen, aber auch simultan und dreidimensionalen Sprachphänomenen zu beleuchten und (3) solche Erkenntnisse für einen Sprachvergleich mit „Lautsprachen“[2] (auch unter Einbeziehung mit verschriftlichten Formen) nutzen zu können. Darüber hinaus kann so auch das Zusammenspiel mehrerer Modalitäten innerhalb eines Sprachsystems untersucht werden (siehe beispielsweise die Forschung zu der lautsprachbegleitenden Gestik und damit einhergehender Theorienbildungen).

Dieser Artikel soll die vorgestellten Gedankengänge anhand der praktischen Arbeit am Korpus zur Österreichischen Gebärdensprache (ÖGS) weiter ausführen und näher beschreiben. Nachdem die Wichtigkeit und Relevanz der Erstellung eines ÖGS Korpus, im Sinne der einleitenden Gedanken vorgestellt wurde, wird im Kapitel 1 auf die internationale Entwicklung und die sich damit etablierenden Kriterien für Gebärdensprachkorpora eingegangen. Im Kapitel 2 werden Prämissen und Grundlagen für die Erstellung von Gebärdensprachkorpora erläutert. Hierzu wird die sozio-kulturelle wie auch die sprachstrukturelle Dimension von Gebärdensprachen ausführlich beschrieben. Beide müssen stets während des Aufbaus eines Gebärdensprachkorpus mitbedacht werden. Wie das gelingen kann wird anhand der ÖGS veranschaulicht. Kapitel 3 widmet sich der Erstellung und korpusbasierten Annotation des ÖGS-Korpus. In Kapitel 4 werden erste Ergebnisse aufgezeigt.

2 In der Gebärdensprachforschung hat sich der Begriff „Lautsprachen“ eingebürgert, um bei Bedarf ein Pendant zu „Gebärdensprachen“ zu haben. An dieser Stelle sei ergänzt, dass zum einen die Verwendung des Terminus Lautsprache für gesproche und geschriebene Sprachpraxis irreführend ist und zum anderen Laute nicht Gebärden entsprechen.

1 Erstellung von Gebärdensprachkorpora

Seit den letzten fünfzehn Jahren gibt es zahlreiche Gebärdensprachforschungsgruppen, die begonnen haben, zu ihrer untersuchten Gebärdensprache einen Korpus aufzubauen.[3]

Ziel solcher Korpuserstellungen ist es, repräsentative Aufnahmen einer Gebärdensprache einer Gehörlosengemeinschaft (meist sind dies Aufzeichnungen verschiedener Varietäten einer Gebärdensprache eines Staates) zu sammeln und diese mit Metadaten und Annotationen zu versehen. Insbesondere das Annotieren der gebärdeten Einheiten in Form von ID-Glossen und (ausgewählter Aspekte) nicht-manuellen Komponenten wie auch das Annotieren linguistischer und/oder interaktiver Informationen ermöglicht Gebärdensprachkorpora „lesbar" und „durchsuchbar" zu machen (vgl. Johnston 2010, 2014; Fenlon et al. 2015; Johnston & Schembri 2006b).[4]

Exemplarisch für das Erstellen und Beschreiben von Gebärdensprachkorpora sei das Korpus der Australischen Gebärdensprache (AUSLAN-corpus) und das Korpus der Deutschen Gebärdensprache (DGS Korpus) genannt. Johnston archivierte seit 2008 rund 100 gehörlose Natives[5] bzw. beinahe Natives von ganz Australien im Endangered Language Archive (ELAR). Als Erster unter den Gebärdensprachforschern formuliert er Annotationsrichtlinien, die die Sprachbeschreibung systematisch und mit anderen Gebärdensprachen

3 Ein Überblick zu Sprachdokumentation und korpuslinguistischen Ansätzen in der Gebärdensprachforschung mit Auflistung der ersten Gebärdensprachkorpusprojekte ist in Fenlon et al. (2015) zu finden.

4 Hier sei angemerkt, dass es zu verschiedenen Gebärdensprachen – auch zur ÖGS – unterschiedliche Datensammlungen gibt, die von Sammlungen einzelner Fachgebärden bis zur Sammlung von Nachrichtenvideos reichen. Da diese Datensammlungen keinen sprachtheoretischen und korpuslinguistischen Ansätzen wie dem hier beschriebenen Gebärdensprachkorpus entsprechen, wird im Rahmen dieses Artikels nicht näher auf sie eingegangen.

5 In diesem Artikel werden die Begriffe „Native Signer" und „gehörlose MuttersprachlerInnen" gleichbedeutend verwendet. Wenn eine gehörlose Person sich für Korpusaufnahmen zur Verfügung stellt, wird sie als „InformantIn" bezeichnet, wenn sie im Prozess der Korpusannotation mitwirkt, als „AnnotatorIn". Grundsätzlich sind „GebärdensprachbenutzerInnen" all jene hörenden und gehörlosen Personen, die eine bestimmte Gebärdensprache verwenden. In diesem Artikel wird der Ausdruck aber stets in Bezug auf „gehörlose MuttersprachlerInnen", die im Korpusaufbau als InformantIn oder AnnotatorIn fungieren, verwendet.

vergleichbar machen sollen. Überdies erstellt er eine korpusbasierte Lexikondatenbank (*Auslan Signbank Dictionary*). Er ist damit als Pionier auf diesem Gebiet zu nennen.

Kommt es zur Masse an gesammelten Daten, ist das DGS Korpus ein relevantes Beispiel, an dem seit 2009 gearbeitet wird. Mit Stand 2018 enthält der Korpus 330 InformantInnen (rund 560 Aufnahmestunden) und 480.000 Gebärdeneinträge. Dieses Korpus zeichnet sich dahingehend aus, dass eine phonetische Gebärdenbeschreibung, das Hamburger Notationssystem für Gebärdensprachen (HaNoSys), zur Korpusbeschreibung eingesetzt wird.[6] Das DGS Korpus wird mit Hilfe des Annotationsprogrammes iLex, der Auslan-Korpus mit ELAN annotiert.[7]

Hinsichtlich des sprachtheoretischen und methodischen Zugangs der Annotation nicht-manueller Elemente in Gebärdensprachkorpora wurden innovative Wege im Zuge der ÖGS-Korpusforschung in Angriff genommen (siehe Lackner 2019a und 2017 sowie die beiden Workshops *SignNonmanuals* im Jahre 2014 und 2019[8]).

2 Prämissen und Grundlagen für die Erstellung von Gebärdensprachkorpora

Grundlegend für die Erstellung eines Gebärdensprachkorpus ist die Berücksichtigung gebärdensprachlicher Spezifika, auf die nun eingegangen werden soll. Unter solchen Spezifika versteht man die sequenzielle, simultane und dreidimensionale Sprachnatur von Gebärdensprachen. Sie soll nun genauer vorgestellt werden, um alle Aspekte, die beim Aufbau und der Annotation ei-

6 HamNoSys ist öffentlich zugänglich und kann auch in ELAN verwendet werden. Die erste Version vor über 20 Jahren wurde in enger Tradition zu Stokoe-basierten Systemen erstellt. NamNoSys ist Grundlage für eine Reihe von Avatarsteuerungen.

7 ilex (abgekürzt für „integriertes Lexikon“) wurde dahingehend entwickelt, dass parallel zum Korpusaufbau auch eine Lexikondatenbank erstellt wird (vgl. Hanke & Storz 2008). ELAN (*EUDICO Linguistic Annotator*) wurde am Max-Planck-Institut in Nijmegen zur Annotation von Video- und Audiodaten entwickelt. Es wird in zahlreichen Gebärdensprachen zur Korpus-Annotation verwendet (vgl. Crasborn & Sloetjes 2008).

8 Vgl. http://signnonmanuals.aau.at/node/589 und http://signnonmanuals.aau.at/node/678

nes Gebärdensprachkorpus mitbedacht werden müssen, zu veranschaulichen. Das Kapitel wird dazu in zwei größere Themengebiete unterteilt. Zuerst soll auf die sozio-kulturelle Dimension von Gebärdensprachen und die damit notwendige Innensicht der Gebärdensprachbeschreibung eingegangen werden. Hierbei soll am Beispiel der ÖGS näher beschrieben werden, wie sich das auf einzelne Teilaspekte dieser sozio-kulturellen Dimension auswirkt. Danach soll auf die Architektur der ÖGS und deren Auswirkungen auf den Aufbau eines Korpus näher eingegangen werden.

2.1 Soziokulturelle Einbettung von Gebärdensprachen am Beispiel der ÖGS

Gebärdensprachen sind eng mit gehörlosen Menschen und deren Gehörlosengemeinschaft verwoben. Um ein Gebärdensprachkorpus zu erstellen, ist die enge Einbindung dieser Gemeinschaft und ihrer SprachbenutzerInnen während der Videoaufzeichnungen natürlicher gebärdensprachlicher Diskurse wesentlich. Das gilt auch für ihre Einbindung in den Annotationsprozess[9] der gebärdensprachlichen Äußerungen. Dadurch kann die Innensicht auf diese Sprache samt ihrer sprachlichen und kulturellen Spezifika gewährleistet und die Einbeziehung der gehörlosen MuttersprachlerInnen für die zukünftige Nutzung des Korpus und daraus resultierender Forschungsergebnisse sichergestellt werden. Wie diese Einbindung im Detail aussieht, soll nun näher beschrieben werden.

2.1.1 Österreichische Gebärdensprache als Sprache der Gehörlosengemeinschaft

Wie bereits erwähnt, ist die ÖGS in Österreich seit 2005 offiziell als Sprache anerkannt.[10] Sie wird von rund 8.000 – 10.000 gehörlosen Gebärdensprach-

9 Einträge zur Beschreibung von Gebärdensprachkorpora werden als „Annotationen“ definiert (Johnston 2010).

10 An dieser Stelle sei auf den Artikel „Die Relikte von Oralismus und Behindertendiskriminiuerng in Österreich“ von Dotter et al. (2019) hinzuweisen, der zeitgleich mit diesem Artikel entstand / in Druck war und wesentliche Aspekte zur Gebärdensprachrechte in Österreich beschreibt.

benutzerInnen verwendet (vgl. Schalber 2015: 105). Überdies ist die ÖGS die Sprache der Österreichischen Gehörlosengemeinschaft. Sie definiert sich aus den zahlreichen regionalen Varietäten der ÖGS, welche durch spezifische Charakteristika geformt und beeinflusst sind. Wesentliche Faktoren solcher Formen und Einflüsse sind das Verwenden bzw. Nicht-Verwenden dieser Sprache in bestimmten sozialen Kontexten, Einfluss und Netzwerk der Sprachnutzung zwischen Generationen, sowie lokale und regionale Gegebenheiten. Vor allem aber sind das ständige Umgebensein von Lautsprache und daraus resultierender Einflüsse, wie das Spracherwerbsalter, früh-/kleinkindliche und schulische Ausbildung (und das damit verbundene Verwenden, teilweise oder nicht Verwenden der ÖGS), sowie soziale und kulturelle Komponenten wesentliche Einflussfaktoren auf die Sprache.Viele Gebärdensprachen sind eng mit der Gründung von Gehörlosenschulen aber auch mit Gehörlosenverbänden/Gehörlosenvereinen verbunden, da es dadurch zu einer kritischen Gruppe an potentiellen GebärdensprachbenutzerInnen kam und die Sprache sich (weiter-) entwickeln konnte (vgl. Dotter 2012). Die erste Gehörlosenschule, das *k.k. Taubstummeninstitut*, wurde 1779 in Wien gegründet. Gehörlose SchülerInnen der gesamten österreichisch-ungarischen Monarchie wurden dorthin entsandt. Über die folgenden Jahrzehnte wurden Tochterschulen[11], über die gesamte Monarchie verteilt, gegründet. Die Lehrerausbildung verblieb in Wien oder wurde von Schule zu Schule weitergetragen (vgl. Venus 1854; Schott 1995; Rössl 1956; Wasserstein 2012; Dotter 2012; Schalber 2015). Diese historische Weiterentwicklung dürfte für eine mögliche strukturelle Ähnlichkeit bzw. Nähe zwischen etlichen nicht-manuellen Elementen, die syntaktische Konstruktionen begleiten, oder einzelnen lexikalischen Einheiten innerhalb

11 Gehörlosenschulen in Österreich wurden in Wien (1779), in Linz (1812), in Mils (Nahe Hall in Tirol, 1830), in Salzburg (1831), in Graz (1831), in St. Pölten (1846) und in Klagenfurt (1847) gegründet (vgl. Venus 1854; Schott 1995; Rössl 1956; Wasserstein 2012; Dotter 2012). In Mikulov/Nikolsburg in der Tschechischen Republik, wurde 1844 eine Jüdische Gehörlosenschule gegründet, die 1852 nach Wien verlegt, aber 1926 geschlossen wurde (vgl. Wasserstein 2012: 189; Schott 1999; Venus 1854: 92). Zur selben Zeit wurden auch Gehörlosenschulen in den folgenden Gebieten der ehemaligen österreichisch-ungarischen Monarchie gegründet: Prag, Tschechische Republik (1786), Vác, Ungarn (1802), Milan, Italien (1805), Brno, Tschechische Republik (1829), Brixen/Bressanone, Italien (1830), Lviv, Ukraine (1830), Bratislava, Slowakei (1833), Gorica/Gorizia, Slowenien/Italien (1842), Trient, Italien (1842) (vgl. Venus 1854; Dotter 2012; Corazza & Lerose 2008a).

der ÖGS und anderen Gebärdensprachen der ehemaligen Donaumonarchie sein (vgl. Dotter 2012 und Schalber 2015: 106 zur Historie der ÖGS; Corazza & Lerose 2008b zu ähnlichen Klassifikatorskonstruktionen in der ÖGS und der Italienischen Gebärdensprache, LIS, verwendet in Triest; Šarač Kuhn et al. 2007 zu gemeinsamen Merkmalen zwischen ÖGS und der Kroatischen Gebärdensprache, HZJ). Ausführlichere Studien zu dieser Nähe zwischen den einzelnen Gebärdensprachen unserer Nachbarländer werden mit zunehmender Erstellung von Gebärdensprachkorpora in diesen Ländern in naher Zukunft möglich sein. Es sei aber erwähnt, dass auch andere Gründe, wie die Vererbung von Gehörlosigkeit und die Nutzung einer visuell-gestischen Sprache über Familiengenerationen hinweg, für die ÖGS wesentlich waren und es noch immer sind.

In der Lexikondatenbank zur ÖGS, LedaSila[12] gibt es rund 31.500 Einträge (Stand 05/2019), die aufgrund der Herkunft der in den Videos zu sehenden GebärdensprachbenutzerInnen regionalen Varietäten zugeordnet sind. Schalber (2015) illustriert, dass in den jeweiligen österreichischen Bundesländern, die diesbezügliche Varietät der ÖGS verwendet und nach dem Namen des Bundeslandes bezeichnet wird. Eingehende Untersuchungen zu regionalen Varietäten der ÖGS – basierend auf regionalen, sozialen, altersabhängigen und weiteren möglichen Einflussfaktoren – sind derzeit erst in Ansätzen zu finden (siehe Lackner 2007, 2013 und 2017 zu einer lokalen Variante der Salzburger Varietät der ÖGS, verwendet im Großarltal).

Soweit aus ersten Untersuchungen zur ÖGS bekannt und aus LedaSila ersichtlich, sind strukturelle Unterschiede zwischen den regionalen Varietäten der ÖGS im Lexikon als auch im Fingeralphabet (vgl. Schalber 2015) zu finden.

Daneben gibt es auch Forschung, die die Varietäten als gesamte ÖGS untersucht. Vorliegende auf erste Korpus-Sammlungen gestützte Studien zu nicht-manuellen Elementen – kurz „*nonmanuals*“ – der ÖGS zeigen, dass bestimmte nicht-manuelle Elemente regelmäßig bestimmte syntaktische Konstruktionen begleiten und mit bestimmten Sprachfunktionen in Verbindung

12 Vgl. https://ledasila.aau.at

gebracht werden (vgl. Lackner 2013 und 2017). Die Ergebnisse zeigen aber auch auf, dass nicht-manuelle Elemente in bestimmten Sprachkonstruktionen hinsichtlich der Häufigkeit ihres Auftretens, der Kookkurrenz mit anderen nicht-manuellen Elementen und der Verwendung unterschiedlicher nicht-manueller Elemente zwischen einzelnen GebärdensprachbenutzerInnen variieren (vgl. Stalzer 2014, Lackner 2013 und 2017). Hier bedarf es korpusbasierter Untersuchungen, um das Potential an Variation nicht-manueller Elemente zwischen einzelnen GebärdensprachbenutzerInnen (interpersonal), zwischen Generationen von MuttersprachlerInnen einer Varietät, zwischen regionalen Varietäten usw. zu ermitteln. Auch das mögliche Potential an Variation aufgrund der Gesprächsorganisation, Themenauswahl und Gesprächskontexten (bestimmte Situationen und Settings, die die Verwendung unterschiedlicher gebärdensprachlicher Register hervorrufen) soll durch das Sammeln eines breiten Spektrums an unterschiedlichen gebärdensprachlichen Diskursen gewährleistet sein.

2.1.2 „Emische Sicht" auf Sprache

Um Zugang zu einer Sprache zu bekommen und diese aus der Sicht der SprecherInnen bzw. GebärdensprachbenutzerInnen zu beschreiben, braucht es die „Innenperspektive" eines Sprachsystems. Der Begriff „emisch", geläufig in Kultur- und Sozialwissenschaft, definiert eine solche einzunehmende „Innenperspektive" als einen sich im System befindenden Insider und steht dem Begriff „etisch" – der die Sicht von Außenstehenden meint – gegenüber. Die Terminologie „emisch" versus „etisch" stammt ursprünglich aus der Sprachwissenschaft, eingeführt von Pike (1967) und abgeleitet vom Begriffspaar phonemisch versus phonetisch. Pike folgend stellt ein emischer Zugang der Sprachbeschreibung die kategoriale Sicht auf die sprachliche Praxis des/r Sprechers/in bzw. des/der Gebärdensprachbenutzers/in innerhalb seines/ihres Sprachsystems, in welchem diese/r partizipiert, dar (vgl. Maas 2011/12 zur emischen und etischen Sprachbeschreibung von Prosodie).

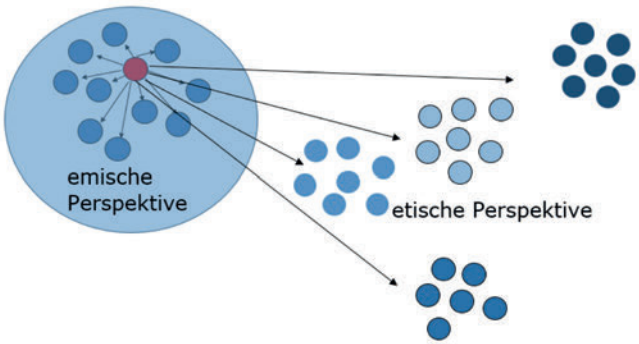


Abb. 1: Emische und etische Perspektive[13]

Wie graphisch verdeutlicht, nimmt eine Gewährsperson (verdeutlicht durch den roten Punkt) eine Innensicht zur Sprachpraxis anderer Personen im System (eingekreist im blauen Feld) ein. Gegenüber SprecherInnen bzw. GebärdensprachbenutzerInnen anderer Gruppen (verdeutlicht durch die Punktansammlungen außerhalb des blauen Kreises) nimmt sie eine etische Perspektive, also eine Sicht außerhalb des Systems, ein. Der sprachtheoretische Ansatz des Einnehmens einer „emischen Perspektive" bildet den Rahmen der Korpuserstellung zur ÖGS und wird sowohl während der Datenaufnahmen als auch während des Annotierens der gebärdensprachlichen Äußerung berücksichtigt. Dieser erstmals emische Zugang zur Sprachbeschreibung von Gebärdensprachen ermöglicht es, manuelle und nicht-manuelle Elemente, die simultan, sequenziell und/oder dreidimensional auftreten, seitens gehörloser MuttersprachlerInnen zu identifizieren und deren Funktion im System zu bestimmen.

Lucas et al. (2001) führte eine methodische Herangehensweise zur Aufzeichnung natürlicher gebärdensprachlicher Diskurse ein. Für den Aufbau des ÖGS Korpus wurde eben diese Herangehensweise genutzt. So konnte der Ansatz der „emischen Perspektive" bereits in der Phase der Materialerstellung gewahrt werden. Gehörlose MuttersprachlerInnen einer Varietät einer Gebärdensprache bestimmen nach Lucas et al. (2001) selbst die zu filmenden gehörlosen

13 Da die Online-Version in Farbe ist, enthalten die graphischen Beschreibungen die jeweiligen Farbhinweise.

InformantInnen, unterstützen die Aufnahme/Niederschrift der Metadaten und leiten den Aufnahmeprozess, indem sie die InformantInnen zu den von ihnen gewünschten zu filmenden Diskursen hinführen und als Gegenüber in Monologen fungieren. Diese Vorgangsweise führt zu in einzelnen Gebärdensprachen üblichen Diskursen und reduziert außersystemische Einflussfaktoren.

Während des derzeitigen Prozesses der Sprachbeschreibung innerhalb des ÖGS-Korpus werden Einheiten mit Satzstatus von gehörlosen AnnotatorInnen segmentiert und deren Bedeutung mittels Paraphrase in ÖGS wiedergegeben. Zusätzlich werden potenzielle Funktionen, die mit der syntaktischen Einheit in Verbindung gebracht werden, beschrieben. Auch Form und Bedeutung/Funktion nicht-manueller Elemente, welche diese satzähnlichen Einheiten begleiten, werden von gehörlosen AnnotatorInnen identifiziert und beschrieben. Nach dieser „emischen" Beschreibung der Sprache erfolgt in der Analyse eine etische Perspektive auf die Annotationsresultate in Hinblick auf den Sprachvergleich.

2.2 Die Architektur von Gebärdensprachen

Um internationale Vergleichbarkeit zu ermöglichen, muss auch die etische Perspektive auf die untersuchte Gebärdensprache gewahrt werden. Die Basis dafür bildet die emische Perspektive, aus der das ÖGS-Korpus entstanden ist. Der Sprachvergleich von ÖGS mit anderen Gebärdensprachen und Lautsprachen setzt Wissen über die allgemeine Architektur von Gebärdensprachen voraus. Es soll daher nun ein Überblick darüber gegeben werden, wie Gebärdensprachen im Allgemeinen aufgebaut sind, wobei besonderes Augenmerk auf der ÖGS liegen soll.

2.2.1 Manuelle und nicht-manuelle Komponenten

Während einer gebärdensprachlichen Äußerung sind sowohl die Hände (inklusive Bewegungen der Arme) als auch weitere Körperteile wie Kopf, Oberkörper, Augenbrauen (nicht-manuelle Artikulatoren) in Bewegung. Die Hände führen dabei die Gebärden aus, stets begleitet von simultan ablaufenden Bewegungen einzelner oder mehrerer nicht-manueller Artikulatoren (nähere

Ausführungen zu nicht-manuellen Elementen in Gebärdensprachen und insbesondere in der ÖGS sind beschrieben in Lackner 2021a, 2021b und Wilcox & Lackner 2021). Da nicht-manuelle Bewegungen gleichzeitig mit manuellen Bewegungen – die vorrangig lexikalische Einheiten konstituieren, aber auch zu anderen Funktionen genutzt werden können, insbesondere was die Ausführungen der zweiten Hand betrifft – ausgeführt werden und da nicht-manuelle Bewegungen bestimmte Funktionen innehaben können, sind viele GebärdensprachforscherInnen geneigt, diese nicht-manuellen Bewegungen mit Intonation in Lautsprachen zu vergleichen (vgl. Crasborn 2006; Dachkovsky & Sandler 2009).

Bei Betrachtung von gebärdensprachlichen Diskursen fällt jedem Laien sofort auf, dass GebärdensprachbenutzerInnen beide Hände zum Artikulieren verwenden. Das Zusammenspiel beider Hände – um beispielsweise lexikalische Einheiten oder morpho-syntaktische Klassifikatorkonstruktionen zu konstruieren – kann sich unterschiedlich gestalten. Die zweite Hand kann die Ausführung der ersten widerspiegeln, als „nicht-dominante" Komponente fungieren, aber auch selbständig sprachliche Information transportieren. Es zeigt sich bei GebärdensprachbenutzerInnen, dass meist eine Hand (rechte oder linke) die dominante ist, es aber durch diskursive oder interaktive Erfordernisse zu einem „*Handshift*" – einem Wechsel der Handdominanz – kommen kann.[14] Dies führt in der ÖGS-Korpus-Annotation dazu, dass im Regelfall lediglich die „dominante Hand" annotiert wird. Erst wenn die zweite Hand zusätzliche oder andere Funktionen innehat, wird auch diese annotiert.

Nicht-manuelle Komponenten können unterschiedliche Sprachfunktionen in Bezug auf das Lexikon, die Morphologie, Syntax und Pragmatik innehaben (vgl. Herrmann & Steinbach 2013).[15] Untersuchungen zur ÖGS zeigen, dass etliche nicht-manuelle Elemente Negation und Assertion, Interrogativi-

14 Einige Untersuchungen zum möglichen Zusammenspiel der beiden Hände in Gebärdensprache ist zu finden u.a. bei Crasborn & Sáfár (2016) oder Papadatou-Pastou & Sáfár (2016) oder Siyavoshi (2017).

15 Sandler (2011) oder Dotter (to appear) folgend, sollten *Nonmanuals* nicht als natürliche Klasse gesehen werden, da einzelne (Gruppen an) nicht-manuelle Elemente unterschiedlich auftreten können. *Nonmanuals* können als adjektivische oder adverbielle Morpheme interpretiert werden, Teil des Lexikons sein, als Diskurselemente fungieren, paralinguistischen oder affektiven Ausdrücken entsprechen oder eine äquivalente Funktion zu Prosodie und Intonation in gesprochenen Sprachen besitzen.

tät, Konditionalität, epistemische Modalität und andere Funktionen innehaben können oder in solchen funktionalen Satzkonstruktionen auftreten (vgl. Lackner 2013 und 2017; Schalber 2006; Stalzer 2014). Eine Untersuchung zu Segmentierungsanzeigern in gebärdensprachlichen Diskursen der ÖGS zeigt, dass die Wahrnehmung von Grenzsignalen im gebärdensprachlichen Diskurs (oftmals verglichen mit Intonationseinheiten in gesprochener Sprache) zwischen GebärdensprachbenutzerInnen und jenen, die keine Gebärdensprachkompetenz aufweisen, stark variieren. Die Ergebnisse zur Auflistung möglicher Grenzsignale zeigt, dass 90 % aller Kopfbewegungen von GebärdensprachbenutzerInnen genannt werden, aber lediglich 10 % von Personen ohne Gebärdensprachkompetenz (vgl. Lackner et al. 2017). Dies zeigt, dass die emische Perspektive in der Identifikation möglicher nicht-manueller Elemente eine wichtige Rolle in der Korpus-Annotation spielt. Die Artikulationsbereiche samt möglicher Bewegungen in der ÖGS sind in der folgenden Abb. 2 dargestellt. In Lackner (2017; 2019a) und Lackner et al. (2019) sind alle bis dato von gehörlosen MuttersprachlerInnen identifizierten nicht-manuellen Elemente nach Artikulationsbereichen und bezüglich der Kopf- und Körperbewegungen entsprechend der möglichen Ausführungsachsen aufgelistet.

Artikulationsbereich	Bewegungselemente
Körper	Neigen und Drehen des Oberkörpers
Kopf	Neigen und Drehen des Kopfes
Augenbrauen	Heben, Senken und Zusammenziehen der Augenbrauen
Augen	Blickrichtung, Blinzelbewegung, Öffnungsgrad der Augen
Gesicht	Naserümpfen, Stirnrunzeln, Aufblasen/Einziehen der Wangen
Mund	Bewegungen des Mundes (Mundbild, Mundgestik)

Abb. 2: Nicht-manuelle Artikulationsbereiche und Bewegungselemente im Überblick

Um die Vielfalt der Möglichkeiten des Auftretens nicht-manueller Elemente und damit einhergehender Herausforderungen in der Korpus-Annotation zu

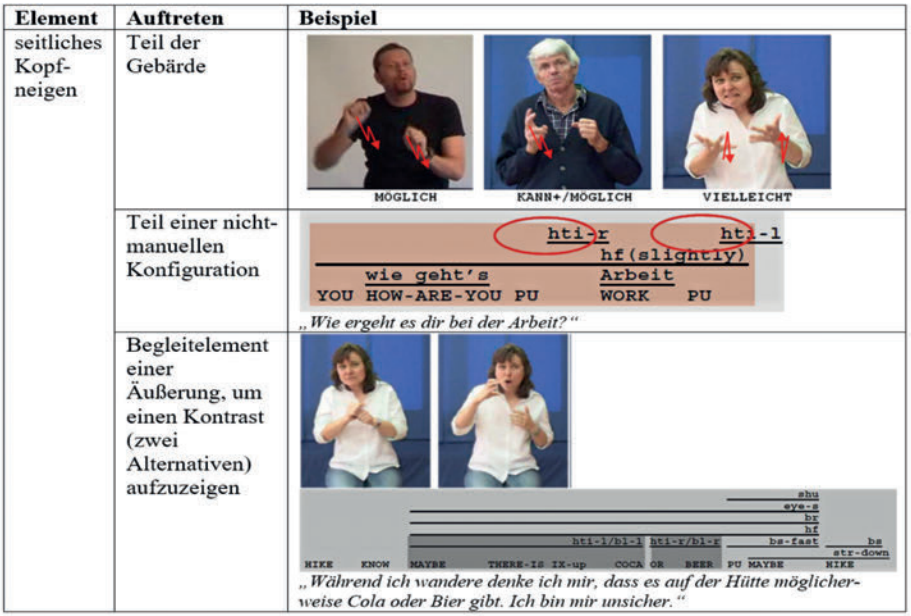


Abb. 3: Nicht-manuelles Element „seitliches Kopfneigen"[16]

verdeutlichen, sei exemplarisch ein nicht-manuelles Element und einige mögliche Kontexte seines Auftretens im Folgenden veranschaulicht.

2.2.2 Gebärdensprachlich-strukturelle Dimensionen

Gebärdensprachen stellen eine visuell-gestische Sprachmodalität dar und weisen die Besonderheit auf, dass sprachliche Elemente nicht nur sequenziell, sondern auch simultan ausgeführt werden. Überdies machen sie sich den dreidimensionalen Raum zu Nutze. Diese genannten, den Gebärdensprachen zugrundeliegenden Strukturphänomene betreffen sowohl die manuellen Komponenten der Sprache, also die Ausführungen der Hände (inklusive Arme) – meist als „Gebärden" benannt –, als auch die nicht-manuellen Komponenten. Nicht-

16 Abbildungen der Gebärden stammen aus LedaSila [Screenshootsbearbeitung von Lackner Andrea] bzw. aus Veröffentlichung von Lackner (2013; 2017; in Druck).

manuelle Elemente sind Haltepositionen des Kopfes, des Körpers (vorrangig Bewegungen des Oberkörpers) und Bewegungen einzelner Artikulatoren des Gesichtsfeldes – das sind spezielle Bewegungen des Mundes, der Augenbrauen und der Nase, das Schließen und Öffnen der Augenlider, das Zusammenziehen der Augenpartie, die Blickrichtung und deren Richtungswechsel.

Die Spezifität der drei sprachstrukturellen Dimensionen *„Simultanität“*, „*Sequenzialität“* und *„Dreidimensionalität“* in der Ausführung der manuellen und nicht-manuellen Komponenten wird nun im Einzelnen erläutert und zur Veranschaulichung mit ÖGS-Beispielen graphisch dargestellt. Diese explizite Ausführung gebärdensprachlicher Architektur ist deswegen von Interesse, weil diese Aspekte bei der Erstellung eines systematisch aufgebauten und nach Sprachelementen durchsuchbar gemachten Korpus stets mitzuberücksichtigen sind.

2.2.2.1 Simultanität

Die simultane Dimension in Gebärdensprachen zeigt sich in der Ausführung der einzelnen Gebärden und in der Umsetzung gesamter gebärdensprachlicher Äußerungen. Wie im Folgenden näher ausgeführt, wird die Simultanität innerhalb der manuellen (teilweise auch innerhalb der nicht-manuellen) Komponenten einer Gebärde aufgezeigt und im Anschluss das simultane Zusammenspiel manueller und nicht-manueller Komponenten in gebärdensprachlichen Äußerungen veranschaulicht.

Seit Stokoe[17] (1960) werden Gebärden und folglich insbesondere die manuelle Komponente der Gebärdenspracharchitektur im Sinne eines Parametermodells gesehen. Diese Parameter umfassen die Konfiguration der Hand (Handform bzw. Handkonfiguration), die Orientierung der Hand-/Fingerrücken (Handorientierung), die Ausführungsposition in Beziehung zum Oberkörper/Kopf (Ausführungsstelle) und die Bewegungskomponente/n (Bewegung). Diesen Parametern ist eine potentiell (un)endliche Anzahl an Elementen zugeordnet.

17 Stokoe (1960) führte als Erster in der Gebärdensprachforschung ein Parametermodell, definiert als Aspektmodell, ein. Er ging davon aus, dass eine Gebärde aus drei gleichzeitig realisierten Komponenten der Parameter Handform („DEZ“ von *designator*), Ausführungsstelle („TAB“ von *tabula*) und Bewegung („SIG“ von *signation*) bestehen. Stokoe führt den Begriff „*Chereme*“ (griechisch χείρ – „Hand“) ein und vergleicht diese mit Phonemen in Lautsprachen.

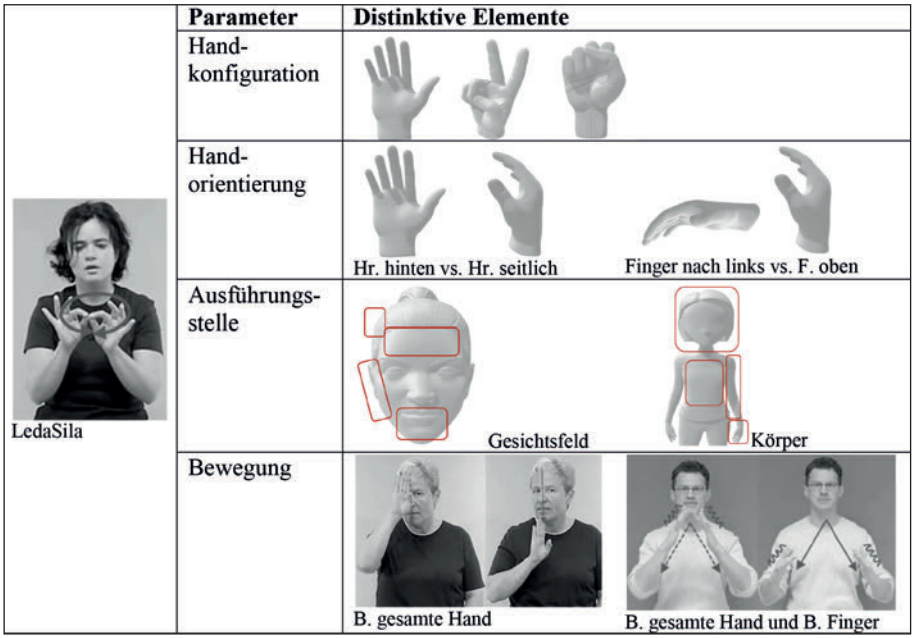


Abb. 4: Parametermodell zur Gebärdenausführung

In einer bestimmten Gebärdensprache gibt es allerdings ein begrenztes Set an Handkonfigurationen. LedaSila führt für die ÖGS in seiner Auswahlmaske verschiedene Handformen an (164 insgesamt).[18] Die Orientierung kann sowohl die des Handrückens als auch die des Fingerrückens betreffen. Die Ausführung der Gebärde wird mit bestimmten möglichen Lokationen am Körper in Beziehung gesetzt. Es gibt sowohl Bewegungen der gesamten Hand (mit/ohne Armbewegung) als auch Bewegungen der Handorientierung und der Finger (vgl. Crasborn 2017 zur sublexikalischen Struktur in Gebärdensprachen, van der Kooij & Crasborn 2008 zur Silbe in Gebärdensprachen).

18 Eine ausführliche Untersuchung zur genauen Anzahl an Handformen der ÖGS ist bis dato noch ausständig. Riemer-Kankkonen (2019) zeigt in seiner Studie zur Anzahl und Verwendung von Handformen zweier regionaler Varietäten der ÖGS auf, dass es genauerer Kriterien und Kategorien zur Bestimmung von Handkonfigurationen bedarf, um eine systematische Beschreibung dieser vornehmen zu können.

In der ÖGS gibt es zahlreiche Gebärden, die sich durch eine Veränderung einzelner manueller Parameter voneinander unterscheiden. Die steirische Gebärde für SOMMER und VATER unterscheidet sich beispielsweise nur durch die Handorientierung, VATER und PAPA durch die Handkonfiguration. In Abb. 5 sind regionale Varianten für SOMMER veranschaulicht. Diese unterscheiden sich durch Veränderung eines Elementes eines Parameters. Die ersten beiden Gebärden unterscheiden sich aufgrund der Bewegungsrichtung, die zweite und dritte aufgrund unterschiedlicher Handkonfiguration, die erste gegenüber der vierten aufgrund der Handrückenorientierung, die vorletzte und letzte aufgrund des Bewegungselementes.

Abb. 5: Varianten für die Gebärde SOMMER

Simultan treten nicht nur die einzelnen Parameter in „einer Hand“ auf – wie dies bei einer einhändigen Gebärdenausführung der Fall ist. Auch die zweite Hand wird bei vielen Gebärden – in Form einer beidhändigen Gebärdenausführung – miteinbezogen. Diese beidhändige Ausführung unterliegt einem bestimmten Regelwerk, wie in Abb. 6 exemplarisch dargestellt.

Battison (1978) stellt grundlegend fest, dass die Handkonfiguration (einschließlich der Handorientierung) ident ist, sobald zwei Hände während einer Gebärdenausführung in Bewegung sind. Wenn die Handkonfiguration (einschließlich der Handorientierung) unterschiedlich ist, dann ist eine Hand in Bewegung, während die zweite Hand eine statische Position einnimmt. Dieser Erkenntnis folgend, unterteilt van der Hulst (1996) Gebärden in symmetrische und asymmetrische Gebärden. In asymmetrischen Gebärden fungiert eine Hand als „dominante Hand“, die zweite als „nicht dominante Hand“.

Nicht nur die zweite Hand, sondern auch nicht-manuelle Elemente können simultan während einer Gebärde ausgeführt werden und Teil einer Ge-

bärde sein. Solche Nonmanuals werden in der Literatur oft als „lexikalische Marker“ bezeichnet (vgl. Lackner 2017: 103 zu TİD).

Eine korpusbasierte Untersuchung zu Kopf- und Körperbewegungen der ÖGS zeigt, dass die Modalgebärden VIELLEICHT, MÖGLICH, KANN/MÖGLICH stets von einem seitlichen Kopfneigen begleitet werden (Lackner 2017: 180). Korpusbasierte Untersuchungen zu Negationsgebärden und semantisch-pragmatischen negativ gewerteten Gebärden in der ÖGS verdeutlichen, dass einige dieser Gebärden stets von einer Ausblasbewegung des Mundes begleitet werden (vgl. Stalzer 2014; Lackner 2014).

Abb. 6: Gebärden begleitet von der Mundbewegung „Ausblasen“

In einer gebärdensprachlichen Äußerung können unterschiedliche Elemente der manuellen und nicht-manuellen Parameter simultan auftreten, einzelne oder kompositionelle Sprachstrukturen aufzeigen und damit einhergehend unterschiedlichen Funktionen dienlich sein.

Die nicht-manuellen Elemente können eine oder mehrere Gebärden in einer Äußerung begleiten. Sie können allein oder kompositorisch (mehrere zusammen) auftreten. Es besteht die Möglichkeit, dass ein Artikulator wie der Kopf mehrere Bewegungen, die distinktiv wahrgenommen und denen unter-

schiedliche Funktionen zugewiesen werden, ausführt. Die genannten Phänomene sollen im folgenden Beispiel (Abb. 7) verdeutlicht werden.

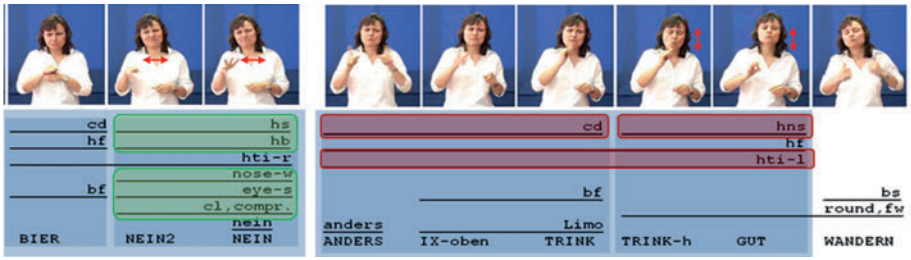


Abb. 7: Nicht-manuelle Elemente begleiten eine gebärdensprachliche Aussage

Die Bilderreihe in Abb. 7 zeigt eine Abfolge an Gebärden, die während einer gebärdeten Überlegung ausgeführt werden. In diesem Beispiel drückt eine Person aus, dass sie wandert und gleichzeitig überlegt, welches Getränk sie wohl später (auf einer Almhütte) trinken wird. Einzelne Gebärden werden ausgeführt und von einigen nicht-manuellen Elementen begleitet. In den ersten beiden grünmarkierten Blöcken werden zwei unterschiedliche, aufeinanderfolgende Negationsgebärden von einem Runzeln der Nasen (*nose-w*), zusammengezogener Augenpartie (*eye-s*) und zusammengepressten Lippen (*cl.compr.*) begleitet. Gleichzeitig führt der Kopf zwei unterschiedliche Bewegungen aus, ein Zurückneigen des Kopfes (*hb*) und eine länger andauernde Schüttelbewegung (*hs*). Die darauffolgenden nicht-manuellen Elemente zeigen, dass der Kopf sowohl eine längere Bewegung ausführen kann, nämlich ein seitliches Neigen des Kopfes (*hti-l*, langgezogener roter Balken), während zwei kürzere nicht-manuelle Bewegungen mit dem Kopf – ein Neigen des Kinns (*cd*, erster kürzerer rotmarkierter Balken) und ein länger andauerndes Nicken (*hns*, zweiter kürzerer rotmarkierter Balken) – zusätzlich ausgeführt werden. Diese möglichen simultan auftretenden Bewegungen müssen für eine Annotation in einem Gebärdensprachkorpus mitberücksichtigt werden.

2.2.2.2 Sequenzialität

Eine Korpus-Annotation von Gebärdensprachen soll auch die Sequenzialität der Gebärdenarchitektur widerspiegeln. Diese zeitliche Abfolge zeigt sich sowohl in der Gebärdenausführung als auch in der sequenziellen Anordnung manueller und nicht-manueller Elemente in einer gebärdensprachlichen Äußerung.

Seit Liddell & Johnson (1989) richtet sich der Fokus verstärkt auf die sequenzielle Anordnung von Halte- und Bewegungsphasen einer Gebärde. Die optimale Vorstellung davon wäre eine H-M-H (*hold-move-hold*) Sequenz (siehe Abb.8).[19]

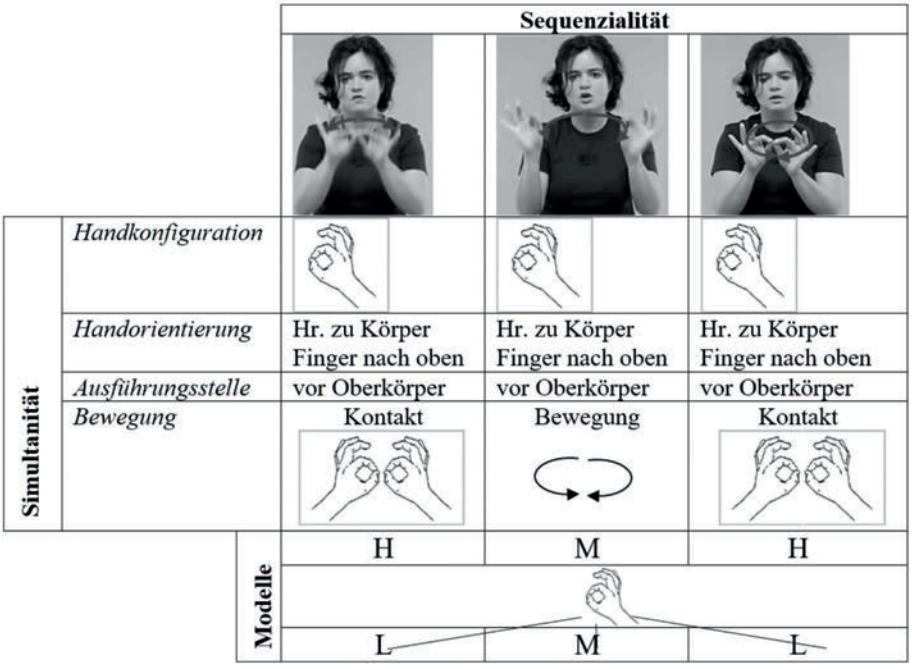


Abb. 8: Sequenz-Modell für Gebärdenausführung

19 Die verstärkte Sicht auf die sequenzielle Seite einer Gebärde führte zu Vergleichen mit lautsprachlichen Silbenkonzepten. Hier wurden Haltesegmente mit Konsonanten verglichen/gleichgesetzt (vgl. Perlmutter 1992).

Die weitere Entwicklung des vorgestellten Optimals führt zur Einführung des L-H-L (*location-move-location*) Modells bzw. des *hand-tier-Modells* (vgl. Sandler 2008). Hier wird die Handform ins Zentrum gerückt und der Anfang sowie das Ende einer Gebärde mit der Lokation (Halten wäre nur eine der vielen Möglichkeiten) in Verbindung gebracht. Brentari (1998) setzt ein Modell an, in welchem eine Gebärde „inhärente" und „prosodische" Merkmale besitzt. Die inhärenten Merkmale sind unveränderlich, die prosodischen (das sind Bewegungen der gesamten Hand, die Veränderung des Settings, der Handorientierung oder des Handöffnungsgrads) sich verändern.

Wie bereits aus Abb. 7 ersichtlich, werden Gebärden im Gebärdenraum, zeitlich nacheinander ausgeführt. Das betrifft auch die nicht-manuellen Elemente.

Abb. 9: Dreidimensionaler Gebärdenraum

2.2.2.3 Dreidimensionalität

Jede Handkonfiguration zeigt sich in dreidimensionaler Gestalt und wird im dreidimensionalen Gebärdenraum durch die Orientierung des Handrückens und der Finger, durch die Ausführungsstelle der Hände und die Bewegungskomponente spezifiziert.

Hinsichtlich des Parameters Bewegung stellen van der Kooij & Crasborn (2008) fest, dass in der Gebärdenausführung aller Gebärden der NGT (Gebärdensprache der Niederlande) eine Bewegung in Form einer „lokalen Bewegung" (*local movement*) oder einer „Wegbewegung" (*path movement*) enthalten ist. Bei Letzterer bewegt sich der gesamte Artikulator im Gebärdenraum. Bei lokalen Bewegungen sind entweder die Finger(gelenke) aktiv oder die Orientierung der Hand (meist resultierend aus der Bewegung des Unterarms, aber auch des Handgelenks). Diese einfachen Bewegungen können zu komplexen Bewegungstypen kombiniert werden. Beispiele hierfür sind Schließen und Öffnen der Hand-, Finger-, Rotations- oder Kontaktbewegungen.

Wegbewegungen können auch in der ÖGS prototypische Muster wie geradlinig, bogenförmig, kreisförmig, wellenförmig oder spiralförmig aufweisen. Lokalbewegungen sind bestimmte Veränderungen der Fingerkonfiguration wie Öffnen/Schließen, Ausstrecken/Einziehen der Finger oder das Bewegen der gesamten Finger. Wenn sich die Handkonfiguration verändert, führt das in den bisher untersuchten Gebärdensprachen stets zu einer Veränderung der gesamten Fingerkonfiguration, nicht aber der einzelnen Finger.[20] Veränderungen in der Orientierung sind schwer in Form von kontrastierenden Merkmalen zu umschreiben, da etliche dieser Bewegungen gemeinsam auftreten.

Abb. 10: Weg- und Lokalbewegungen in einzelnen Varietäten der Gebärde Dezember21

Die Dreidimensionalität beschränkt sich in der Gebärdenausführung nicht nur auf die manuelle Komponente. Auch die nicht-manuelle besitzt Räumlichkeit, veranschaulicht durch mögliche Oberkörperbewegungen (siehe Abb.11). Räumliche Veränderungen nicht-manueller Elemente können Einfluss auf die oben genannten Parameter der manuellen Komponente haben, wie beispiels-

20 Eine Ausnahme stellen Begriffe dar, die in Form eines „Finger-Alphabets" wiedergegeben werden bzw. daraus lexikalisierte Gebärden, welche noch zum Teil diese sequenzielle Abfolge aufweisen.

21 Interessant ist, dass unterschiedliche Bewegungstypen zu Bewegungsmustern, die visuell als gleiche Form wahrgenommen werden, führen können (vgl. DEZEMBER 2, 5 und 6).

weise die relative Beziehung zur Aufführungsstelle, wenn sich der Oberkörper anders positioniert oder bewegt. Die folgende Abb. 11 veranschaulicht die nicht-manuelle Komponente „Oberkörper“ und mögliche distinktiv wahrgenommene Bewegungen oder Positionierungen des Oberkörpers in Form von seitlicher Neigung (erstes Bild), Vor- oder Rückwärtsneigung (zweites Bild) oder von seitlicher Drehbewegung (drittes Bild).

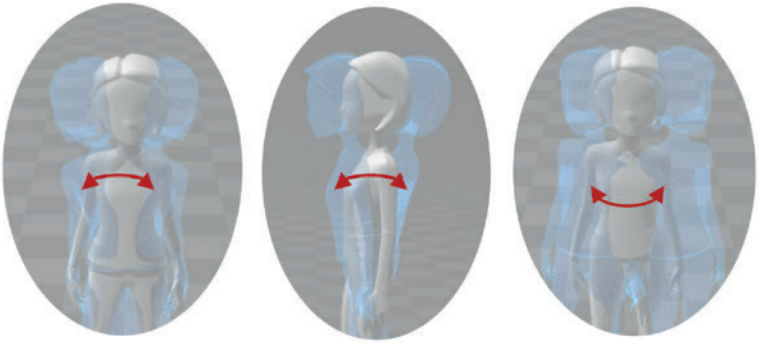


Abb. 11: Dreidimensionalität in der nicht-manuellen Komponente „Oberkörper“

Einige Gebärden können im dreidimensionalen Raum modifiziert werden. Außerdem kann die Dreidimensionalität des Gebärdenraums während einer sequenziell-ablaufenden gebärdensprachlichen Äußerung unterschiedlich genutzt werden. Hierzu zählt eine topographische, referenzielle und temporale Nutzung des Raumes. Überdies kann der Gebärdenraum selbst als autonomer Raum für sprachliche Funktionen genutzt werden (Lackner 2017). Zur Veranschaulichung werden einige Beispiele angeführt (Abb.12 und Abb. 13).

Abb. 12: Modifizierung einer Gebärde im 3-D-Raum

In Abb. 12 ist die Gebärde für AUSTAUSCH hinsichtlich ihrer Richtung und Dimension der Bewegungsausführungen modifiziert. Dadurch werden unterschiedliche Funktionen, wie ein durative oder reziprokes Handeln, ausgedrückt.

Abb. 13: Referenzielle Nutzung des 3-D-Raumes

In Abb. 13 wird mittels Ausführung der Gebärde ZUSAMMENHANG bzw. durch Hinzeigen des Indexfingers (IX-sie) auf die Ausführungsstelle der Gebärde GEBÄRDEN+SPRACHE verwiesen (jeweils mit rotem Kreis illustriert). Zusätzlich wird mittels Blickes auf diese räumliche Stelle im Gebärdenraum (angedeutet mit den Pfeilen) und damit auf den zu referierenden Inhalt – GEBÄRDEN+SPRACHE – hingewiesen.

Für die Korpus-Annotation stellt die systematische Mitberücksichtigung der dreidimensionalen funktionalen Nutzung des Gebärdenraums eine besondere Herausforderung dar, da sich in Annotationsprogrammen wie ELAN die Sequenzialität durch die lineare Zeileneintragung der ID-Glossen und die Simultanität einer gebärdensprachlichen Äußerung durch Einträge in die zweidimensional übereinander angeordneten Zeilen gut veranschaulichen lässt, während die Dreidimensionalität jeweils diesen beiden ein- und zweidimensionalen Optionen zugeordnet werden muss.

3 Das ÖGS-Korpus im Entstehen

Die vorgestellten sozio-kulturellen Dimensionen, in denen die ÖGS eingebettet ist und ihre Architektur als Sprache, die visuell-manuell organisiert ist, beeinflussen den Umgang mit ihr als Forschungsgegenstand.

Die Entstehung und der Aufbau des ÖGS-Korpus sollen nun vorgestellt werden, wobei die genannten Faktoren mitberücksichtigt werden. Der Aufbau wird allgemein beschrieben, danach soll genauer auf den Annotationsvorgang der gesammelten Daten eingegangen werden.

3.1 Aufbau des ÖGS-Korpus

Das ÖGS-Korpus umfasst derzeit Aufnahmen von 50 gehörlosen MuttersprachlerInnen der ÖGS (Stand 05/2019) in einem Zeitumfang von zirka 50 Stunden, gerechnet in Rohaufnahmen (Lackner 2020a). Die gesammelten Korpusdaten werden am *Max Planck Institute for Psycholinguistics* (MPI) in Nijmegen langzeitarchiviert[22]. Im Rahmen des FWF-Projekts „Der Beitrag nicht-manueller Elemente zur Syntax der ÖGS“ und vorausgehender Korpusdatensammlungen der Autorin dieses Artikels (vgl. Lackner 2017) ist es das Ziel, zumindest fünf Varietäten der ÖGS mit je sieben bis acht VertreterInnen im Ausmaß von 30 Stunden aufzunehmen. Die Aufnahmen werden mit drei Kameras durchgeführt (frontal, 45° und 90° Perspektive) und enthalten sowohl Monologe als auch Dialoge (teils mit mehreren beteiligten Personen). Diese Sammlung an Aufzeichnungen stellt die erste umfassendere Korpussammlung der ÖGS dar und soll mit voranschreitender, systematischer Datenannotation als Grundlage für künftige ÖGS-Untersuchungen dienen. Die geschnittenen und zu annotierenden Aufnahmen werden hierzu mit dem Annotationsprogramm (ELAN)[23] verlinkt.

22 https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0016_670C_E

23 Dieses multimediale Annotationsprogramm ermöglicht es, die unterschiedlichen Kameraperspektiven simultan ablaufen zu lassen. Die Annotationen können in *time-aligned* Zeilen eingetragen werden, die die Möglichkeit bieten, simultan auftretende gebärdensprachliche Elemente einzutragen und deren sequenziell-zeitliches und sim+ultan-kompositorisches Zusammenspiel zu veranschaulichen (vgl. Crasborn & Sloetjes 2008; ELAN 2021).

Im derzeit machbaren Umfang werden die im Korpus enthaltenen manuellen Komponenten mittels ID-Glossen annotiert (vgl. 3.2.1). Mit Stand 05/2019 sind rund 20 % der geschnittenen und für die Annotationstätigkeit aufbereiteten Aufzeichnungen geglosst.

Der Fokus des momentanen Forschungsvorhabens liegt auf der Annotation nicht-manueller Elemente und deren Funktionen in „satzähnlichen Einheiten". Die Annotation erfolgt gemeinsam mit gehörlosen MuttersprachlerInnen, basierend auf den Annotationsrichtlinien zu *Nonmanuals* von Lackner (2017; 2019a) und Lackner et al. (2019), die im Folgenden näher erläutert werden sollen. Auch wenn dieser Prozess zeitaufwendig ist, sollen bis einschließlich 2021 je 45 Minuten einer gebärdeten Variation hinsichtlich aller möglichen nicht-manuellen Parameter annotiert sein, um künftig erste aussagekräftige Untersuchungen zu einem *usage-based* und *user-annotated approach* durchführen zu können.

Die Annotation pro Video wird von je drei gehörlosen MuttersprachlerInnen der zu annotierenden Varietät der ÖGS, getrennt voneinander, ausgeführt. Dies ermöglicht das Aufzeigen der Sprachinterpretation aus Sicht mehrerer GebärdensprachbenutzerInnen (*multiple views*) und deren Übereinstimmungen (*interrater reliability*).

3.2 Annotation des Korpus

Wie in 2.2.2 beschrieben, konstituiert sich eine gebärdensprachliche Äußerung aus der sequenziellen Abfolge von Gebärden, die stets eine simultane und dreidimensionale Dimension aufweisen und dieses Potential teils auch aus morpho-syntaktischen Gründen nutzen, und auch aus simultan ausgeführten und sequenziell ablaufenden nicht-manuellen Elementen. Hinzu kommen teilweise auch zusätzliche manuelle Elemente, wenn die zweite Hand andere sprachliche Funktionen innehat.

Johnston & Schembri (2006a)[24] folgend, bedarf eine Analyse semantisch/pragmatisch-propositionaler Einheiten deren Definition als „satzähnliche

24 Die Autoren diskutieren beispielsweise den syntaktischen Status von komplexen Prädikaten, sofern diese in sequenzieller und/oder simultaner Anordnung auftreten und zwei Prädikationen aufweisen.

Einheiten“ (*clause-like units*). Dies schließt die Annotation und Analyse der sequenziellen Abfolge und der syntaktischen Rollen von Gebärden zueinander und innerhalb einer Satzkonstruktion mit ein. Eine Annotation gebärdensprachlich ausgeführter „satzähnlicher Einheiten“ bedarf auch einer genauen Annotation der zweiten Hand und der begleitenden nicht-manuellen Elemente (vgl. hierzu Vermeerbergen et al. 2007; Miller 1994 zu unterschiedlichen Simultankonstruktionen, Sáfár & Crasborn 2013 zu Simultanität der zweiten Hand, Wilbur 2000 zu nicht-manuellen Elementen).

Um diesen Aspekten gerecht zu werden, benötigt eine Satzanalyse zu Beginn die Annotation der einzelnen Gebärden.

3.2.1 Annotation der Gebärden

Es hat sich für Gebärdensprachkorpora eingebürgert, dass jede Gebärde zeitlich segmentiert und als *Token* in Form einer ID-Glosse annotiert wird. Eine ID-Glosse dient zum Identifizieren einer Gebärde und wird in Form einer bestimmten, auszuwählenden schriftsprachlichen Glosse für eine Gebärde eingetragen, die in der Annotation durch ein zusammengestelltes kontrolliertes Vokabular ausgewählt werden kann. Dies dient dazu, dass polysemische Schreibweisen (beispielsweise „Dusche“ versus „Brause“) oder vorauseilende Zuweisung von Wortarten (beispielsweise „gehörlos“ versus „Gehörlose/r“[25]) vermieden werden. Ebenso werden teilweise spezielle Gebärden eines Typs in der Glosse vermerkt, um als zusätzliches Hilfsmittel das Durchsuchen des Korpus zu erleichtern. Im ÖGS-Korpus sind beispielsweise alle Negationsgebärden mit „NEG.xxx“ (zum Beispiel „NEG.NICHT“, „NEG.NEIN“, „NEG.DARF-NEIN“) gekennzeichnet. In Anlehnung an Johnstons (2014) Annotationsrichtlinien für das *Auslan*-Korpus gibt es auch im ÖGS-Korpus die Möglichkeit, zu jeder Gebärde in ELAN weitere abhängige Zeilen mit Annotationen zu versehen. Solche abhängigen Zeilen dienen dazu, semantisch-pragmatisch spezifische, sublexikalische, morpho-syntaktische oder grammatische Informationen zusätzlich zur jeweiligen Gebärde eintragen zu können.

25 Beide Begriffe können in einer gemeinsamen Auswahlmöglichkeit als Eintrag verwendet werden.

Eine Basisannotation von Gebärdensprachkorpora beinhaltet die durchgehende Annotation der Gebärden mittels ID-Glossen, Annotation des Mundbilds und einer schriftsprachlichen Übersetzung der gebärdensprachlichen Äußerungen. Entsprechend der einzelnen Forschungsziele werden die abhängigen Zeilen der ID-Glossen mit weiteren Einträgen versehen.

Um eine syntaktische Analyse eines Korpus durchführen zu können, werden derzeit ausgewählte Gebärdensequenzen glossiert. Jede Glosse wird hinsichtlich ihrer Argumentstruktur, ihrer syntaktischen Makrorolle und ihrer semantischen Rolle innerhalb einzelner „satzähnlicher Einheiten“[26] analysiert, indem diese Rollen in die einzelnen abhängigen Zeilen zur Glosse eingetragen werden. In Anlehnung an Lehmann (2005), der sich auf Foley & Van Valin (1984) stützt, entsprechen im derzeitigen Annotationsprozess Makro-Rollen den zentralen funktionalen Rollen im syntaktischen Gefüge und damit *„actor, undergoer, indirectus“*, während „semantische Rollen“ den thematischen Rollen entsprechen.

3.2.2 Annotation der „satzähnlichen Einheiten“

Etliche Studien zu Gebärdensprachen zeigen, dass das Zusammenspiel prosodischer und syntaktischer Segmentation durch die Verflechtung der manuellen und nicht-manuellen Ausführungen in einer gebärdensprachlichen Äußerung eine besondere Herausforderung darstellt – sowohl für GebärdensprachbenutzerInnen als auch für LinguistInnen. Es bedarf daher der Bestimmung möglicher Satz-/Diskurseinheiten prosodischer, syntaktischer und semantisch-propositionaler Indikatoren (vgl. Hansen/Heßmann 2006; 2008; Fenlon 2010; Nicodemus 2009; Omel/Crasborn 2012; Mallinger 2012 und Lackner et al. 2017 zur ÖGS). Einer semantisch-funktionalen Perspektive folgend, definiert Maas (2004: 361) die gegebene Bedingung einer kontextfreien und folglich unabhängigen semantischen Interpretation eines Satzes als semantische Finitheit eines Satzes. Givón (1998) unterstreicht, dass aus einer funktionalen Perspektive syntaktische und Diskurseinheiten funktionale Bereiche aus-

26 Dieser Vorgang folgt Johnstons (2014) Annotationsrichtlinien, umgesetzt in einer Untersuchung von Hodge (2013) zur Untersuchung von *„claus-like-units“* im Auslan-Korpus.

drücken, d.h. proposition-semantische und diskurspragmatische Information. Selting (2005) wiederum unterstreicht, dass Syntax mit interaktiven Funktionen interagiert. Dieser funktionalen Sicht folgend, werden im ÖGS-Korpus Annotationen zur Bestimmung von propositionalen Einheiten durchgeführt, hier in Anlehnung an Hodge (2013) als „satzähnliche Einheiten" definiert.

a) Gehörlose MuttersprachlerInnen werden instruiert, „Bedeutungseinheiten" zu bestimmen, welche sich aus gebärdensprachlichen Ausdrücken konstituieren und unabhängig voneinander funktional interpretierbar sind.
b) Im Anschluss werden die GebärdensprachbenutzerInnen gebeten, jede einzeln bestimmte „satzähnliche Einheit" zu paraphrasieren (wird gefilmt) und eine oder mehrere sprachliche Funktionen zu bestimmen, welche mit der jeweiligen Einheit assoziiert werden. Dieser Ablauf wird anhand einer Vorlage von möglichen auszuwählenden Sprachfunktionen, die einzelnen funktionalen Domänen zugeordnet sind, durchgeführt. Lehmann & Maslova (2004) folgend, werden funktionale Domänen wie beispielsweise „Kontrast", „Illokution", „Modalität" verwendet (vgl. Lackner 2017: 65–69 für einen Überblick aller funktionalen Domänen; Lackner 2017: 27–38 zur Definition und Beschreibung von funktionalen Domänen). Um die Annotation für gehörlose MuttersprachlerInnen zu erleichtern, sind mögliche Funktionen auf einer Homepage[27] mit Videobeispielen dargestellt.

3.2.3 Annotation der nicht-manuellen Elemente

Crasborn (2014) stellt fest, dass die Annotation nicht-manueller Aktivität[28] davon abhängt, welcher Status und welche Eigenart (*nature*) nicht-manueller Aktivität in einer Theorie zuerkannt wird, welche Mittel der Operationalisierung für deren Bestimmung eingesetzt werden und wie die Reliabilität zur Bestimmung nicht-manueller Aktivität getestet wird.

27 Vgl. http://signnonmanuals.uni-graz.at/
28 Crasborn (2014) bezieht sich vorrangig auf Bewegungen des Kopfes, seine Aussage betrifft aber jeglichen Umgang mit nicht-manuellen Elementen.

Die Annotation nicht-manueller Elemente im ÖGS-Korpus erfolgt daher in zwei Schritten.

a) Gehörlose MuttersprachlerInnen werden gebeten, nicht-manuelle Elemente, insbesondere jene, die mit propositionaler Semantik, Diskurs-Pragmatik und/oder interaktiven Funktionen in Verbindung gebracht werden, zu identifizieren. Basierend auf Untersuchungen von Lackner (2013/2017) wurde eine Liste aller möglichen nicht-manuellen Elemente der ÖGS, die seitens gehörloser InformantInnen genannt wurden und die mit Sprachfunktionen in Verbindung gebracht wurden, zusammengestellt (Lackner et al. 2019). In der Korpus-Annotation sind die einzelnen nicht-manuellen Elemente den einzelnen Artikulatoren, mit welchen sie produziert werden, zugeordnet. Die Artikulatoren Kopf und Körper werden zusätzlich in alle möglichen Bewegungsrichtungen des Artikulators untergliedert, da gleichzeitig ausgeführte Bewegungen möglich sind. Beispielsweise kann der Kopf vorgestreckt und gleichzeitig eine Nickbewegung ausgeführt werden, beide nicht-manuellen Elemente haben aber ihre eigene Funktion inne. Die ELAN-Vorlage besitzt folglich die Zeilen: Augen Öffnungsgrad, Blick, Augenbrauen, Gesicht, Mundbild, Mundgestik, Kopf vor/zurück, Kopf auf/ab, Kopf drehen, Kopf neigen, Kopf Sonstige, Körper vor/zurück, Körper drehen, Körper neigen, Körper Sonstige, Schultern.

   Jede Zeile besitzt ein kontrolliertes Vokabular, resultierend aus allen bis dato genannten möglichen nicht-manuellen Bewegungen bzw. Positionen, die von den GebärdensprachbenutzerInnen mit einer sprachlichen Funktion in Verbindung gebracht werden. Beispielsweise umfasst das kontrollierte Vokabular für die bis dato identifizierten nicht-manuellen Elemente des Parameters „Augenbrauen“ die folgenden Einträge: gehoben, gehoben-stark, zusammengezogen, zusammengezogen-innen-gehoben, gehoben-zusammengezogen, eine-Braue-gehoben.

b) Als zweiten Schritt werden gehörlose InformantInnen gebeten, Funktionen, die sie mit den jeweiligen nicht-manuellen Komponenten verbinden, zu beschreiben und diese (wenn möglich) der Liste der Sprachfunktionen zuzuordnen, die bereits zur Bestimmung jener Funktionen diente, welche mit „satzähnlichen Einheiten“ assoziiert werden. Zusätzlich können die

InformantInnen ihre individuelle Beschreibung und weitere Kommentare, wie mögliche vergleichbare Kontexte des Auftretens eines Elementes, ergänzen.

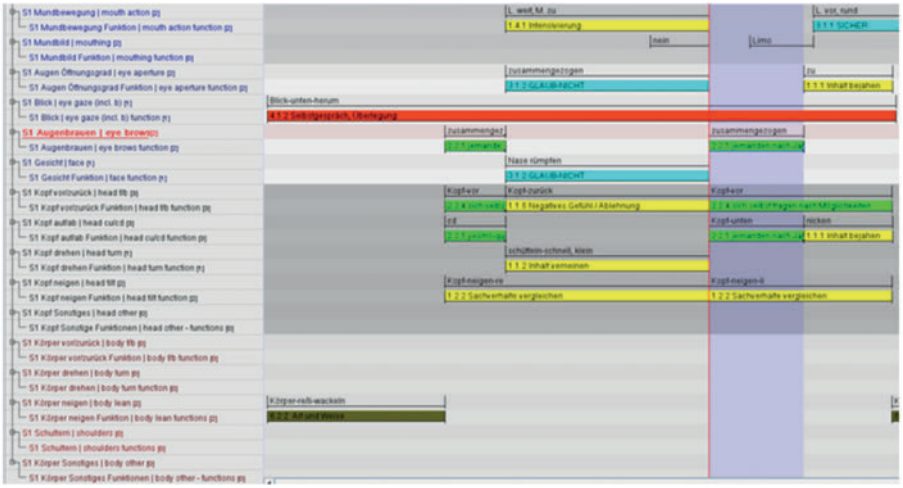


Abb. 14: Annotation der Form und zugeschriebener Funktionen in ELAN

In der Darstellung Abb.14 wird eine Annotation in ELAN veranschaulicht. In der linken Spalte sind alle möglichen Zeilen für die Annotation nicht-manueller Parameter aufgelistet. In der jeweils übergeordneten Zeile eines Zeilenpaares wird das vom gehörlosen Native wahrgenommene nicht-manuelle Element entsprechend eines auszuwählenden (kontrollierten) Vokabulars eingetragen. In der jeweiligen abhängigen Zeile kann die zugeschriebene Funktion ausgewählt werden. Die Farben signalisieren die funktionale Domäne, der bestimmte Funktionen zugeordnet sind. Daraus ergibt sich für die Analyse ein gut übersichtliches Bild. In dieser Abbildung (Abb. 14) werden beispielsweise ein Naserümpfen und ein Zusammenziehen der Augenpartie mit (epistemischer) Modalität in Verbindung gebracht (türkis unterlegt), während bestimmte Augenbrauen- und Kopfpositionen mit Interrogativität in Verbindung gebracht werden, die der funktionalen Domäne „Illokution“ zugeordnet ist (grün unterlegt).

4 Erste Ergebnisse aus der Annotation

Aus der bisherigen Annotation und Analyse das ÖGS-Korpusdaten lassen sich zahlreiche Aspekte untersuchen. So dient der Korpus dazu, bestimmte gebärdensprachliche Elemente aus emischer Sicht zu identifizieren und ihre Funktionen in der jeweiligen Konstruktion zu beschreiben. Beispielhaft sollen zwei Aspekte herausgestellt werden, die zeigen sollen, welche Analysemöglichkeiten ein Gebärdensprachkorpus bietet.

4.1 Emische Sicht auf gebärdensprachliche Elemente

Wie unter 2.1.2 und 3.2 beschrieben, wird das Korpus von gehörlosen Natives annotiert. In der Auswertung der Annotationen zeigt sich die emische Innensicht auf unterschiedliche Elemente. In Abb. 15 wird das anhand der funktionalen Verwendung des Gebärdenraumes veranschaulicht.

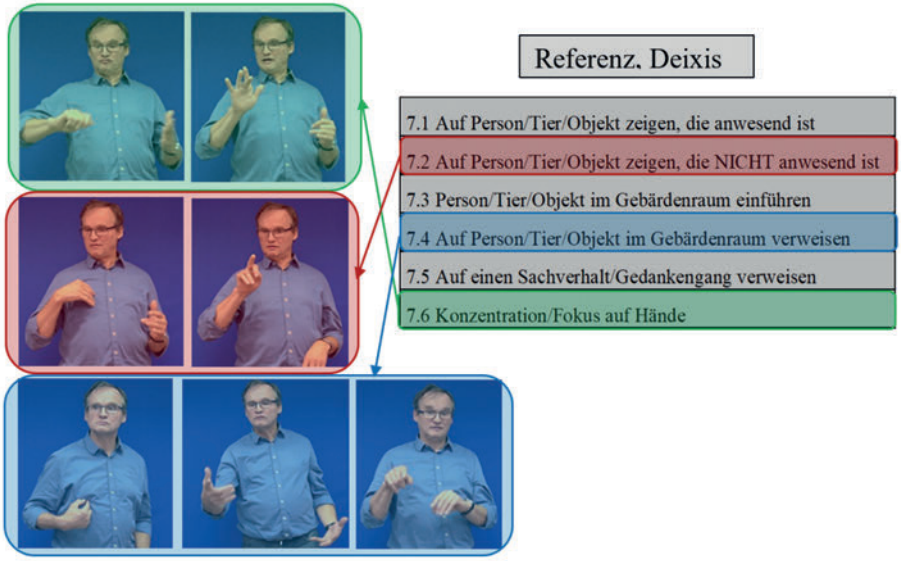


Abb. 15: Emische Perspektive auf Sprachelemente, die mit Referenz assoziiert werden

Die Bilder in Abb. 15 stammen aus einer Erzählung über einen Wanderausflug, zu welchem der Informant mit dem Auto anreiste. Sein neues Auto hat keine Schlüsselzündung, sondern wird mit Knopfdruck ein- und ausgeschaltet. In den beiden Bildern der ersten Zeile blickt der Informant auf seine rechte Hand, um auf die Gebärde – die gleichzeitig die Handhabung/Durchführung der Aktivität zeigt (Starten mit Schlüssel bzw. mit Knopfdruck) – zu referieren (grün markiert). Im Laufe der Erzählung blickt der Informant immer wieder auf den rechten Bereich im Gebärdenraum, wo sich die Knopfschaltung befindet. Dies wird als „auf ein Objekt zeigen, das nicht anwesend ist" gewertet (blau markiert). Während der Erzählung stellt sich heraus, dass der Informant vergessen hat, vor Beginn der Wanderung auf den Ausschaltknopf zu drücken und der Motor während seiner Wanderung ständig lief. Als er von einer anderen Person darauf hingewiesen wird, blickt bzw. verweist der Erzähler mit dem Indexfinger stets auf den zuvor angezeigten rechten Bereich im Gebärdenraum. Dies wird als ein Verweis auf „sein Auto" seitens der gehörlosen AnnotatorInnen interpretiert (blau markiert). Obwohl es sich immer um ein Hinzeigen/Verweisen auf denselben Punkt/Platz im Gebärdenraum und aus „etischer Perspektive" um dieselben Elemente handelt, werden den nicht-manuellen (Blickrichtung nach rechts unten) und den manuellen Elementen (Index) unterschiedliche Funktionen zugewiesen („emische Perspektive") (vgl. Wilcox & Lackner 2021).

4.2 Interrater-Reliabilität und Variation von gebärdensprachlichen Elementen

Da die Annotationstätigkeit von mehreren AnnotatorInnen zu denselben ausgewählten gebärdensprachlichen Diskursen durchgeführt wird, zeigen sich interessante Auswertungen zur Urteilsübereinstimmung bzw. -abweichung zwischen den einzelnen Personen.

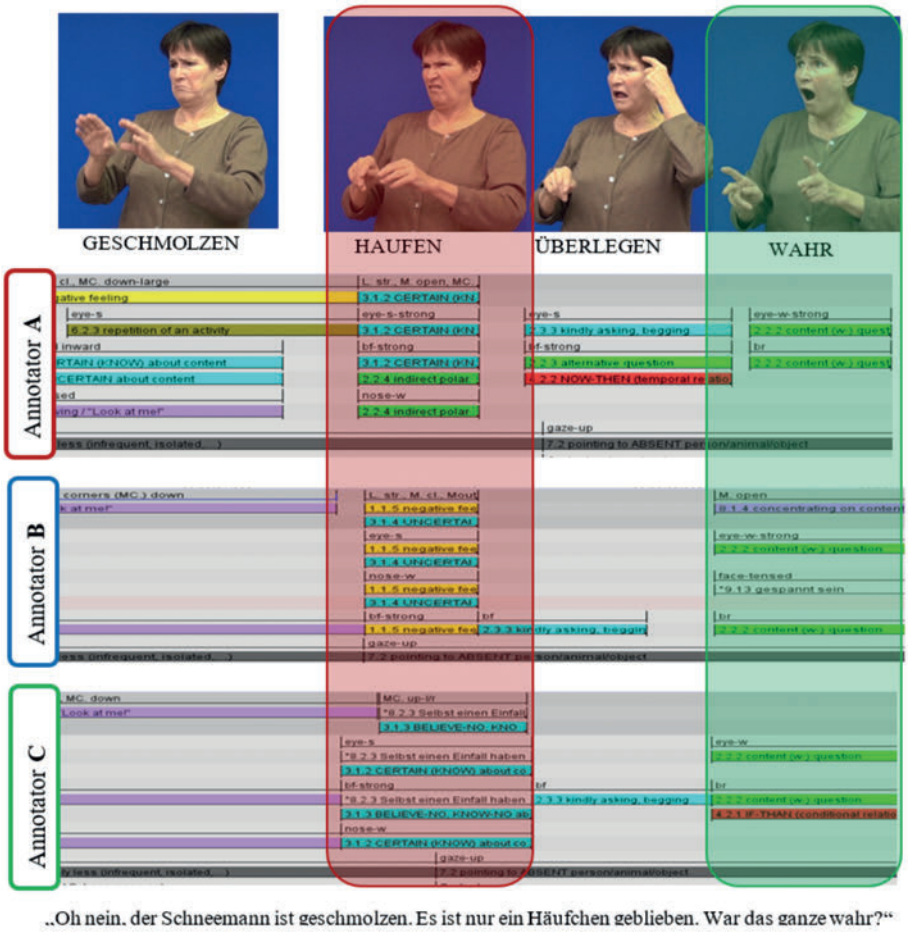


Abb. 16: Interrater-Reliabilität einer gebärdensprachlichen Äußerung

Die Abb. 16 ist der *„Snowman Story“* entnommen. Ein Junge baut einen Schneemann, der um Mitternacht zum Leben erwacht. Mit dem Schneemann unternimmt der Junge eine Reise ins Land der Schneemänner, wo er auch Santa Claus trifft, der ihm einen blauen Schal mit kleinen Schneemännern darauf schenkt. Nachdem der Junge von seiner Reise nach Hause kommt, trennt er sich vom Schneemann, der im Garten stehen bleibt, und geht schlafen. Am

Morgen ist der Schneemann geschmolzen. Der Junge überlegt, ob er die Reise nur geträumt hat und zieht den Schal aus seiner Hosentasche. Nun ist ihm klar, dass die Geschichte wahr ist. Die Gebärdenfolge in Abb.16 zeigt eine Informantin, die sich in der Erzählung soeben fragt, ob die Geschichte geträumt oder wahr sei. Dieses Sich-selbst-Befragen wird von unterschiedlichen nichtmanuellen Elementen begleitet, welche von den AnnotatorInnen (A, B und C) mit Interrogativität und (epistemischer) Modalität in Beziehung gebracht werden.

Wie auf der Abbildung ersichtlich, haben die drei AnnotatorInnen starke Übereinstimmungen in der Identifikation und funktionalen Bestimmung nichtmanueller Elemente, wie aus den rot und grün markierten Blöcken ersichtlich ist. Es zeigt sich, dass die Konfiguration aus Naserunzeln, Zusammenziehen der Augenpartie und der Augenbrauen mit „Unsicherheit" (epistemischer Modalität) verbunden wird (rot-markierter Block), das Hochziehen der Augenbrauen und die weitgeöffneten Augen mit Interrogativität (grün-markierter Block). Die Darstellung zeigt ebenso, dass einige nicht-manuelle Elemente von dem einen oder der anderen AnnotatorIn wahrgenommen wird. Inwiefern diese Unterschiede mit einem unbewussten Wahrnehmen („Übersehen") oder anderen Faktoren zusammenhängen, werden genauere Analysen der Korpus-Annotation ergeben.

Da die soeben beschriebene Geschichte von mehreren gehörlosen InformantInnen wiedergegeben wurde (es gibt eine Vielzahl an vergleichbaren aufgezeichneten Diskursen), werden künftige Auswertungen des Korpus zeigen, inwiefern die einzelnen gehörlosen Natives, aber auch einzelne Varietäten der ÖGS sich hinsichtlich der Verwendung einzelner Gebärden, bestimmter gebärdensprachlicher (syntaktischer) Konstruktionen und nicht-manueller Elemente unterscheiden (vgl. Lackner 2019b).

5 Ausblick[29]

Das ÖGS-Korpus[30] befindet sich derzeit im Aufbau. Einige Varietäten der ÖGS werden in naher Zukunft noch in das Korpus aufgenommen. Ein großer Schwerpunkt in den kommenden Jahren ist die Weiterführung der Basis-Annotationen sowie der Annotation nicht-manueller Elemente in ausgewählten gebärdensprachlichen Diskursen. Die Auswertung und Analyse ist hinsichtlich der ÖGS-Syntax und dem Beitrag nicht-manueller Elemente zu syntaktischen Konstruktionen vielversprechend. Korpusdaten werden Aufschluss über die Art des Auftretens nicht-manueller Elemente (Form, Häufigkeit, Kookkurrenz mit anderen Nonmanuals), also auch über ihre Position und Dauer des Auftretens in bestimmten Typen an Satzkonstruktionen (wie beispielsweise unterschiedlichen Interrogativen) geben. Die Form-Funktion-Assoziation von nicht-manuellen Elementen wird zeigen, welche Nonmanuals tendenziell mit welchen Funktionen in Beziehung gebracht werden. Ziel der Satzanalyse wird es sein, sequenziell-simultan-dreidimensionale Strukturen an Satzkonstruktionen und Typen von Satzkonstruktionen in Modellen zu veranschaulichen. Wesentlich wird es auch sein, die Satzanalyse über den Satz hinaus und folglich im Rahmen einer Diskursanalyse zu betrachten.

Der im ÖGS-Korpus gewählte sprachtheoretische und methodische Ansatz bietet eine gute Basis für künftige Analysen zur Sprachvariation innerhalb der ÖGS. Überdies bietet das entstehende ÖGS-Korpus die Möglichkeit

29 Dieses ÖGS-Korpus entstand aus einer Eigeninitiative im Sinne der Erstellung eines eigenen Korpus und der zweimaligen Antragstellung für Drittmittel (beim FWF). Um die Gebärdensprachforschung mit einem stabilen, systematisch aufgebauten und annotierten Gebärdensprachkorpus einen großen Schritt weiterzubringen, wird die finanzielle und institutionelle Absicherung der ÖGS-Forschung im Fokus der künftigen Arbeit liegen.

30 Zusätzlich zum ÖGS-Korpus wurde im Rahmen des SignNonmanuals-Teams in kleinem Rahmen ein erstes „ÖGS-Korpus Kindersprache“ begonnen, der eine erste Sammlung gebärdensprachlicher Diskurse hörbeeinträchtigter Kinder, die die ÖGS verwenden, ist. Zusätzlich wurde begonnen, ein erstes „ÖGS-L2 Korpus“ aufzubauen, der eine Sammlung gebärdensprachlicher Diskurse hörender LernerInnen, die die ÖGS als Zweitsprache (L2) erwerben, darstellt. Beide Korpora sind am Beginn des Aufbaus und erst teilweise mit Annotationen (Glossierung, Annotation manueller und nicht-manueller Elemente, Übersetzung in deutsche Schriftsprache) versehen (vgl. Lackner 2020b und 2020c).

des Sprachvergleichs zwischen Gebärdensprachen – insbesondere hinsichtlich der funktionalen Nutzung nicht-manueller Elemente.

Eine Ergänzung zur Korpus-Annotation wird Aufschluss über den Sprachstatus einzelner manueller und nicht-manueller Elemente geben. Folglich wird das Abfragen und Beurteilen des Sprachstatus bestimmter nicht-manueller Elemente, indem gehörlose MuttersprachlerInnen beurteilen, ob bestimmte satzähnliche Einheiten ohne Kookkurrenz bestimmter nicht-manueller Elemente wohlgeformt sind und ob bzw. welche nicht-manuellen Elemente sich gegenseitig ergänzen/ersetzen können, ein künftiges Forschungsziel sein. Diese Abfrage soll Information über die Akzeptierbarkeit, die Variation und den Grammatikalisierungsstatus einzelner Elemente innerhalb syntaktischer Konstituten geben.

Zusätzlich zum Sprachvergleich zwischen Gebärdensprachen und funktionalen Vergleichen mit Lautsprachen ist in Zukunft auch ein Vergleich mit sprachbegleitenden gestischen Elementen ein essentielles Desiderat der Sprachwissenschaft.

Korpora sollen aber nicht nur der Erforschung der Sprachstrukturen dienen. Ebenso ist es ihre Aufgabe als Grundlage für die Erstellung einer fundierten Sprachbeschreibung in Form einer benutzerfreundlichen Grammatik zu fungieren. Das Korpus-Material kann auch für Lehrende als Vorlage zum Einsatz kommen, indem unterschiedliche gebärdensprachliche Realisierungen einzelner Diskurse als Modelle möglicher sprachlicher Umsetzung dienen oder indem das Material selbst für unterschiedliche Übungszwecke eingesetzt wird. Hier braucht es aber einen behutsamen Umgang mit den Korpus-Daten. Deren Handhabung und kompetenzorientierter Einsatz, der sich auf Richtlinien und Umgangsformen stützt, muss erst erarbeitet werden.

Referenz

Battison, R. (1978): *Lexical borrowing in American Sign Language*. Silver Spring, Md.: Linstok Press.

Brentari, D. (1998): *A Prosodic Model of Sign Language Phonology*. Cambridge, MA: MIT Press.

Crasborn, O. (2006): Nonmanual structures in sign language. In: Keith Brown (Hg.*), Encyclopedia of language and linguistics*. Vol. 8, 2nd Edition. Oxford: Elsevier, 668–672.

Crasborn, O./Sloetjes, H. (2008): Enhanced ELAN functionality for sign language corpora. In: O. Crasborn/T. Hanke/E. Efthimiou/I. Zwitserlood/E. Thoutenhoofd (Hg.), *Construction and Exploitation of Sign Language Corpora*. *3rd Workshop on the Representation and Processing of Sign Languages.* Paris: ELRA, 39–43.

Crasborn, O. (2014): *Annotating the head*. Presentation at the Workshop SignNonmanuals, Klagenfurt. Manuscript.

Crasborn, O./S*áfár*, A. (2016): An annotation scheme to investigate the form and function of hand dominance in the Corpus NGT. In: R. Pfau/M. Steinbach/A. Hermann (Hg.), *A Matter of Complexity: Subordination in Sign Languages*. Berlin: Mouton de Gruyter, 231–251.

Crasborn, O. (2017): Sublexical structure. In: J. Quer/C. Cecchetto/C. Donati/C. Ceraci/M. Kelepir/R. Pfau/M. Steinbach (Hg.), *SignGram Blueprint. A Guide to Sign Language Grammar Writing*. Berlin/Bosten: De Gruyter Mouton, 22–36.

Corazza, S./Lerose, L. (2008a): L'origine della Lingua die Segni Italiana, variante triestina. In: C. Bagnara/S. Corazza/S. Fontana/A. Zuccalà (Hg.*), I segni parlano. Prospettiva di ricerca sulla Lingua die Segni Italiana*. Milano: Franco Angali, 132–139.

Corazza, S./Lerose, L. (2008b): Vergleich von Klassifikatoren in der Österreichischen und der Triestiner Gebärdensprache. In: *Gebärdensprachlinguistik und Gebärdensprachkommunikation. Referate der VERBAL-Sektion "Gebärdensprachlinguistik und -Kommunikation" innerhalb der 34. Österreichischen Linguistiktagung an der Universität Klagenfurt am 8. 12. 2006*. Klagenfurt: ZGH, 31–36.

Dachkovsky, S./Sandler, W. (2009): Visual intonation in the prosody of a sign language. *Language and Speech*, 52(2/3), 287–314.

Dotter, F. (2012): Eine kurze Geschichte von Gebärdensprache und Gehörlosenbildung in Österreich. In: *Ž. Ribičič/W. Mojca Polak/A. Trtnik Herlec (Hg.)*, *Gebärdenkommunikation mit Babys und Kleinkindern: Handbuch*. *Projekt Tiny Signers*. Ljubljana/Leeds: Zavod za gluhe in naglušne/EuroVia, 67–68.

Dotter, F./ Jarmer, H./Huber, L. (2019): Die Relikte von Oralismus und Behindertendiskriminierung in Österreich. In: M. Schmidt/A. Werner (Hg.), *Zwischen*

Fremdbestimmung und Autonomie. Neue Impulse zur Gehörlosengeschichte in Deutschland, Österreich und der Schweiz. Bielefeld: transcript Verlag, 373–422.

Dotter, F. (to appear). Sign Language. Comparing modalities: differences and commonalities. In: D. Jung/J. Helmbrecht (Hg.), *WSK 13 Linguistic Typology*. Berlin: De Gruyter Mouton.

ELAN (Version 6.2) [Computer software] (2021): Nijmegen: Max Planck Institute for Psycholingusitics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Fenlon, J. (2010): *Seeing sentence boundaries: The production and perception of visual markers signalling boundaries in sign languages.* UCL (University College London): PhD thesis.

Fenlon, J./Schembri, A./Johnston, T./Cormier, K. (2015): Documentary and Corpus Approaches to Sign Language Research. In: E. Orfanidou/B. Woll/G. Morgan (Hg.), *Research Methods in Sign Language Studies: A Practical Guide*. Oxford: Wiley-Blackwell, 156–172.

Foley, W. A./Van Valin, R. D. (1984): *Functional Syntax and Universal Grammar.* Cambridge: Cambridge University Press.

Givón, T. (1998): The Functional Approach to Grammar. In: M. Tomasello (Hg.), *The new psychology of language. Cognitive and Functional Approaches to Language Structure*. Lawrence Erlbaum Associates, 41–68.

Hanke, T./Storz, J. (2008): iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In: O. Crasborn/T. Hanke/E. Efthimiou/I. Zwitserlood/E. Thoutenhoofd (Hg.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages.* Paris: ELRA, 64–67.

Hanke, T. (2010): *HamNoSys 4 Handshapes Chart*. Drawings by Heiko Zienert, Olga Jeziorski, Andreas Hanß. In: https://www.sign-lang.uni-hamburg.de/ dgs-korpus/ files/inhalt_pdf/HamNoSys_Handshapes.pdf [25.06.2019].

Hansen, M./Heßmann, J. (2006): *Reanalysing sentences in German Sign Language (DGS)* texts. Präsentation zur 28. Jahrestagung der DGfS, Bielefeld.

Hansen, M./Heßmann, J. (2008): Matching propositional content and formal markers. Sentence boundaries in a DGS text. *Sign Language and Linguistics* 10(2), 145–175.

Herrmann, A./Steinbach, M. (2013): *Nonmanuals in sign language.* Amsterdam/Phil.: Benjamins.

Hodge, G. (2013): *Patterns form a signed language corpus: Clause-like units in Auslan (Australian sign language)*. Macquarie University: PhD thesis.

Hulst, H. van der (1996) On the other hand. *Lingua* (98), 121–143.

Johnston, T. (2010): From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15(1), 104–129.

Johnston, T. (2014) [Manuskript in ständiger Überarbeitung, Stand November 2016]: *Auslan Corpus Annotation Guidelines*. In: http://www.auslan.org.au/about/annotations/ [25.06.2019].

Johnston, T./Schembri, A. (2006a): *Identifying clauses in signed languages: applying a functional approach.* Paper presented at the DGfS workshop: How to recognise a sentence when you see one: methodological and linguistic issues in the creation of sign language corpora, 23–24 February, Bielefeld.

Johnston, T./Schembri, A. (2006b): Issues in the creation of a digital archive of a signed language. In: L. Barwick/N. Thieberger (Hg.), *Sustainable data from digital fieldwork.* Sydney: Sydney University Press, 7–16.

Kooij, E. v. d./Crasborn, O. (2008): Syllables and the word-prosodic system in Sign Language of the Netherlands. *Lingua*, 118, 1307–1327.

Lackner, A. (2007): *Turn-Taking in der österreichischen Gebärdensprache: Eine Gesprächsanalyse der Salzburger Variante*. Karl- Franzens-Universität Graz: Diplomarbeit.

Lackner, A. (2013): *Linguistic functions of head and body movements in Austrian Sign Language (ÖGS). A corpus-based analysis*. Karl-Franzens-Universität Graz: Dissertation.

Lackner, A. (2014): *Semantic relations of particular nonmanuals and the signs with which they co-occurring. Discussion paper.* Präsentation im Workshop SignNonmanuals, Klagenfurt.

Lackner, A. (2017): *Functions of and body movements in Austrian Sign Language (ÖGS).* Berlin: De Gruyter Mouton.

Lackner, A. (2019a): Describing Nonmanuals in Sign Language. *Grazer Linguistische Studien*, 91, 45–103.

Lackner, A. (2019b): *The Contribution of Nonmanuals to Clauses in Austrian Sign Language (ÖGS). Describing Nonmanuals by Using an Emic, Functional Approach.* Presentation at the Workshop „What can sign languages and deaf research add to science, art and society?“, Nov. 13, 2019, Budapest.

Lackner, A. (2020a): ÖGS-Korpus. Korpus der Österreichischen Gebärdensprache. ÖGS-Datenbank.

Lackner, A. (2020b): ÖGS-Korpus Kindersprache in Österreichischer Gebärdensprache. ÖGS-Datenbank.

Lackner, A. (2020c): ÖGS-L2 Korpus. Korpus von L2-Lerner*innen der Österreichischen Gebärdensprache. ÖGS-Datenbank.

Lackner, A. (2021a, Hg.): Sign Languages' Nonmanuals. Linguistic contributions from the international sign language workshop on nonmanuals and other sign language related issues held at the University of Graz in May 2019. *Grazer Linguistische Studien* (Special Issue), 93.

Lackner, A. (2021b): Nonmanuals in sign languages: a research dsideratum. In: Lackner, A. (Hg.), *Sign Languages' Nonmanuals.* Grazer Linguistische Studien (Special Issue), 93, 1–27.

Lackner, A. (Manuskript eingereicht zur Publikation): Semantic concepts associated with time-related signs in Austrian Sign Language (ÖGS). In: K. Richterová (Hg.), *Calendaric terms in sign languages.* Berlin: Mouton de Gruyter.

Lackner, A./Dotter, F./Stalzer, C./Mallinger, L./Pirker, A./Hausch, C./Unterberger, N./Riemer-Kankkonen, N./Dürr, X./Auersperg, B./Wiener, A./Koppendorfer, M./Graf, I./Recheis, C. (2017): *SignNonmanuals. Segmentierung und Strukturierung von Texten in Österreichischer Gebärdensprache (ÖGS).* Klagenfurt: Veröffentlichungen des Zentrums für Gebärdensprache und Hörbehindertenkommunikation. Band 24.

Lackner, A./Graf, I./Raffer, L./Scharfetter, E./Riemer-Kankkonen, N./Stalzer, C./Hausch, C./Unterberger, N./Bergmeister, E. (2019): *Austrian Sign Language (ÖGS) Corpus Annotation / Annotation des ÖGS-Korpus.* Klagenfurt: Veröffentlichungen des Zentrums für Gebärdensprache und Hörbehindertenkommunikation. Band 25.

Lehmann, C. (2005): Participant roles, thematic roles and syntactic functions. In: T. Tsunoda/T. Kageyama (Hg.), *Voice and Grammatical Relations: Festschrift for Masayoshi Shibatani.* (153-174). Amsterdam: J. Benjamins.

Lehmann, C./Maslova, E. (2004): Grammaticography. In: G. Booij/C. Lehmann/J. Mugdan/S. Skopeteas (Hg.), *HSK: Vol. 17,2. Morphologie. Ein Handbuch zur Flexion und Wortbildung*. Berlin: Mouton de Gruyter, 1857–1882.

Liddell, Scott K./Johnson, Robert E. (1989): American Sign Language: The Phonological Base. *Sign Language Studies*, 64, 195–278.

Lucas, C./Bayley, R./Valli, C. (2001): *Sociolinguistics variation in American Sign Language* (Vol. 7). Washington, D.C.: Gallaudet University Press.

Maas, U. (2004): "Finite" and "nonfinite" from a typological perspective. *Linguistics,* 42(2), 359–385.

Maas, U. (2011/12): *Sprachausbau. Skript zur Vorlesung. Graz WS 2011/12*. Manuskript, Graz.

Mallinger, L. (2012): *Grenzsignale in der Österreichischen Gebärdensprache*. Karl-Franzens-Universität Graz: Diplomarbeit.

Miller, C. (1994): Simultaneous Constructions in Quebec Sign Language. In: M. Brennan/G.H. (Hg.) *Word Order Issues in Sign Languages*. Durham: ISLA, 89–109.

Nicodemus, B. (2009): *Prosodic markers and utterance boundaries in American sign language interpretation*. (Studies in Interpretation 5). Washington: Gallaudet University Press.

Ormel, E. & Crasborn, O. (2012): Prosodic Correlates of Sentence in Signed Languages: A Literature Review and Suggestions for New Types of Studies. *Sign Language Studies*, 12(2), 109–145.

Papadatou-Pastou, M./Sáfár, A. (2016): Handedness prevalence in the deaf: Meta-analyses. *Neuroscience & Biobehavioral Reviews*, 60, 98–114. https://doi.org/10.1016/j.neubiorev.2015.11.013.

Perlmutter, D. M. (1992): Sonority and syllable structure in American Sign Language. *Linguistic Inquiry*, 23, 407–442).

Pike, Kenneth L. (1967[2]): *Language in relation to a unified theory of the structure of human behavior.* Den Haag: Mouton.

Riemer-Kankkonen, N. (2019): *Lexicon-based analysis of hand configurations*. Presentation at the SignNonmanuals Workshop 2, 3rd–4th May 2019, Graz.

Rössl, E. (1956): Geschichte der Landes-Taubstumen-Lehranstalt (1831–1956). In: Landes-Taubstummenanstalt in Graz (Hg.) *125 Jahre Taubstummenbildung in Steiermark.* Graz.

Šarač Kuhn, N./Schalber, K./Alibašić, T./Wilbur, R. B. (2007): Cross-linguistic comparison of interrogatives in Croatian, Austrian, and American Sign Languages. In: P. M. Perniss/R. Pfau/M. Steinbach (Hg.), *Visible variation: Comparative studies on sign language structure*. Berlin: Mouton de Gruyter, 207–244.

Sáfár, A./Crasborn, O. (2013): A corpus-based approach to manual simultaneity. In: L. Meurant/A. Sinte/M. Van Herreweghe/M. Vermeerbergen (Hg.), *Sign Language Research, Uses and Practices: Crossing views on theoretical and applied sign language linguistics*. Berlin: Mouton de Gruyter, 179–203.

Sandler, W. (2008): The syllable in sign language: Considering the other natural modality. B. Davis/K. Zajdo (Hg.), *The syllable in speech production*. *Perspectives on the Frame Content Theory*. New York: Taylor Francis, 379–408.

Sandler, W. (2011) Prosody and syntax in sign language. *Transactions of the Philological Society*, 108(3), 298–328. https://doi.org/10.1111/j.1467- 968X.2010.01242.x

Schalber, K. (2006): What is the chin doing? An analysis of interrogatives in Austrian Sign Language. *Sign Language and Linguistics*, 9(1/2), 133–150.

Schalber, K. (2015): Austrian Sign Language. In: J. Bakken Jepsen/G. De Clerck/S. Lutalo-Kiingi/W. B. McGregor (Hg.), *Sign Languages of the World. A Comparative Handbook.* Berlin: De Gruyter Mouton, 105–128.

Schott, W. (1995): *Das K.K. Taubstummen-Institut in Wien 1779–1918: Dargestellt nach historischen Überlieferungen und Dokumenten mit einem Abriß der wichtigsten pädagogischen Strömungen aus der Geschichte der Gehörlosenbildung bis zum Ende der Habsburgermonarchie.* Wien: Böhlau.

Schott, W. (1999): *Das Allgemeine Österreichische Israelitische Taubstummen-Institut in Wien 1844-1926. Dargestellt nach historischen Überlieferungen und Dokumenten mit einer Einleitung über die Entwicklungsgeschichte der Gehörlosenbildung.* Wien: Eigenverlag Walter Schott.

Selting, M. (2005): Syntax and prosody as methods for the construction and identification of turn-constructional units in conversation. In: A. Hakulinen/Selting, M. (Hg.), *Syntax and Lexis in Conversation.* Amsterdam: Benjamins, 17–44.

Siyavoshi, S. (2017): The Role of the Non-dominant Hand in ZEI Discourse Structure. *Sign Language Studies* 18(1), 58–72.

Stalzer, C. (2014): *Negation in der Österreichischen Gebärdensprache (ÖGS)*. Karl-Franzens-Universität Graz: Diplomarbeit.

Stokoe, W. C. (2005/1960): Sign Language Structure: An Outline of the Visual Communication System of the American Deaf. *Journal of Deaf Studies and Deaf Education*, 10(1), 3–37. [Auch in: http://jdsde.oxfordjournals.org/content/10/1/3.full.pdf+html [25.06.2019].

Venus, A. (1854): *Das kaiserl. königl. Taubstummen-Institut in Wien, seit seiner Gründung bis zum gegenwärtigen Zeitpunkte: nebst einer einleitenden Geschichte des Taubstummen-Unterrichtes und einer kurzen historisch-statistischen Darstellung der in dem österreichischen Kaiserstaate bestehenden Taubstummen-Anstalten; mit dem Grundrisse des Gebäudes*. Wien: Braumüller. [Auch in: https://reader.digitale-sammlungen.de/resolve/display/bsb10761366.html [25.06.2019]

Vermeerbergen, M./Leeson, L./Crasborn, O. (2007): *Simultaneity in signed languages: Form and function*. Amsterdam: John Benjamin.

Wasserstein, B. (2012): *On the eve: Jews in Europe before the second world war*. New York: Simon & Schuster Paperbacks.

Wilbur, R. B. (2000): Phonological and prosodic layering of non-manuals in American Sign Language. In: K. Emmorey/H.L. Lane (Hg.), *The signs of language revisited: An anthology to honor Ursula Bellugi and Edward Klima*. Hillsdale, NJ: Lawrence Erlbaum, 190–214.

Wilcox, S./Lackner, A. (2021): Language is an "activity of the whole body": A memorial to Franz Dotter. In: A. Lackner (Hg.) *Sign Languages' Nonmanuals*. Grazer Linguistische Studien. Special Issue, 93, 225–259.

Autor*innen und Herausgeber*innen

Gavin Brookes ist UKRI Future Leader Fellow an der Abteilung für Linguistik und Englisch an der Universität Lancaster, UK. Seine Forschungsschwerpunkte sind Korpuslinguistik, Diskursforschung, Multimodalität und Gesundheitskommunikation. Zu seinen Werken in diesem Bereich gehören *Analysing Health Communication: Discourse Approaches* (mit D. Hunt), *Obesity in the News: Language and Representation in the Press* (mit P. Baker) und *Corpus, Discourse and Mental Health* (mit D. Hunt). Brookes ist Mitherausgeber der Buchreihe *Corpus and Discourse* (Bloomsbury) und des *International Journal of Corpus Linguistics* (John Benjamins).

Ausgewählte Publikationen:

Brookes, G. (2020): Corpus linguistics in illness and healthcare contexts: a case study of diabulimia support groups. In: Demjén, Z. (Hg.), *Applying linguistics in illness and healthcare contexts.* Contemporary Studies in Linguistics, Bloomsbury, 44–72. https://doi.org/10.5040/9781350057685.0009

Brookes, G./Wright, D. (2020): From burden to threat: A diachronic study of language ideology and migrant representation in the British press. In: Rautionaho, P./Nurmi, A./Klemola, J. (Hg.), *Corpora and the Changing Society: Studies in the Evolution of English.* Studies in Corpus Linguistics, Vol. 96, John Benjamins, 113–140. https://doi.org/10.1075/scl.96.05bro

Brookes, G./Putland, E./Harvey, K. (2021): Multimodality: Examining Visual Representations of Dementia in Public Health Discourse. In: Brookes, G./Hunt, D. (Hg.), *Analysing Health Communication: Discourse Approaches.* Palgrave Macmillan, 241–269.

Federico Collaoni ist Lehrbeauftragter an der Università degli Studi di Udine/Italien im Bereich der Germanistischen Linguistik im Bachelorstudium, bis 2019 leitete er auch das Seminar *Übersetzung – Deutsch I Masterstudium: Einführung in die Fachübersetzung*. 2015 erlangte er die Lehrberechtigung für das Fach *Deutsch als Fremdsprache* in der Sekundarstufe des italienischen Schulsystems, und seit 2019 ist er Stammlehrer am Sprachgymnasium „E. L. Martin" (Latisana, Udine). Nach seiner Promotion an der Alpen-Adria-Universität Klagenfurt nahm er an verschiedenen Tagungen in Deutschland, Österreich, Polen, Dänemark, Rumänien und Ungarn aktiv teil. Zu seinen Forschungsinteressen zählen Medien- und Diskurslinguistik, Kontaktlinguistik und Mehrsprachigkeit sowie Sprach- und Literaturdidaktik. Seit 2017 ist er Co-Autor des am Dipartimento di Lingue e Letterature, Comunicazione, Formazione e Società der Universität Udine angesiedelten Projekts PRAGER (*Pragmatica del discorso sociale sulle energie rinnovabili*).

Ausgewählte Publikationen:

Collaoni, F. (2015): Deutsch als Wissenschaftssprache: Terminologieentwicklungen im Bereich ‚Erneuerbare Energien'. In: Szurawitzki, M./Busch-Lauer, I./Rössler, P./Krapp, R. (Hg.), *Wissenschaftssprache Deutsch: international, interdisziplinär, interkulturell*. Tübingen: Narr Francke Attempto, 153–162.

Collaoni, F. (2017): Grenzen der Sprachen und Grenzen der Sprachwissenschaft in der Ökolinguistik. In: Bartoszewicz, I./Szczęk, J./Tworek, A. (Hg.), *Grenzen der Sprache – Grenzen der Sprachwissenschaft I*, Linguistische Treffen in Wrocław 13, Wrocław/Dresden: Neisse Verlag, 43–54.

Collaoni, F. (2019): Angewandte Germanistische Linguistik: Ein interdisziplinärer Ansatz am Beispiel des Themas Energiewende. In: Philipp, H./Weber, B./Wellner, J. (Hg.), *Kosovarisch-rumänische Begegnung: Beiträge zur deutschen Sprache in und aus Südeuropa*, »Forschungen zur deutschen Sprache in Mittel-, Ost- und Südeuropa« 8, Open Access Schriftenreihe der Universitätsbibliothek Regensburg, 118–130.

Collaoni, F. (2020): Linguistik und Literatur: Ein interdisziplinärer Ansatz zur Analyse des Themas „Emotionen" am Beispiel Wolf Biermanns „Ermutigung". In: Bartoszewicz, I./Szczęk, J./Tworek, A. (Hg.), Linguisti-

sche Treffen in Wrocław 18 (Bd. II), Wrocław/Dresden: Neisse Verlag, 61–69.

Collaoni, F. (in Vorb.): Kap. 1.1.2 Lexikographische Fragestellungen und Fachwort, Kap. 2.3.3 Merkmale des Subkorpus „Collaoni“, Kap. 3.3. Diskursorganisierende Rolle des Sach- und Fachwortschatzes. In: Jammernegg, I./Kuri, S. (in Vorb.): *Diskurssteuerung und Wissensmanagement: Eine kontrastive linguistische Analyse zum Energiewendediskurs in Deutschland und in Italien*. Wien/Zürich: LIT.

Bernhard Glodny ist an der Universitätsklinik für Radiologie der Medizinischen Universität Innsbruck tätig. Sein klinisches Hauptarbeitsgebiet ist die Interventionelle Radiologie. Dementsprechend liegen seine wissenschaftlichen Schwerpunkte in der kardiovaskulären Grundlagenforschung, in der kardiovaskulären Bildgebung und in der interventionellen Radiologie. Er ist außerdem der Hauptkoordinator des Clinical-PhD-Programms an der Medizinischen Universität Innsbruck und Koordinator des Programms *Clinical Imaging Science*. Glodny arbeitete am Universitätsklinikum der Westfälischen Wilhelms Universität in Münster und an der Chemischen Fakultät an der Aufreinigung von blutdrucksenkenden Substanzen aus Nierenmark und antiproliferativen Substanzen aus Juglans Regia, der gemeinen Walnuss. Glodny und seine KollegInnen konnten erstmals das Vorhandensein von Pregnanen in dieser Gefäßpflanze nachweisen. Einige seiner wichtigsten Beiträge im Bereich der kardiovaskulären und interventionellen Radiologie beziehen sich auf neue Techniken zur Vermeidung von Luftembolien bei der Biopsie der Lunge. Durch sein gendermedizinisches Interesse wurde das interdisziplinäre Projekt *MedCorpInn* angestoßen, welches im vorliegenden Band beschrieben wird.

Ausgewählte Publikationen:

Glodny, B./Pauli, G. F. (2006): The vasodepressor function of the kidney: prostaglandin E2 is not the principal vasodepressor lipid of the renal medulla. *Acta Physiol (Oxf)*, 187(3): 419–30. https://doi.org/10.1111/j.1748-1716.2006.01578.x

Glodny, B./Pauli, G. F. (2005): Medullopressin: A New Pressor Activi-

ty from the Renal Medulla. *Hypertension Research* 28, 827–836. https://doi.org/10.1291/hypres.28.827

Pauli, G. F./Friesen, J. B./Gödecke, T./Farnsworth, N. R./Glodny, B. (2010): Occurrence of progesterone and related animal steroids in two higher plants. *Journal of Natural Products*, 73(3): 338–45. https://doi.org/10.1021/np9007415

Freund, M. C./Petersen, J./Goder K. C. et al. (2012): Systemic air embolism during percutaneous core needle biopsy of the lung: frequency and risk factors. *BMC Pulmonary Med*icine 12(2). https://doi.org/10.1186/1471-2466-12-2

Glodny, B./Schönherr, E./Freund, M. C./Haslauer, M./Petersen, J./Loizides, A./Grams, A. E./Augustin, F./Wiedermann, F. J./Rehwald, R. (2017): Measures to Prevent Air Embolism in Transthoracic Biopsy of the Lung. *American Journal of Roentgenology,* 208(5): W184–W191.

Leonhard Gruber ist Oberarzt an der Abteilung für Radiologie an der Medizinischen Universität Innsbruck. Seine Forschungsschwerpunkte liegen in den Bereichen Gender und muskuloskelettale Bildgebung, sowie Radiomics und neuartiger statistischer Ansätze für große Datensätze. Neben Vorträgen auf internationalen Kongressen umfassen seine wissenschaftlichen Publikationen ein breites Feld innerhalb der Radiologie, z.B. Projekte zu Gender-Imaging, zur Bildgebung von peripheren Neuropathien sowie zu Weichteiltumoren. Gruber ist außerdem PI des interdisziplinären ÖAW-geförderten und in diesem Band beschriebenen Projekts *MedCorpInn*.

Ausgewählte Publikationen:

Gruber, L./Gruber, H./Luger, A. K./Glodny, B./Henninger, B./Loizides, A. (2017): Diagnostic hierarchy of radiological features in soft tissue tumours and proposition of a simple diagnostic algorithm to estimate malignant potential of an unknown mass. *European Journal of Radiology*, 95: 102–110. https://doi.org/10.1016/j.ejrad.2017.07.020

Gruber, L./Loizides, A./Löscher, W./Glodny, B./Gruber, H. (2017): Focused high-resolution sonography of the suprascapular nerve: A simple surro-

gate marker for neuralgic amyotrophy? *Clinical Neurophysiology*, 128(8): 1438–1444. https://doi.org/10.1016/j.clinph.2017.04.030

Gruber, L./Jiménez-Franco, L. D./Decristoforo, C./Uprimny, C./Glatting, G./Hohenberger, P./Schoenberg, S. O./Reindl, W./Orlandi, F./Mariani, M./Jaschke, W./Virgolini, I. J. (2020): MITIGATE-NeoBOMB1, a Phase I/IIa Study to Evaluate Safety, Pharmacokinetics, and Preliminary Imaging of 68Ga-NeoBOMB1, a Gastrin-Releasing Peptide Receptor Antagonist, in GIST Patients. *Journal of Nuclear Medicine*, 61(12): 1749–1755. https://doi.org/10.2967/jnumed.119.238808

Gruber L./Loizides. A/Peer S./Walchhofer L. M./Spiss V./Brenner E./Stahl K./Gruber H. (2020): Ultrasonography of the Peripheral Nerves of the Forearm, Wrist and Hand: Definition of Landmarks, Anatomical Correlation and Clinical Implications. *RoFo*, 192(11): 1060–1072.

Elisabeth Gruber-Tokić studierte Allgemeine und Angewandte Sprachwissenschaft (Diplomstudium) und Sprach- und Medienwissenschaften (Doktoratsstudium) an der Universität Innsbruck und der Universidad de Oviedo, Spanien. Während des Studiums arbeitete sie als wissenschaftliche Mitarbeiterin in verschiedenen linguistischen Forschungsprojekten. 2016 promovierte sie mit einem DOC-Stipendium der Österreichischen Akademien der Wissenschaften zu den onymischen Umfeldern ausgewählter Tiroler Bergbauareale. Aktuell ist sie Projektleiterin des Innsbrucker Forschungsprojektes *Text Mining Medieval Mining Texts* (2019-2022). Zu ihren zentralen Forschungsgebieten zählen Onomastik, Korpuslinguistik, Paläographie und Wissenschaftskommunikation. Seit 2016 ist Gruber-Tokić Mitglied der Tiroler Nomenklaturkommission sowie der Arbeitsgemeinschaft für Kartographische Ortsnamenkunde. Außerdem unterrichtet sie seit 2016 Deutsch als Fremdsprache am BFI Tirol.

Ausgewählte Publikationen:

Gruber-Tokić, E./Adamski, I. (2019): Für die politische Rede typische rhetorische Figuren. In: Burkhardt, Armin (Hg.), *Handbuch der politischen Rhetorik*. Band 10, Berlin/Boston: De Gruyter, 583–602.

Gruber, E. (2017): Ausgewählte Grubennamen der Hs. Dipauliana 1164. In: Bichlmeier, H./Pohl, H. (Hg.), *Akten des XXX. Namenkundlichen Symposiums in Kals am Großglockner*. Hamburg: Baar-Verlag, 135–158.

Gruber, E. (2016): Grubennamen des Bergbauareales Silberberg im Verleihbuch der Rattenberger Bergrichter (1460-1463). In: Anreiter, P./Rampl, G. (Hg.), *8. Tagung des Arbeitskreises für Bayerisch Österreichische Namenforschung vom 25. bis 27. September 2014 in Innsbruck*. Wien: Praesens, 31–54.

Gruber, E. (2013): *Die Namen von Ebbs (Tirol)*. Innsbrucker Beiträge zur Onomastik. Wien: Praesens Verlag.

Gerald Hiebel ist seit 2020 Senior Scientist am Institut für Archäologien und am Digital Science Center an der Universität Innsbruck. Nach dem Geographiestudium an der Universität Wien war er von 1998-2007 in der Privatwirtschaft bei verschiedenen Telekommunikationsunternehmen tätig. Ab 2007 war er Mitarbeiter am Arbeitsbereich Vermessung der Universität Innsbruck und promovierte 2012. Seine Hauptforschungsgebiete sind Geoinformation, Ontologien und semantische Technologien. Mehrjährige Forschungsaufenthalte am ICS-FORTH in Heraklion und an der University of Southern California (USC) in Los Angeles vertieften seine Kenntnisse in diesen Bereichen. Hiebel leitete Projekte des Österreichischen Wissenschaftsfonds (FWF) und der Österreichischen Akademie der Wissenschaften (ÖAW) zu sprachwissenschaftlichen und archäologischen Fragestellungen.

Ausgewählte Publikationen:

Hiebel, G./Aspöck E./Kopetzky K. (2021): Ontological Modeling for Excavation Documentation and Virtual Reconstruction of an Ancient Egyptian Site. *ACM Journal on Computing and Cultural Heritage*, 14(3), Article 32, 1–14. https://doi.org/10.1145/3439735

Rampl, G./Gruber-Tokić, E./Posch, C./Hiebel, G. (2021): Toponomastik und Korpuslinguistik. Bergnamen im (Kon-)Text. In: Dräger, K./ Heuser, R./Prinz, M.: *Toponyme*. Berlin, Boston: De Gruyter, 225–248. https://doi.org/10.1515/9783110721140-012

Hiebel, G./Doerr, M./Eide, Ø. (2017) CRMgeo: A Spatiotemporal Extension of CIDOC-CRM. Together with. *International Journal on Digital Libraries Special Issue* 18, 271–279. https://doi.org/10.1007/s00799-016-0192-4

Anna Lena Huber ist Nachwuchswissenschaftlerin an der Universitätsklinik für Augenheilkunde und Optometrie der Medizinischen Universität Innsbruck. Ihr Dissertationsprojekt befasst sich mit neuen Therapien für neoproliferative Netzhauterkrankungen. Als Projektmitarbeiterin des in diesem Band beschriebenen, interdisziplinären Projekts *MedCorpInn* hat sie sich insbesondere mit gendermedizinischen Fragestellungen beschäftigt und u.a. Studien mit radiologischen medizinischen Texten erfolgreich in den Kontext der Augenheilkunde überführt. Mehrere Publikationen dazu sind in Vorbereitung.

Karoline Irschara ist Universitätsassistentin am Institut für Sprachwissenschaft der Universität Innsbruck und war in ihrer Studienzeit in mehreren dort angesiedelten linguistischen Forschungsprojekten tätig. In ihrem laufenden Dissertationsprojekt beschäftigt sie sich mit einer intersektionalen korpuslinguistischen Analyse medizinischer Befunde. Ihre Forschungsinteressen liegen in der Korpus- und Genderlinguistik sowie im Bereich der medizinischen Kommunikation. Sie ist PI des ÖAW-geförderten Projekts *MedCorpInn*, welches sich mit der Erstellung und Analyse eines umfangreichen Korpus aus medizinischen Befunden befasst.

Ausgewählte Publikationen:

Irschara, K./Huber, B./Ilić, S./Prossliner, L./Kienpointner, M. (2015): Dependenzanalyse interlingual – Zur Beschreibung der Struktur von komplexen Sätzen mithilfe der Dependenzgrammatik. *Wiener Linguistische Gazette* 79, Institut für Sprachwissenschaft/Universität Wien, 1–36.

Irschara, K. (2016): Manipulatives Argumentieren in TV-Diskussionen. Ein Fallbeispiel aus der ZIB 2. *Verbal – Zeitschrift des Verbands für Angewandte Linguistik*, Jahrgang XVII, 1/2016, 8–25.

Iris Jammernegg betreut als wissenschaftliche Mitarbeiterin *(ricercatrice)* an der Università degli Studi di Udine/Italien den übersetzungs- und kommunikationswissenschaftlich ausgerichteten Fachbereich *Deutsche Sprachwissenschaft*. Ihre Forschungsschwerpunkte sind sprachlich-textuelle Diskurs- und Fachinformationsstrategien in der Unternehmenskommunikation sowie der politischen Öffentlichkeitsarbeit aus kontrastiv-translatorischer Perspektive (Länder-Varietäten innerhalb des deutschen Sprachraums bzw. im Vergleich zu Italien), Wissensvermittlung und Kompetenzaufbau sowie fachspezifische DaF-Didaktik. Sie koordinierte die evaluierende Testung der Lernumgebung für das nationale PRID-Projekt TransLab (*Laboratorio on-line per la didattica della traduzione specializzata verso l'italiano da ceco, russo e tedesco*, *www.translab-project.eu*, Laufzeit 2017-2019). Seit 2017 ist sie Co-Autorin des am Dipartimento di Lingue e Letterature, Comunicazione, Formazione e Società der Universität Udine angesiedelten Projekts PRAGER (*Pragmatica del discorso sociale sulle energie rinnovabili*).

Ausgewählte Publikationen:

Jammernegg, I. (2019): Introduzione metodologica alla parte sperimentale. In: Perissutti, A. M./Kuri, S. (Hg.), *TransLab. Un progetto didattico per la traduzione specializzata*, LAM – Lingue antiche e moderne Strumenti, 1/2019, 167–180.

Jammernegg, I. (2019): *Gesellschaft verstehen und vermitteln. Eine interkulturelle und crosslinguale Diskursanalyse*. Udine: Forum.

Jammernegg, I. (2020): Rationale und emotionale Synergien in deutschen Energiewende-Diskursstrategien: Das Beispiel RWE. In: Carobbio, G./Desoutter, C./Fragonara, A. (Hg.), *Macht, Ratio und Emotion: Diskurse im digitalen Zeitalter / Pouvoir, raison et émotion: les discours à l'ère du numérique*. Bern: Peter Lang, Linguistic Insights 275, 143–161.

Jammernegg, I. (2020): Quantitativ-qualitative Exploration von Diskursmerkmalen und crosslingualen Aspekten. In: Roni, R. (Hg.), *Mantua Humanistic Studies. Volume XII*, collana *Mantua Humanistic Studies*, Mantova: Universitas Studiorum, 229–257.

Jammernegg, I./Kuri, S. (in Vorb.): *Diskurssteuerung und Wissensmanagement: Eine kontrastive linguistische Analyse zum Energiewendediskurs in Deutschland und in Italien.* Wien/Zürich: LIT.

Sonja Kuri ist assoziierte Professorin für Germanistische Linguistik an der Università degli Studi di Udine/Italien. Ihre Forschungsinteressen gelten der Text- und Diskurslinguistik, der Anwendung text- und gesprächslinguistischer Fragestellungen auf die Analyse literarischer Texte und dem Sprachenlehren und -lernen in mehrsprachigen und multimedialen Kontexten. Sie ist Co-Autorin des EU-Projekts WRILAB2 (*On-line Reading and Writing Laboratory for Czech, German, Italian and Slovenian as L2, www.wrilab2.eu,* Laufzeit 2014-2016) und des nationalen PRID-Projekts TransLab (*Laboratorio online per la didattica della traduzione specializzata verso l'italiano da ceco, russo e tedesco*, *www.translab-project.eu*, Laufzeit 2017-2019). Seit 2017 ist sie Co-Autorin des am Dipartimento di Lingue e Letterature, Comunicazione, Formazione e Società der Universität Udine angesiedelten Projekts PRAGER (*Pragmatica del discorso sociale sulle energie rinnovabili*).

Ausgewählte Publikationen:

Perissutti, A. M./Kuri, S./Doleschal, U. (Hg.) (2016): *WRILAB2. A Didactical Approach to Develop Text Competences in L2*. Wien: LIT.

Perissutti, AM./Kuri, S. (Hg.) (2019): *TransLab. Un progetto didattico per la traduzione specializzata*, LAM – Lingue antiche e moderne Strumenti, 1/2019.

Kuri, S. (2019): *Die Rezension. Entwicklung, Bestimmung, Geltung*. Udine: Forum.

Kuri, S. (2019): "Sehr geehrter Herr Sektionschef!" Oder: Wie aus einem Appell (fast) eine Obligation wird. Der Brief in Franz Werfels Novelle Eine blaßblaue Frauenschrift. In: Polledri, E./Costagli, S. (Hg.), *La lettera nella letteratura tedesca ed europea*. Cultura tedesca 56, 209–228.

Jammernegg, I./Kuri, S. (in Vorb.): *Diskurssteuerung und Wissensmanagement: Eine kontrastive linguistische Analyse zum Energiewendediskurs in Deutschland und in Italien*. Wien/Zürich: LIT.

Andrea Lackner leitet seit 2017 das FWF-Projekt *The Interplay of Nonmanuals and Clauses in Austrian Sign Language (ÖGS) Texts* am Institut für Sprachwissenschaft der Universität Graz. Davor initiierte und führte sie die wissenschaftliche Forschung des FWF-Projekts *Segmentation and Structuring Austrian Sign Language Texts* (2011-2015) aus, das von Franz Dotter (†) geleitet wurde und am Zentrum für Gebärdensprache und Hörbehindertenkommunikation, Universität Klagenfurt, angesiedelt war. Sie schloss 2007 ihr Studium der Sprachwissenschaft ab und promovierte 2013 im selbigen Fach.

Ausgewählte Publikationen:

Lackner, A. (2017): *Functions of head and body movements in Austrian Sign Language*. Berlin, Boston: De Gruyter.

Lackner, A. (2019): Describing Nonmanuals in Sign Languages. *Grazer Linguistische Studien (gls)* 91, 45–103.

Lackner, A. (2020): *ÖGS-Korpus. Korpus der Österreichischen Gebärdensprache / Corpus ÖGS. Austrian Sign Language Corpus.* Access at The Language Archive, MPI Nijmegen, https://hdl.handle.net/1839/14ac9c60-59b2-4171-9869-0a49f51f4de5

Wilcox, S./Lackner, A. (2021): Language is an „activity of the whole body“: A memorial to Franz Dotter. *Grazer Linguistische Studien (gls)* 93, 230–264.

Rosemarie Lühr ist Honorarprofessorin an der Humboldt Universität zu Berlin, ihre Forschungsschwerpunkte sind Indogermanistik und Historische deutsche Sprachwissenschaft.

Sie studierte Latein, Sport, Indogermanistik und Kirchenmusik an der Universität Erlangen und promovierte dort 1977 im Bereich der Vergleichenden indogermanischen Sprachwissenschaft. 1984 erfolgte die Habilitation im Bereich der Deutschen Philologie (Sprachwissenschaft) an der Universität Regensburg. Von 1990 bis 1994 war sie als Universitätsprofessorin für Vergleichende Sprachwissenschaft an der Justus-Liebig Universität Gießen tätig, bis 2013 als Universitätsprofessorin für Indogermanistik an der Friedrich-Schiller-Universität Jena. Ihre Drittmittelprojekte beschäftigen sich u.a. mit der indogermanischen Konkurrenzsyntax, der historischen jiddischen Syntax,

der deutschen Wortfeldetymologie, dem Lexikonkonzept des Indogermanischen, etymologischen Wörterbüchern und weiteren mehr.

Ausgewählte Publikationen:

Lühr, R. (1982): *Studien zur Sprache des Hildebrandliedes*. Zugl. Erlangen-Nürnberg, Univ., Diss., 1977. Frankfurt am Main: Lang (Europäische Hochschulschriften Reihe 1, Deutsche Sprache und Literatur).

Lühr, R. (1988): Expressivität und Lautgesetz im Germanischen. *Monographien zur Sprachwissenschaft* 15, Heidelberg: Winter.

Lühr, R. (2000): Gedichte des Skalden Egill. *Jenaer indogermanistische Textbearbeitungen* 1, Dettelbach: J. H. Röll.

Stephanie Mangesius ist Ärztin und forscht an der Abteilung für Neuroradiologie und Radiologie sowie der Neuroimaging Core Facility an der Medizinischen Universität Innsbruck, und ist in mehreren nationalen und internationalen Kooperationen beteiligt. Ihre früheren Forschungsprojekte befassen sich mit der Entwicklung von Hirntumoren unter Verwendung von In-vitro- und In-vivo-Ansätzen. In den momentan laufenden Projekten beschäftigt sie sich mit fortschrittlichen Bildgebungstechniken zur Prognoseabschätzung und dem Therapieansprechen von Tumoren und Metastasen im Gehirn. Weitere Forschungsansätze konzentrierten sich auf die Untersuchung von Biomarkern für neurodegenerative Parkinson-Syndrome mit Hilfe von Labortechniken sowie klinischen und bildgebenden Methoden. Rezenter beschäftigt sie sich auch mit neuroinflammatorischen Erkrankungen, sowie auch der Erforschung seltener Krankheiten wie der Ataxie. Mangesius ist außerdem Teilprojektleiterin und PI im interdisziplinären Projekt *MedCorpInn*, das sich mit gendermedizinischen Fragestellungen zu radiologischen Befundtexten befasst.

Ausgewählte Publikationen:

Mangesius, S./Hussl, A./Krismer, F./Mahlknecht, P./Reiter, E./Tagwercher, S./Djamshidian, A./Schocke, M./Esterhammer, R./Wenning, G./Müller, C./Scherfler, C./Gizewski, E. R./Poewe, W./Seppi, K. (2018): MR planimetry in neurodegenerative parkinsonism yields high diagnostic

accuracy for PSP. *Parkinsonism & Related Disorders*, 46, 47–55. https://doi.org/10.1016/j.parkreldis.2017.10.020

Mangesius, S./Mariotto, S./Ferrari, S./Pereverzyev, S. Jr./Lerchner, H./Haider, L./Gizewski, E.R., Wenning, G./Seppi, K./Reindl, M./Poewe, W. (2020): Novel decision algorithm to discriminate parkinsonism with combined blood and imaging biomarkers. *Parkinsonism & Related Disorders*, 77: 57–63. https://doi.org/10.1016/j.parkreldis.2020.05.033

Mangesius S./Hussl, A./Tagwercher, S./Reiter, E./Müller, C./Lenhart, L./Krismer, F./ Mahlknecht, P./Schocke, M./Gizewski, E. R./Poewe, W./Seppi, K. (2020): No effect of age, gender and total intracranial volume on brainstem MR planimetric measurements. *European Radiology*; 30(5): 2802–2808. https://doi.org/10.1007/s00330-019-06504-1

Haider, L./Chan Wei-Shin, E./Olbert, E./Mangesius, S./Dal-Bianco, A./ Leutmezer, F./Prayer, D./Thurnher, M. (2019): Cranial Nerve Enhancement in Multiple Sclerosis Is Associated With Younger Age at Onset and More Severe Disease. *Frontiers in Neurology*, 10: 1085. https://doi.org/10.3389/fneur.2019.01085

Straub, S./Mangesius, S./Emmerich, J./, et al. (2020): Toward quantitative neuroimaging biomarkers for Friedreich's ataxia at 7 Tesla: Susceptibility mapping, diffusion imaging, R2 and R1 relaxometry. *Journal of Neuroscience Research*, 98: 2219–2231.

Tony McEnery ist außerordentlicher Professor für Englische Sprache und Linguistik an der Universität Lancaster und Inhaber des Changjiang-Lehrstuhls an der Xi'an Jiaotong Universität, China. Zuvor war Tony McEnery Forschungsdirektor und Interimsgeschäftsführer des britischen Economic and Social Research Council (ESRC) und Direktor des ESRC Centre for Corpus Approaches to Social Science (CASS) in Lancaster. Er hat zahlreiche Publikationen zur Korpuslinguistik veröffentlicht.

Ausgewählte Publikationen:

McEnery, A./Baker, H./Brezina, V. (2021): Slavery and Britain in the 19th century. In: A. Čermáková/Egan, T./Hasselgård, H./Rørvik, S. (Hg.), *Time*

in Languages, Languages in Time. Studies in Corpus Linguistics; 101, John Benjamins, 9–38. https://doi.org/10.1075/scl.101.02mce

McEnery, A./Baker, H./Dayrell, C. (2021): Analysing the impacts of 19th-century drought: A corpus-based study. In: Fuster-Márquez, M./Gregori-Signes, C./Santaemilia, J./ Rodríguez-Abruñeiras, P. (Hg.), *Exploring discourse and ideology through corpora.* Linguistic Insights; 276, Peter Lang Publishing Group. https://doi.org/10.3726/b17868

McEnery, A./Hardie, A. (2020): Neo-Firthian corpus linguistics. In: Joseph, J./Waugh, L./Monville-Burston, M. (Hg.), *The Cambridge History of Linguistics* Cambridge University Press.

Claudia Posch ist Assistenzprofessorin am Institut für Sprachwissenschaft der Universität Innsbruck. Ihre Forschungsschwerpunkte liegen in der Korpuslinguistik sowie der Feministischen Linguistik. Derzeit arbeitet sie an ihrer Habilitationsschrift, in welcher sie Fragestellungen aus der Diskurslinguistik mit Korpuslinguistik kombiniert. Sie leitet seit 2019 das ÖAW- und go!digital-geförderte Projekt *MedCorpInn,* in welchem ein großes Korpus medizinischer Befunde kompiliert wird. Zuvor leitete sie das ebenfalls im Bereich Korpuserstellung angesiedelte Projekt *Alpenwort* und war bei mehreren Nachfolgeprojekten federführend mit dabei. Unter anderem wurden in diesen Projekten verschiedensprachige Korpora von Zeitschriften der Alpenvereine erstellt (AV-Österreich, Neuseeland, Club Andino Bariloche) und semantische sowie Namenannotationen weiterentwickelt.

Ausgewählte Publikationen:

Posch, C./Rampl, G. (2020): Lima or cima? Structure recognition and OCR in building the corpus of the Austrian Alpine Club Journal. *International Journal of Corpus Linguistics* 25, 489–503. https://doi.org/10.1075/ijcl.19094.pos

Rampl, G./Gruber-Tokić, E./Posch, C./Hiebel, G. (2021): Toponomastik und Korpuslinguistik. Bergnamen im (Kon-)Text. In: Dräger, K./ Heuser, R./Prinz, M.: *Toponyme*. Berlin/Boston: De Gruyter, 225–248. https://doi.org/10.1515/9783110721140-012

Posch, C./Kegyes, E. (in press): Periphrastische Geschlechtsspezifizierung im Ungarischen und im Deutschen. In: Kegyes, E./Zipser, K. (Hg.) *Kontrastive Studien im Sprachenpaar Deutsch und Ungarisch*. Kovac: Hamburg.
Posch, C./Rampl, G. (2018): *Alpenwort – Corpus of the Almanac of the Austrian Alpine Club* (data set, Version 1.0.0).

Gerhard Rampl ist seit 2013 Senior Scientist am Institut für Sprachwissenschaft an der Universität Innsbruck. Nach dem Studium der Allgemeinen und Angewandten Sprachwissenschaft und Romanistik an der Universität Innsbruck war er von 2005-2013 Mitarbeiter am Institut für Österreichische Dialekt- und Namenlexika an der Österreichischen Akademie der Wissenschaften (ÖAW). Seine Hauptforschungsgebiete sind Onomastik und Korpuslinguistik, daneben beschäftigt er sich aber auch mit Interaktionslinguistik. Zu Letzterem führte er im akademischen Jahr 2011/12 das Forschungsprojekt *Communicating Location in Emergency Calls* an der University of California, Los Angeles (UCLA) durch. Ein Lehraufenthalt führte ihn im Herbstsemester 2016 an die University of New Orleans, wo er u. a. zum romanischen Erbe in den Ortsnamen von Louisiana unterrichtete. In jüngster Zeit baute er in Zusammenarbeit mit Claudia Posch das Korpus Alpenwort der Zeitschrift des Österreichischen Alpenvereins auf. In diesem wurden die Jahrbücher des Alpenvereins (jährlich erschienen von 1869/70 bis in die Gegenwart) digital erfasst und linguistisch annotiert, wobei die Erkennung und Auszeichnung von Toponymen einen Schwerpunkt bildete.

Ausgewählte Publikationen:

Rampl, G./Gruber-Tokić, E./Posch, C./Hiebel, G. (2021): Toponomastik und Korpuslinguistik. Bergnamen im (Kon-)Text. In: Dräger, K./Heuser, R./Prinz, M.: *Toponyme*. Berlin/Boston: De Gruyter, 225–248. https://doi.org/10.1515/9783110721140-012
Posch, C./Rampl, G. (2020): Lima or cima? Structure recognition and OCR in building the corpus of the Austrian Alpine Club Journal. *International Journal of Corpus Linguistics* 25, 489–503. https://doi.org/10.1075/ijcl.19094.pos

Rampl, G. (2014): „leitstelle tirol notruf-wo genau ist der einsatzort? Die Frage nach dem Einsatzort in der Eröffnung von Telefon-Notrufen.“ In: Schwarze, C./ Konzett, C. (Hg.): *Interaktionsforschung: Gesprächsanalytische Fallstudien und Forschungspraxis*. Berlin: Frank und Timme, 83–103.

Birgit Waldner studierte Technische Chemie an der Technischen Universität Wien und schloss ihr Studium 2010 mit Auszeichnung ab. Anschließend promovierte sie im Bereich der theoretischen Strukturbiologie und des virtuellen Wirkstoffdesigns an der Universität Innsbruck und absolvierte gleichzeitig ein Masterstudium der Material- und Nanowissenschaften. Nach Abschluss des Doktoratsstudiums 2018 arbeitete sie für kurze Zeit in der pharmazeutischen Industrie, bevor sie eine Karriere in der Medizin anstrebte. Seit 2019 ist sie Medizinstudentin und Postdoc im Bereich des Natural Language Processing an der Medizinischen Universität Innsbruck.

Ausgewählte Publikationen:

Waldner, B. J./Machalett, R./Schönbichler, S./Dittmer, M./Rubner, M. M./Intelmann, D. (2020): Fast evaluation of herbal substance class composition by relative mass defect plots. *Analytical Chemistry* 92(19), 12909-12916. https://doi.org/10.1021/acs.analchem.0c01447

Waldner, B. J./Kramml, J./Kahler, U./Spinn, A./Schauperl, M./Podewitz, M./Fuchs, J. E./Cruciani, G./Liedl, K. R. (2018): Electrostatic Recognition in Substrate Binding to Serine Proteases. *Journal of Molecular Recognition*, 31(10), e2727. https://doi.org/10.1002/jmr.2727

Schauperl, M./Podewitz, M./Waldner, B. J./Liedl, K. R. (2016): Enthalpic and Entropic Contributions to Hydrophobicity. *Journal of Chemical Theory and Computation*, 12(9), 4600–4610. https://doi.org/10.1021/acs.jctc.6b00422

Waldner, B. J./Fuchs, J. E./Schauperl, M./Kramer, C./Liedl, K. R. (2016): Protease Inhibitors in View of Peptide Substrate Databases. *Journal of Chemical Information and Modeling*, 56(6), 1228–1235. https://doi.org/10.1021/acs.jcim.6b00064

Waldner, B. J./Fuchs, J. E./Huber, R. G./von Grafenstein, S./Schauperl, M./Kramer, C./Liedl, K. R. (2015): Quantitative Correlation of Conformational Binding Enthalpy with Substrate Specificity of Serine Proteases. *The Journal of Physical Chemistry B*, 120(2), 299–308. https://doi.org/10.1021/acs.jpcb.5b10637

Index der Korpora

Index

Der Index wurde basierend auf einer Keyword-Analyse aller Beiträge erstellt und händisch ergänzt.

Korpora, korpuslinguistische Methoden und Instrumentarien der Korpuserstellung haben in den vergangenen Jahren einen beispiellosen Aufschwung innerhalb der Linguistik erlebt. Der vorliegende Band stellt sich die Frage inwieweit inzwischen von einer Digitalen Linguistik gesprochen kann. Neben der Erstellung und Zusammensetzung linguistischer Textkorpora thematisieren die Beiträge dieses Bandes unterschiedliche datengeleitete wie korpusgestützte Verfahren und reflektieren die methodische Vielfalt digitaler linguistischer Disziplinen.

ISBN 978-3-99106-061-1
9 783991 060611