IDS OPEN
ONLINE-ONLY PUBLIKATIONEN
DES LEIBNIZ-INSTITUTS FÜR DEUTSCHE SPRACHE

Deutsch im Wandel

Beiträge zur Methodenmesse der IDS-Jahrestagung 1

Herausgegeben von

Annelen Brunner
Sandra Hansen
Christian Lang
Ngoc Duyen Tanja Tu
Sascha Wolfer

IDS OPEN
Band 16 (2026)

Inhalt

https://doi.org/10.21248/idsopen.16.2026.66

Annelen Brunner / Sandra Hansen / Christian Lang / Ngoc Duyen Tanja Tu / Sascha Wolfer

Einleitung

Dieser Sammelband präsentiert die Beiträge zur Methodenmesse der 61. Jahrestagung des Leibniz-Instituts für Deutsche Sprache 2025, die unter dem Motto „Deutsch im Wandel" das Thema Sprachwandel – vor allem im neueren Deutsch, aber auch in seiner Ausprägung über mehrere Sprachstufen hinweg – in den Fokus nahm. Durch die zunehmende Digitalisierung auch historischer Ressourcen und die Entwicklung neuer technologischer Möglichkeiten zur Erhebung, Erfassung und Analyse von Daten eröffnen sich auch der Sprachwandelforschung neue Möglichkeiten. Ein besonderes Anliegen der Methodenmesse war es darum, neuartige methodische Zugänge (z. B. Werkzeuge und Ressourcen) vorzustellen, die die Erforschung von Sprachwandel unterstützen oder überhaupt erst ermöglichen.

Für die Methodenmesse gab es eine freie Ausschreibung. Aus den 27 Einreichungen wurden die vorliegenden 12 Beiträge ausgewählt und auf der Methodenmesse in Form von Postern und/oder Tech-Demos präsentiert. Kriterien für die Auswahl waren u. a. Aktualität und tatsächliche Verwendbarkeit der vorgestellten Methoden für die Sprachwandelforschung. Für diesen Sammelband wurden die Beiträge verschriftlicht, so dass sie neben Beschreibungen von Methoden, Ressourcen oder Werkzeugen auch konkrete Fallstudien enthalten. Damit bietet der Band neben einem aktuellen Überblick über Methoden der Sprachwandelforschung auch Anregungen und Unterstützung für die eigene Forschung.

Felicitas Kleber, **Christoph Draxler**, **Jürgen Trouvain** und **Sven Grawunder** stellen ein halbautomatisch bearbeitetes Radionachrichten-Korpus als nützliche Quelle für diachrone Trendstudien vor. Auf Basis des Korpus nehmen die Autor*innen eine erste explorative Bewertung von Wandel im Bereich Sprechgeschwindigkeit vor. Außerdem zeigen sie, dass während des untersuchten Zeitraums von 60 Jahren der Anteil sogenannter Kernnachrichten zugunsten anderer Nachrichtenelemente reduziert wurde.

Charlotte Rein und **Timo Schürmann** präsentieren die Smartphone-App „PALAVA" und zeigen, wie man mit dieser Daten zum regionalen Sprachgebrauch in Nordrhein-Westfalen sammeln und auswerten kann. Eine Beispielanalyse verdeutlicht, dass die Daten, die mit der App gesammelt werden, nicht nur synchrone Betrachtungen der sprachlichen Variation in den Regionalsprachen ermöglichen, sondern auch die Beobachtung des aktuellen Sprachwandels mit Hilfe von *apparent time*-Analysen.

Das Tool „Annimate", das zum Datenexport aus historischen Referenzkorpora genutzt werden kann, steht im Beitrag von **Matthias Stemmler** und **Sonja Zeman** im Mittelpunkt. Im Rahmen einer Fallstudie zu den Interjektionen *ei*/*ey* demonstrieren sie die Funktionen des Tools und dessen Potenziale bei der Untersuchung von Sprachwandelphänomenen.

In ihrem Beitrag stellen **Laura Duve** und **Valeria Schick** einen Workflow zur softwaregestützten Erstellung, Aufbereitung und Auswertung historischer Korpora vor. Dieser wurde im Kontext der DFG-geförderten Forschungsgruppe „Praktiken der Personenreferenz" entwickelt. Als Beispiel dient dabei eine Untersuchung zur Verwendung der Heische-Formel mit *man* (z. B. *man nehme X*) in unterschiedlichen Gattungen des 15. bis 17. Jahrhunderts.

Thomas Burch, **Jost Gippert**, **Sarah Ihden**, **Sarah Kwekkeboom**, **Ralf Plate**, **Lena Schnee**, **Ingrid Schröder**, **Roland Schuhmann** und **Johanna Wittmack** beschreiben

den Aufbau und die Nutzungsmöglichkeiten einer digitalen Forschungsumgebung zur historischen Wortbildung des Deutschen. In dieser Forschungsumgebung, die im Rahmen des DFG-Langzeitprojekts „Wortfamilien diachron“ entwickelt wird, werden Referenzwortschätze des Althochdeutschen, Altsächsischen, Mittelhochdeutschen und Mittelniederdeutschen miteinander verknüpft und Suchen nach morphologischen Strukturmustern ermöglicht. Auf dieser Grundlage können Themen der historischen Wortbildungsforschung, z. B. Suffixablösung bei Nomina agentis, untersucht werden.

Mit der Webapplikation „Reddit Corpus Keyword Search“ (ReCKS) kann man ein großes Korpus von Kommentaren der Onlineplattform Reddit (Subreddit r/de, 2006–2023, ca. 41 Mio. Tokens) durchsuchen. **Jenia Yudytska** und **Jannis Androutsopoulos** zeigen, wie Suchergebnisse zur Weiterverarbeitung in Tabellenform exportiert oder mit der App selbst visualisiert werden können. In ihrem Beitrag gehen sie zudem auf die technische Architektur ihrer Anwendung ein und stellen eine beispielhafte Studie zu Genderzeichen vor.

Peter Meyer und **Doris Stolberg** skizzieren eine zurzeit im Aufbau befindliche Online-Ressource, mit der u. a. Unterschiede zwischen allgemeinen Sprachwandeltendenzen in deutschbasierten Kontaktvarietäten herausgestellt werden können. Prinzipiell können mit dieser webbasierten Forschungs- und Analyseplattform mündliche Korpora von Diaspora-Varietäten des Deutschen vergleichend im Hinblick auf lexikalische Kontaktphänomene untersucht werden.

Das Projekt „Deutsche Wortfeldetymologie in europäischem Kontext“ (DWEE) wird von **Thomas Eckart**, **Ulrike Ertel** und **Christine Rains** vorgestellt. Sie gehen insbesondere auf die damit verbundene Datenbank ein und zeigen, welche Zugänge den Nutzer*innen zur Verfügung stehen. Als Beispiel präsentieren sie eine Untersuchung zum Bedeutungswandel des Begriffs „Bildung“ sowie des damit verbundenen Wortfelds.

Susanne Kabatnik und **Anne Klee** beleuchten pandemiebedingten lexikalischen Sprachwandel am Beispiel der Cholera anhand einer Korpusstudie. Zudem wird die lexikografische Aufbereitung pandemierelevanter Begriffe als digitales Wörterbuch sowie der Aufbau einer *Linked-Open-Data*-Datenbank zu historischen Pandemien vorgestellt.

Der Beitrag von **Daniel Knuchel**, **Xenia Bojarski**, **Davide Ventre**, **Sonja Huber** und **Gunilla Kaibel** stellt Methoden vor, Diskurswandel mit Hilfe von quantitativen Methoden auf der Grundlage von Korpora zu visualisieren und zu analysieren. Dies geschieht anhand eines von den Autor*innen entwickelten und frei verfügbaren Toolkits, das es erlaubt, solche Analysen ohne große technische Expertise durchzuführen, und wird mit mehreren Anwendungsbeispielen illustriert.

Der Erstellungsprozess des Editionsprojektes des Darmstädter Tagblatts (1740 bis 1986) steht im Beitrag von **Jörn Stegmeier**, **Kevin Kuck**, **Anna Christina Kupffer**, **Anne Christine Günther**, **Angela Hammer**, **Dario Kampkaspar**, **Marcus Müller** und **Thomas Stäcker** im Fokus. Sie gehen auf technische und rechtliche Aspekte bei der Erstellung der Ressource ein und präsentieren anhand einer Fallstudie, wie diese für diachrone Untersuchungen genutzt werden kann.

Der Beitrag von **Dennis Beitel**, **Lisa Dücker**, **Salome Lipfert** und **Georg Oberdorfer** zeigt, wie die Plattform „REDE SprachGIS“ genutzt werden kann, um regionalen Sprachwandel vom späten 19. bis ins frühe 21. Jahrhundert mithilfe verschiedener Tonkorpora und schriftlicher Dialekterhebungen zu untersuchen. Unter Einbezug der Daten aus der REDE-Neuerhebung wird ein ausgewähltes Phänomen zusätzlich in drei Generationen in *apparent time* analysiert und mit Selbsteinschätzungsdaten zur Dialektkompetenz aus der

REDE-Infothek verglichen. Die Autor*innen verdeutlichen so das Potenzial einer multiperspektivischen Analyse durch die Kombination unterschiedlicher Datensätze.

Wir danken den Autor*innen für die innovativen Einreichungen und interessanten Präsentationen. Den Leser*innen wünschen wir viel Spaß bei der Lektüre der ausgearbeiteten Beiträge.

Kontaktinformation

Dr. Annelen Brunner
Leibniz-Institut für Deutsche Sprache (IDS)
R5, 6-13
68161 Mannheim
E-Mail: brunner@ids-mannheim.de

Dr. Sandra Hansen
Leibniz-Institut für Deutsche Sprache (IDS)
E-Mail: hansen@ids-mannheim.de

Dr. Christian Lang
Leibniz-Institut für Deutsche Sprache (IDS)
E-Mail: lang@ids-mannheim.de

Dr. Ngoc Duyen Tanja Tu
Leibniz-Institut für Deutsche Sprache (IDS)
E-Mail: tu@ids-mannheim.de

Dr. Sascha Wolfer
Leibniz-Institut für Deutsche Sprache (IDS)
E-Mail: wolfer@ids-mannheim.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.67

Felicitas Kleber/Christoph Draxler/Jürgen Trouvain/Sven Grawunder

Rundfunknachrichten als linguistisch-phonetische Ressource für die Sprachwandelforschung

Ein halbautomatisiertes Verfahren und eine Demosprachdatenbank

Abstract Spoken language is subject to synchronous variation at the segmental and suprasegmental levels, which in turn can lead to diachronic change. This study presents a semi-automatically processed radio news corpus as a useful source for diachronic trend studies and provides an initial exploratory assessment of prosodic change in terms of speech rate. The corpus comprises radio news from 1956 to 2017, obtained from the public service broadcaster *Saarländischer Rundfunk*. An initial analysis revealed that during the 60-year period under investigation, the proportion of core news was reduced in favor of other news elements. Further signal-based and signal+text-based analyses of the core news items showed that speech rate increased over time, which was accompanied by fewer and shorter pauses. The articulation speed, on the other hand, did not change. Advantages and disadvantages of news corpora for prosodic change research are discussed.

Keywords radio news, prosodic change, speech and articulation rate, pauses

1. Einleitung

Sprachwandel findet auf allen linguistischen Ebenen geschriebener und gesprochener Sprache statt. Der vorliegende Beitrag thematisiert neue Untersuchungsmöglichkeiten im Bereich von Wandel gesprochener Sprache. Im Vergleich zu dessen schriftsprachlich basierter Rekonstruktion lassen sich rezente und aktuell stattfindende Wandelprozesse auf segmentaler (Einzellaute) aber auch suprasegmentaler (Prosodie) Ebene experimentalphonetisch erst seit dem Aufkommen von Tonträgern und dem Vorhandensein einer entsprechenden Datengrundlage untersuchen. Diese Datenlage sowie die technologischen Möglichkeiten ihrer Analyse haben sich in den vergangenen Jahrzehnten rapide verbessert. Das *ONZE Corpus* (Gordon/Maclagan/Hay 2007) oder das *Philadelphia Neighborhood Corpus* (Labov/Rosenfelder/Fruehwald 2013) sind Beispiele soziophonetischer Korpora, anhand derer sowohl Ursprung als auch Verbreitung komplexer Lautwandelprozesse (z. B. vokalische Kettenverschiebungen, diachrone Vokalnasalierung) signalphonetisch untersucht werden können (vgl. Kleber 2025). Nach nunmehr 100 Jahren vertonter deutscher Nachrichtensprache (Dussel 2022) bietet sich dieses Medium aufgrund von Verfügbarkeit (Rundfunkarchive) und Vergleichbarkeit (Lesesprache, relative Homogenität der Sprechenden) perspektivisch nicht nur für sprachtechnologische Untersuchungen (z. B. Hecht/Riedler/Backfried 2002) an, sondern auch für Sprachwandelstudien zum Deutschen (Bose/Grawunder/Schwarz 2018), z. B. in Form von *real-time* Trendstudien, in denen ältere und jüngere Daten verschiedener Personen vergleichbarer Altersgruppen verglichen werden (Bülow/Scheutz/Wallner 2019).

Die Grundlage für Studien dieser Art bieten Projekte wie die ARD-Nachrichtenarche (Grawunder 2011). In diesem Beitrag werden Vor- und Nachverarbeitung von Radionachrichten mittels der BAS WebServices (Kisler/Reichel/Schiel 2017) u. a. zur Analyse im EMU-SDMS-Format (Winkelmann/Harrington/Jänsch 2017) anhand einer Sprachdatenbank mit derzeit 49 Nachrichtensendungen vom Saarländischen Rundfunk aufgezeigt.

1.1 Nachrichtensprache

Seit etwa 100 Jahren – mit Aufkommen des Rundfunks als Massenmedium – werden Nachrichten vorgelesen. Sie stellen damit im Vergleich zu anderen Gattungen gesprochener Sprache wie beispielsweise Gedichtrezitationen, Predigten oder öffentlichen Reden ein relativ junges Genre dar.

Typische Merkmale von Radionachrichten sind die Monologform vorgelesener und vorbereiteter Sprache, eine recht komplexe Syntax mit korrekter Grammatik und das Fehlen bestimmter Satzmuster wie Fragen (vgl. Iivonen/Niemi/Paananen 1995). Neben diesen kompositorischen Differenzen unterscheidet sich Nachrichtensprache von anderen Formen gesprochener Sprache zudem in der Sprachproduktion. Normalerweise artikulieren Nachrichtensprechende sehr deutlich und verwenden einen „neutralen“ Tonfall. Ersteres dient einer besseren Verständlichkeit; letzteres soll eine neutrale Haltung der Sprechenden gegenüber den Inhalten zum Ausdruck bringen. Nachrichtensprache weist somit stilistische Merkmale auf, die in Kontrast zur spontanen Sprache in Gesprächen stehen.

Code	Beschreibung
COR	Kernnachricht, Meldung, klassische Nachricht
INT	Anmoderation/Einleitung
LOC	Ortsangabe oder Thema
OTO	O-Ton (Bericht oder Interviewausschnitt; oft mit Hintergrundgeräusch)
OUT	Abmoderation/Outro
PAC	Jingle/Senderkennung, akustische Trenner, Pausen- bzw. Zeitzeichen, akustische Verpackung
REP	Bericht (oft vorproduzierter Einspieler)
SPO	Sport
STO	Börsenmeldungen
TIM	Zeitansage
TOP	Themenübersicht
TRA	Verkehrsnachrichten
WET	Wetterbericht/-vorhersage/-interview

Tab. 1: Beschreibung der Nachrichtenelemente und ihre Kodierung (vgl. Schwiesau/Grawunder/Bose 2011)

Nachrichten, egal ob im Radio oder im Fernsehen vorgetragen, sind nicht nur eine Aneinanderreihung loser Nachrichtenmeldungen, sondern haben in der Regel eine festgelegte Struktur, die aus bestimmten Elementen wie Kernnachrichten, Berichten, Sportnachrichten, Wirtschaftsmeldungen, Wetterberichten und Verkehrsmeldungen besteht. Für eine Übersicht über die Elemente einer Radionachrichtensendung siehe Tabelle 1.

Die Mischung verschiedener textlicher Elemente erfordert eine Diskursstruktur. Die einzelnen Nachrichtenelemente erfüllen verschiedene Funktionen und unterscheiden sich entsprechend in prosodischen und anderen linguistischen Merkmalen. So ist z. B. die Ankündigung eines Ortes unmittelbar vor einer Kernnachricht im Vergleich zu dieser syntaktisch verkürzt und weist eine besondere Form fallender Intonation auf. Was den Unterschied in der Sprechgeschwindigkeit zwischen den Elementen betrifft, so fanden Grawunder/Kettel (2015) im Wesentlichen keinen Unterschied zwischen deutschen Nachrichtenlesungen von Kernnachrichten und Berichten (beide 4,3–4,4 Silben pro Sekunde, im Folgenden s/s), aber einen Rückgang auf 3,3 s/s für das dazwischenliegende überleitende Intro-Element. Insgesamt deuten diese Werte auf eine insgesamt langsamere Sprechgeschwindigkeit hin, insbesondere im Vergleich zu vorgelesener Sprache von nicht-professionellen Sprecherinnen und Sprechern (5,97 s/s in Pellegrino/Coupé/Marsico 2011; zur Variabilität von Sprechgeschwindigkeit siehe Kleber/Schmid im Dr.), aber auch den in Iivonen/Niemi/Paananen (1995) untersuchten Nachrichtensprechenden (5,8 s/s in).

Eine für das Genre möglicherweise generell langsamere Sprechgeschwindigkeit (siehe aber unten) stünde in Einklang mit der Beobachtung, dass der Sprechstil von Nachrichtensprechenden ein Paradebeispiel für *clear speech* ist, die sich eher durch Hyper- als durch Hypoartikulation (Lindblom 1990) auszeichnet. Insbesondere ältere Nachrichtensprechende haben – oft im Zusammenhang mit einer Schauspielausbildung – ein Sprechtraining erhalten. Diese Form der Sprechausbildung ist zwar für Nachrichtensprechende jüngerer Nachrichtensendungen nicht mehr obligatorisch, aber auch heute noch dürfen beim Saarländischen Rundfunk (SR) nach Angaben des Senders nur Sprecherinnen und Sprecher Nachrichten vorlesen, deren Sprechfähigkeit für dieses Medium zuvor geprüft wurde und die eine Sprechfreigabe erhalten haben. Dabei liegt der Schwerpunkt auf Sprechfluss, Verständlichkeit, Klarheit der Sprachleistung und einer standardnahen Aussprache. Dies führt wiederum dazu, dass die Sprache von Nachrichtensprechenden als vorbildliche Standardaussprache gelten kann. Sowohl im Britischen Englisch (BBC-Aussprache, siehe z. B. Ladefoged 2001) als auch im Deutschen (Lameli 2004) dient die Aussprache von Nachrichtensprechenden als Referenz für die Beschreibung von Standardsprache. Im Deutschen finden sich aber auch dort regionalsprachliche Merkmale (Spiekermann 2007). Die Vorbildfunktion von Nachrichtensprechenden zeigt sich besonders an der oft engen Wechselbeziehung zwischen der in Aussprachewörterbüchern kodifizierten Standardsprache und deren lautliche Umsetzung durch Nachrichtensprechende (z. B. Krech et al. 2009). Ihre Aussprache spiegelt damit den jeweils aktuellen Stand der Standardaussprache wider und sollte bei Untersuchungen zu Lautwandelprozessen (siehe 1.2) dieser Varietät mit berücksichtigt werden.

Ungeachtet dessen sollte die Aussprache von Nachrichtensprechenden nicht mit der Aussprache von nicht-professionellen Sprecherinnen und Sprechern der Standardaussprache gleichgesetzt werden, denn Unterschiede zwischen den Gruppen sind für eine Reihe von Sprachen belegt. Mok/Fung/Li (2014) verglichen beispielsweise Kantonesisch bei profes-

sionellen Fernsehsprechenden im Vergleich zu nicht-professionellen Sprechern und Sprecherinnen. Ihre Ergebnisse zeigten, dass die kantonesischen Nachrichtensprechenden nicht nur einen größeren Tonhöhenbereich und eine höhere Variabilität der Silbenlänge aufwiesen als die nicht-professionellen Sprecher und Sprecherinnen, sondern im Gegensatz zu den oben stehenden Beobachtungen zum Deutschen auch deutlich schneller sprachen. Barbosa/Madureira/de Mareüil (2017) zeigten anhand von Daten in vier Sprachen (europäisches und brasilianisches Portugiesisch, Französisch, Deutsch), dass sich professionelle und nicht-professionelle Sprecher und Sprecherinnen beim Lesen von Nachrichten und anderen Genres in mehreren prosodischen Parametern unterscheiden, darunter die Grundfrequenz (f0) und die Intensität. In Bezug auf Pausen stellten Trouvain/Werner/Möbius (2020) bei deutschen Radionachrichtensprechenden fest, dass sie durchschnittlich 16 Pausen pro Minute Sprechzeit produzierten und damit weniger im Vergleich zu anderen Genres. 80% dieser Pausen wurden zudem mit einem hörbaren Einatmungsgeräusch realisiert. Dieser Wert liegt damit deutlich über jenen, die für andere Genres gemessen wurden. Auch die Pausendauer war bei Radionachrichten viel kürzer im Vergleich zu jenen in vorgelesenen Erzählungen oder Gesprächen von nicht-professionellen Sprechern und Sprecherinnen. Im Tschechischen wiederum ist zwar die Zahl der Pausen beim Vorlesen von Nachrichten niedriger im Vergleich zu anderen Genres (hier dem Rezitieren von Gedichten), die Dauer der Pausen unterschied sich jedoch nicht (Šturm/Volín 2023). Nachrichtensprache wie auch Sprachaufnahmen professioneller Sprecherinnen und Sprecher unterscheiden sich also in vielerlei Hinsicht. Dieser Aspekt muss in Untersuchungen wie der vorliegenden berücksichtigt werden, obwohl hier weder das Genre Nachrichtensprache noch der Vergleich zwischen professionellen und nicht-professionellen Sprecherinnen und Sprechern im Vordergrund stehen, sondern Nachrichtensprache aufgrund der Verfügbarkeit für linguistische Fragestellungen – hier im Bereich von Lautwandel – genutzt wird.

1.2 Prosodischer Wandel

Sprache als inhärent dynamisches System unterliegt grundsätzlich Wandel im Laufe der Zeit. Die entsprechenden Lautwandelprozesse können nicht nur im Nachhinein, wenn sie abgeschlossen sind, beschrieben werden, sondern auch, wenn sie aktuell stattfinden (vgl. Kleber 2016). Lautwandel wird in der Regel als Abweichung der Aussprache auf segmentaler Ebene über längere Zeiträume untersucht, die sich über wenige (vgl. Zellou/Tamminga 2014) bis zu vielen Generationen (vgl. Hahn/Siebenhaar 2019, S. 232) erstrecken kann. Er kann jedoch auch prosodische Merkmale beeinflussen (manchmal auch als prosodischer Wandel bezeichnet, vgl. Page 1999; Bose/Grawunder/Schwarze 2018). Neben dem Einfluss systeminterner Faktoren auf Wandelprozesse (z. B. durch Koartikulation, Dekakzentuierung; vgl. Kleber 2016), sind extern motivierte Veränderungen innerhalb eines gesprochenen Genres zu erwarten. Letztere ergeben sich aus einem dynamischen Umfeld, wie ein sich schnell ausbreitendendes Medium in politisch, wirtschaftlich, und sprachtechnologisch sich verändernden Zeiten. Die Aufgabe bleibt jedoch bestehen, Informationen in bestmöglicher Form von einem Sprechenden an viele Zuhörende zu übertragen. Prosodische Veränderungen im Laufe der Zeit sind für andere Rundfunkgenres wie Fußballreportagen dokumentiert (Trouvain 2015). Daher ist auch im Genre der Radionachrichten über einen Zeitraum von mehreren Jahrzehnten mit Veränderungen der prosodischen Merkmale zu rechnen.

De Mareüil/Rilliard/Allauzen (2012) stellten in einer Untersuchung audiovisueller Nachrichten aus Frankreich von 1945 bis 1997 fest, dass in dem ca. 50-jährigen Untersuchungszeitraum eine Reihe prosodischer Merkmale zurückgegangen sind, darunter die durchschnittliche f0, der Tonhöhenanstieg in Verbindung mit der Anfangsbetonung, die Vokaldauer als

Merkmal einer emphatischen Anfangsbetonung sowie die Dauer sogenannter *Inter-Pause Units* (nachfolgend IPUs), d. h. zeitlichen, sprachlich gefüllten Einheiten zwischen Pausen. Die Sprechgeschwindigkeit zeigte in dieser Studie jedoch keine Änderung über die Zeit. Rilliard et al. (2023) untersuchten französische Fernseh- und Radiodaten aus vier, 60 Jahre umspannenden Zeiträumen (1955–56, 1975–76, 1995–96 und 2015–16) und fanden eine Tendenz zu tieferen Stimmen bei den Sprechenden, unabhängig von deren Geschlecht.

Savino/Wehrle/Grice (2024) verglichen in einer *real-time* Untersuchung den prosodischen Stil italienischer Nachrichtensprecher aus den späten 1960er Jahren mit den Produktionen zweier Nachrichtensprechender, die 2005 dieselben Texte vorlasen. Ihre Ergebnisse zeigen, dass die f0 hinsichtlich des Anteils der Anstiege und Abfälle über die Zeit (*wiggliness*) sowie der räumlichen Auslenkung (f0-range, hier *spaciousness*) in beiden Epochen ähnlich waren. In den neueren Aufnahmen war jedoch die durchschnittliche f0 höher und die Sprechgeschwindigkeit schneller, was hauptsächlich auf die Verringerung der Anzahl und Dauer der Pausen zurückzuführen war, wodurch viel längere IPUs entstanden.

Diese Studien liefern sprachübergreifend eher heterogene Beobachtungen und erschweren somit eindeutige Vorhersagen, wie z. B. kürzere Pausen und längere IPUs in jüngerer Zeit. Auch lässt sich daraus nicht ableiten, dass Entwicklungen über einen Zeitraum von mehreren Jahrzenten linear verlaufen, z. B. durch kontinuierlich kürzere Pausen und kontinuierlich längere IPUs. Ergebnisse zur Vokalnasalierung im Amerikanischen Englisch belegen, dass Änderungen zwischen zwei Generationen sich in der dritten Generation wieder umkehren können (Zellou/Tamminga 2014). Auch in dieser *real-time* Studie stehen neben einer Evaluation der Nachrichtenelemente, Sprechgeschwindigkeit und Pausen im Vordergrund, deren Entwicklung über einen vergleichbaren Zeitraum wie in Rilliard et al. (2023) nun für das Deutsche explorativ untersucht werden soll.

2. Methode

Sowohl die Daten als auch alle im folgenden vorgestellten Tools zur Datenverarbeitung stehen für wissenschaftliche Forschungszwecke zur freien Verfügung. Bei der Nachnutzung der Nachrichtendaten können wir uns auf das deutsche Urheberrechtsgesetz (UrhG) berufen, da es sich nicht nur um öffentlich ausgestrahlte, sondern auch um öffentlich finanzierte Daten handelt. Insbesondere §60d UrhG erlaubt die automatische und systematische Vervielfältigung von Quellenmaterial zur Erstellung eines Korpus für wissenschaftliche Forschungszwecke (European Commission/Senftleben 2022; Margoni/Kretschmer 2022).

2.1 Daten

Für diese Pilotstudie wurden 43 (von insgesamt 49) Radionachrichten aus den Jahren 1956 bis 2017 von einem deutschen öffentlich-rechtlichen Radiosender, dem SR, analysiert.[1] Die ersten 31 der analysierten Aufnahmen aus den Jahren 1956 bis 2002 wurden uns vom Archiv des Senders zur Verfügung gestellt. Das Korpus wurde um weitere Aufnahmen aus

1 Sechs im Korpus enthaltene Radionachrichten blieben unberücksichtigt, da es sich um weitere Nachrichtensendungen aus bereits repräsentierten Jahren handelte und diese Jahre dann überrepräsentiert gewesen wären. Standen zwei Nachrichtensendungen pro Jahr zur Verfügung wurde nur die der Europawelle bzw. von SR1 in die Analyse einbezogen.

den Jahren 2003 bis 2017 aus der Sammlung „Nachrichtenarche“[2] (Grawunder 2011) ergänzt, von denen 12 in die Analyse einbezogen worden sind. In der Regel handelt es sich dabei um Nachrichtensendungen der Europawelle bzw. des Nachfolgesenders SR1. In wenigen Fällen um Sendungen der Saarlandwelle, dem Vorgängersender von SR3 und in einem Fall um eine ARD-Nachrichtensendung. Die Vergleichbarkeit der Daten wird dadurch gewährleistet, dass alle Sendungen im SR in Saarbrücken aufgezeichnet worden sind. Wie die Zahlen bereits verraten, stehen uns nicht aus allen Jahren Daten zur Verfügung, auch weil gerade in den früheren Jahrzehnten Radionachrichten nur selten als archivierungswürdig angesehen wurden (Schwiesau/Grawunder/Bose 2011). Im Korpus nicht vertreten sind die Jahre 1958–1960, 1962–1968, 1970–1971, 1977, 1982, 1987, 1989, 2005, und 2014–2015. Für alle anderen Jahre fließt jeweils eine Nachrichtensendung pro Jahr in die Analyse ein (vgl. Fußn. 1). Die verschiedenen Jahre werden durch unterschiedliche Nachrichtensprechende repräsentiert, so dass das Korpus *real-time* Trendstudien ermöglicht, in denen ältere und jüngere Daten verschiedener Individuen verglichen werden (Bülow/Scheutz/Wallner 2019). Die Identifikation der Sprechenden ist derzeit noch nicht abgeschlossen: 76% der Kernnachrichten wurden aber bislang 14 Sprechern und 7 Sprecherinnen zugeordnet. In wenige Sendungen kann dieselbe Person die Kernnachrichten vorgetragen haben; von anderen Sprechenden liegt hingegen nur eine Sendung vor.

2.2 Vorverarbeitung

Ein Ziel der aktuellen Studie war die Spezifikation eines hochautomatisierten und effizienten Arbeitsablaufs, um anschließende Analysen gesprochener Sprachkorpora vorzubereiten. Die Gesamtdauer aller 49 Nachrichtensendungen der Demosprachdatenbank beläuft sich auf ca. 3,5 Stunden, die der hier untersuchten Sendungen auf 3 Stunden und 8 Minuten. Die Größe des Gesamtkorpus beträgt 1,22 GB. In einem ersten vorverarbeitenden Schritt wurden

1. die Audiodateien in WAV-Format mit dem Kommandozeilen-Tool ffmpeg automatisch konvertiert (Mono, Abtastrate mindestens 44,1 kHz, 16 Bit lineare Quantisierung; siehe ffmpeg-Codec unter https://ffmpeg.org/, Stand: 10.12.2025),
2. die Dateinamen angeglichen,
3. orthografische Transkripte mit satzähnlichen Segmenten (Chunks) mithilfe automatischer Spracherkennung (unter Verwendung eines Python-Skriptes) erstellt.

Die Dateinamenangleichung ist nicht nur im Sinne eines übersichtlicheren Korpus, sondern ermöglicht auch eine automatische Verarbeitung über Webdienste (siehe 2.3.1). Die automatische Spracherkennung (ASR) wurde mit einer lokalen Installation von WhisperX (Bain et al. 2023, Version 3.1.1) und mit den in (1) angegebenen Parametern durchgeführt:

2 Die „Nachrichtenarche“ wurde 2003 ins Leben gerufen mit dem Ziel, das Medium Nachrichtensprache u. a. korpuslinguistisch nutzbar zu machen und einer fehlenden Einheitlichkeit bei der Archivierung gesprochener Nachrichten entgegenzuwirken. Einmal jährlich, am 11. November, werden daher seit 2003 die Aufnahmen der 13-Uhr-Nachrichten von allen öffentlich-rechtlichen Radiosendern in Deutschland gesammelt.

(1) whisperx --model large-v3 --lang de
--compute_type float32 --output_dir results *.wav

Dies erzeugte (unter anderem) zeitlich mit dem Sprachsignal alignierte Transkriptdateien im .srt-Format (d. h. einem gängigen Textformat für Untertitel). Die Überprüfung, Korrektur und Weiterverarbeitung dieser Transkripte erfolgte anschließend im Transkriptionseditor Octra (Pömp/Draxler 2017) in Verbindung mit Octra Backend (Draxler/Pömp 2024), einer web-basierten Infrastruktur für Transkriptionsprojekte. Der genaue Arbeitsablauf besteht aus den folgenden Schritten (siehe Draxler et al. 2025 für weitere Details):

4. Manuelle Korrektur des in 3. generierten Transkripts inkl. Zusammenfassen der Chunks in Nachrichteneinheiten (d. h. Löschung von Chunk-Grenzen, siehe Abb. 1);
5. Manuelle Etikettierung der einzelnen in 4. erzeugten Nachrichteneinheiten nach ihrem Elementtyp (vgl. Tab. 1);
6. Automatisches Schneiden der Audiodateien entsprechend der in 4. korrigierten Chunk-Segmentgrenzen.

Der 4. und 5. Schritt in o. g. Aufzählung erfolgte parallel und nicht konsekutiv. Jede geschnittene Nachrichteneinheit besteht aus einem einzigen Nachrichtenelement-Etikett in spitzen Klammern (z. B. <COR>), gefolgt vom orthografischen ASR-Transkript der Nachricht (vgl. Abb. 1).

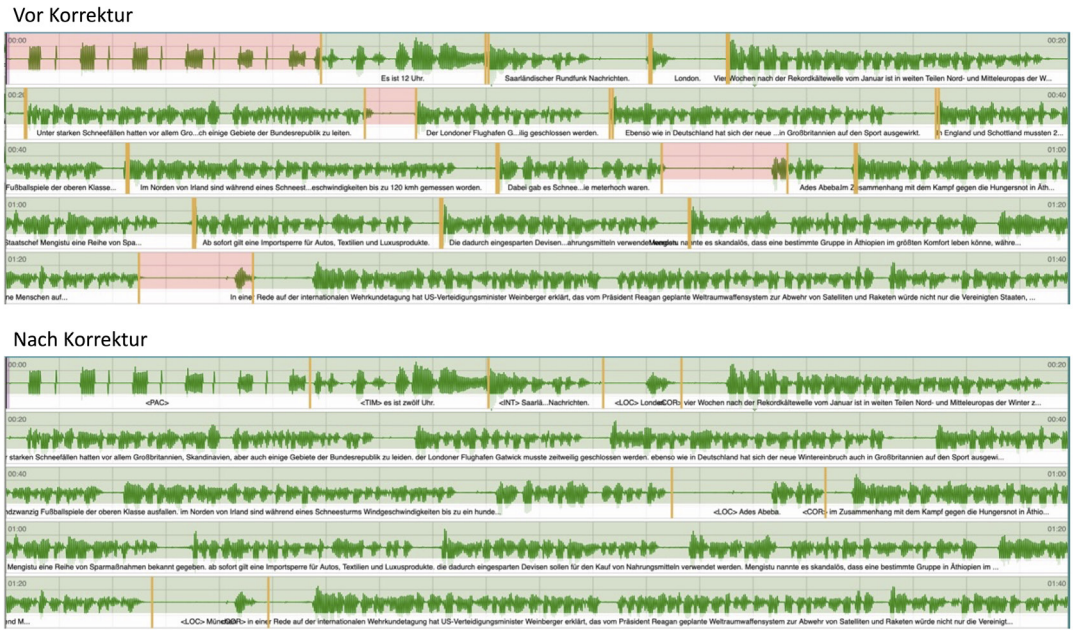

Abb. 1: Mithilfe ASR segmentierte und etikettierte Text- und Pausensegmente vor (oben) und nach (unten) der Korrektur in Octra (Pömp/Draxler 2017). Der Prüfprozess umfasst u. a. die Grenzkorrektur und die Etikettierung der Nachrichtenelemente (vgl. Tab. 1). Rot markiert sind Abschnitte, in denen die ASR aufgrund fehlender Spracherkennung kein Transkript geliefert hat. In der manuellen Korrektur wurden diese entweder mit <PAC> etikettiert (vgl. Tab. 1) oder – im Falle von Pausen – für deren in 2.3.1 beschriebene automatische Erkennung unetikettiert einem Snippet zugeordnet

Für die Berechnung der Wortfehlerrate (WER kurz für *Word Error Rate*), d. h. des Anteils von der ASR falsch erkannter Wörter, wurden die automatisch mit ASR generierten und die in Octra manuell korrigierten Transkripte normalisiert und anschließend mit der

Python-Bibliothek JIWER (Vaessen 2018) berechnet (siehe Draxler et al. 2025). Der WER lag bei 8,85% und war damit überraschend hoch für Sprachaufnahmen, die von professionellen Sprechern und Sprecherinnen unter sehr guten akustischen Bedingungen aufgenommen worden sind. Da die folgenden Arbeitsschritte auf den manuell wortkorrigierten Transkripten basieren, wirkt sich dies jedoch nicht auf die in 3. präsentierten Analysen aus.

Mit Blick auf die geplanten Analysen waren die Transkribierenden angewiesen, die rechte Grenze so nah wie möglich an das Ende des sichtbaren und hörbaren Sprachsignal zu setzen. Dies gewährleistet, dass Segmentgrenzen keine Pausen unterbrechen und somit eine zuverlässige Segmentierung (siehe 2.3.1) und Messung der Pausendauer (siehe 2.3.3; für Beispiele siehe Abb. 1 unten).

2.3 Annotation und Messung

2.3.1 Automatische Annotation

Das Schneiden der 43 analysierten Audiodateien in Octra resultierte in 832 Snippets mit je einem Nachrichtenelement. Mit einem Python-Skript wurde die BAS-Webservice-Pipeline ‚G2P → MAUS→ PHO2SYL'[3] (Kisler/Reichel/Schiel 2017) für jedes Dateipaar (d. h. wav-Datei plus die dazugehörige Textdatei mit orthographischer Transkription auf Phrasenebene) aufgerufen, um alle Snippets automatisch auf drei weiteren Ebenen in Wörter, Silben und Phoneme zu segmentieren und die Silben und Phoneme phonetisch zu etikettieren. Pausen werden in MAUS dabei auf allen drei Ebenen inklusive der Phonemebene segmentiert und durch das Etikett <p:> abgebildet. Das Output-Format entsprach dem EMU-Speech Database Management System (Winkelmann/Harrington/Jänsch 2017). Die 832 resultierenden Kleinstdatenbanken, bestehend aus je einer Audiodatei und der dazugehörigen Transkriptdatei im json-Format, wurden anschließend in einer einzigen EMU-Sprachdatenbank zusammengeführt (vgl. Abb. 2). Auf eine manuelle Korrektur der Segmentgrenzen und Etiketten wurde – im Gegensatz zur weniger zeitaufwendigen Korrektur der automatisch falsch erkannten Wörter (vgl. 2.2) – nicht nur aus Zeitgründen (vorerst) verzichtet: Diachrone Trends lassen sich anhand automatisch segmentierter Daten robust schätzen, da die Streuung innerhalb der Daten aufgrund fehlerhafter Segmentierungen größer ausfällt und Trends eher unter- als überschätzt werden (vgl. Kisler/Kleber 2019). Die Fehleranfälligkeit wurde zudem durch die zuvor erfolgte Wortkorrektur in Octra (vgl. 2.2) reduziert.[4]

3 G2P = Graphem zu Phonem, MAUS= Munich Automatic Segmentation, PHO2SYL = Silbentrennung basierend auf phonemischer Transkription.

4 Ein weiterer Vorteil der automatischen Segmentierung ist, dass diese nicht den natürlichen subjektiven Schwankungen bei einer manuellen Korrektur unterworfen ist. Im Gegensatz zu manuellen Eingriffen ist das Ergebnis automatischer Segmentierung replizierbar. Ein Vergleich zwischen automatischer und manueller Segmentierung ist ferner nur bedingt aussagekräftig, da es auch bei der manuellen Segmentierung keine eindeutig richtige Segmentierung gibt (siehe hierzu Kisler/Kleber 2019).

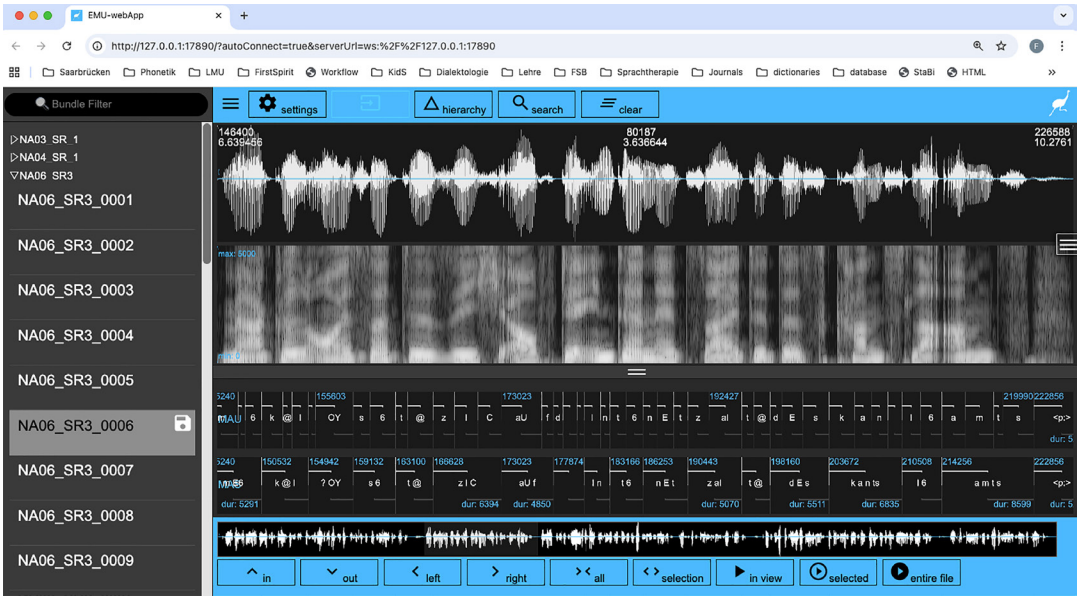

Abb. 2: Ausschnitt einer in der EMU-webApp geöffneten Kernnachricht (NA06_SR3_0006) einer Nachrichtensendung aus dem Jahr 2006 (NA06_SR3). Im Menü links sind weitere Nachrichtenelemente derselben Sendung sowie 2 weitere Nachrichtensendungen sichtbar. Unter dem Signalausschnitt sind die Phonem- und die Silbensegmentierung dargestellt

2.3.2 Signalbasierte Messung

Die Sprechgeschwindigkeit (Anzahl der Silben[5] pro Sprechzeit *inklusive* Pausen) und die Artikulationsgeschwindigkeit (Anzahl der Silben pro Artikulationszeit *exklusive* Pausen) sowie die Pausen pro Minute wurden zunächst auf der Grundlage des unsegmentierten akustischen Signals mit Hilfe des Praat-Skripts von de Jong/Wempe (2009)[6] gemessen. Das Skript erkennt Stimmhaftigkeit durch Periodizitätsanalyse und detektierter Intensitätsgipfel, denen Intensitätsminima vorausgehen, die als potenzielle Silbenkerne betrachtet werden. Das Skript löscht anschließend Peaks, die nicht stimmhaft sind. So werden manchmal Gipfel mit geringer Intensität übersehen, während Passagen mit hoher Intensität manchmal Doppelgipfel zugeordnet werden. Die Maße wurden mit linearen Regressionsmodellen in R statistisch ausgewertet (siehe 3.2 für Details).

2.3.3 MAUS-basierte Messung

Zur Validierung der rein signalbasierten Messung der Sprechgeschwindigkeit und des obigen Arbeitsablaufs wurden zusätzlich die Sprechgeschwindigkeit und die Pausen auf Grundlage einer Signal+Text-basierten MAUS-Segmentierung der Phoneme und Silben gemessen. MAUS segmentiert und etikettiert Sprachlaute auf Grundlage sowohl akustischer Merkmale, die aus dem Signal extrahiert werden, als auch lexikalischer Informationen und phonotaktischer Regeln, die aus Trainingsdaten gewonnen werden (vgl. Kisler/Reichel/Schiel 2017). Dadurch können phonologische Prozesse wie Assimilationen oder Elisionen, die u. a. zu den spontansprachlich häufigen Abweichungen von kanonischer Silbifizierung führen, Rechnung getragen werden. Mit MAUS segmentierte Daten bieten

5 Weder hier noch bei dem in 2.3.3 beschriebenen Ansatz entsprechen die Silben notwendigerweise den kanonischen, sondern den im Signal erkannten Silben. MAUS berücksichtigt die Wahrscheinlichkeit, mit der Silben realisiert werden und trägt Silbenelisionen Rechnung.

6 https://github.com/FieldDB/Praat-Scripts/blob/main/praat-script-syllable-nuclei-v2file.praat (Stand: 10.12.2025).

sich daher besonders für den Vergleich mit der in 2.3.2 beschriebenen Methode an. Zu diesem Zweck wurde die EMU-Sprachdatenbank in R unter Verwendung des Pakets emuR (Winkelmann/Harringont/Jänsch 2017) nach verschiedenen Labels (z. B. Nachrichtenelemente, Pausen, Silben) abgefragt. Zusätzlich zu den oben genannten Maßen wurde die Dauer der von MAUS segmentierten Pausen und die Dauer aller Segmente zwischen den Pausen, d. h. die IPUs, berechnet. Die Maße wurden mit linearen Regressionsmodellen in R geprüft (für Details siehe 3.2).

3. Ergebnisse

3.1 Verteilung der Nachrichtenelemente

Vor der Analyse der Sprechtempi und Pausen wurde das Korpus zunächst auf die Verteilung der Nachrichtenelemente hin explorativ-deskriptiv untersucht. Dabei stand einerseits der proportionale Anteil der Elemente *an* sowie deren Anordnung *innerhalb* einer Nachrichtensendung im Fokus. Nach der Sichtung aller Nachrichtensendungen und aller Nachrichtenelemente werden im Folgenden die Entwicklung der Kernnachrichten und Reportagen im Vergleich zu allen anderen Elementen anhand ausgewählter Jahre (= Sendungen) beschrieben. Letzeres dient der Übersichtlichkeit. Die Auswahl der Jahre erfolgte so, dass der Zeitraum breitestmöglich (1957–2017) und gleichmäßig (MW = 10 Jahre, SD = 1,1) abgebildet wird und jedes Jahrzehnt vertreten ist (vgl. 2.1).

Abbildung 3 zeigt zwei Maße für diese Stichprobe von Nachrichtensendungen. In der linken Grafik in Abbildung 3 ist zu sehen, dass der prozentuale Anteil der Kernnachrichten (COR) im Laufe der Zeit ab- und der der Berichte (REP) zunimmt. Der Anteil aller anderen Elemente (OTH) hat abgenommen.

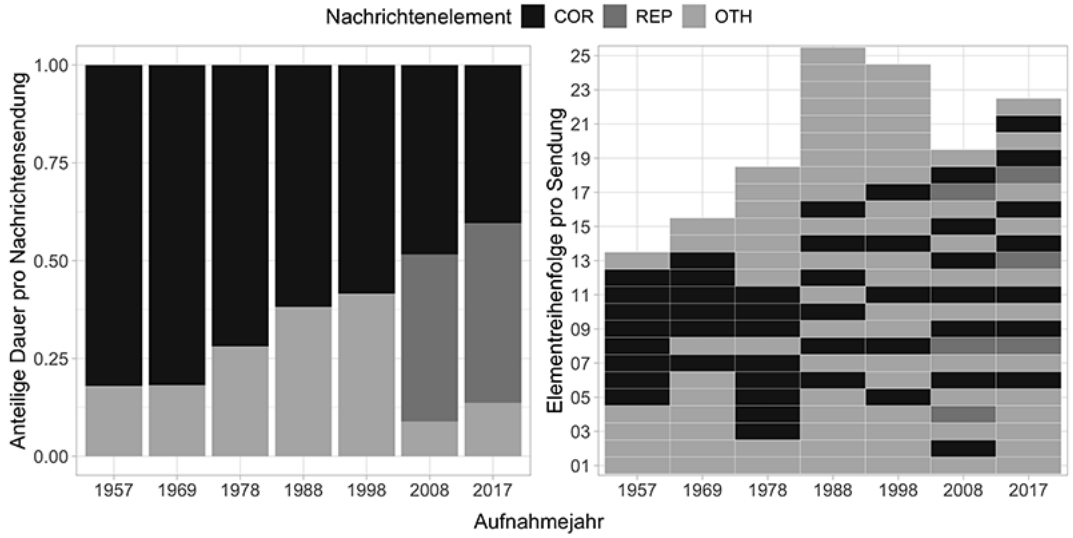

Abb. 3: An einer Nachrichtensendung anteilige Dauer ausgewählter Nachrichtenelemente (links) und deren Reihenfolge innerhalb einer Sendung (rechts; COR = Kernnachricht, REP = Bericht, OTH = andere)

Mit Fokus auf Kernnachrichten (COR) zeigt die rechte Grafik in Abbildung 3, dass diese Art von Elementen in den ältesten Sendungen früher einheitlicher am Ende einer Sendung vorgetragen wurden. Seit den 1980er-Jahren werden Kernnachrichten mit anderen Nachrichtenelementen durchsetzt präsentiert. Während die Zahl der Kernnachrichten innerhalb

einer Sendung ungefähr gleich geblieben ist, hat die Zahl anderer Nachrichtenelemente zudem zugenommen.

3.2 Tempo- und Pausenanalysen der Kernnachrichten

Die Analysen in diesem Abschnitt konzentrieren sich auf die Teilmenge der 251 Kernnachrichten aller 43 hier analysierten Sendungen. Mithilfe linearer Regressionen wurde der Zusammenhang zwischen Aufnahmejahr (= Sendung) und Sprechgeschwindigkeit bzw. Pausen pro Minute geprüft. Zuvor wurden die für jede Kernnachricht vorliegenden Werte der abhängigen Variable gemittelt, so dass pro Sendung (= Aufnahmejahr) nur ein Wert pro Sprecher bzw. Sprecherin vorlag, und die Verteilung der Residuen geprüft.

In Abbildung 4 ist die aus der signal- und der MAUS-basierten Analyse abgeleitete Sprechgeschwindigkeit dargestellt. Beide Analysen zeigen einen allmählichen Anstieg der Sprechgeschwindigkeit im Zeitraum von 1956 bis 2017, was sich in signifikanten Korrelationen widerspiegelt (signalbasierten Analyse: $R^2 = 0{,}29$, $F[1{,}41] = 16{,}56$, $p < 0{,}001$; MAUS-basierte Analyse: $R^2 = 0{,}17$, $F[1{,}41] = 8{,}68$, $p < 0{,}01$). Bei Betrachtung einzelner Zeiträume scheinen dabei Phasen der Sprechgeschwindigkeitsveränderung (1956–1957, 2006–2017) auf Phasen der Stabilität zu folgen (1986–1995). Eine statistische Analyse dieser Phasen ist erst bei Vorlage mehrerer Sprecherdaten sinnvoll. Die Veränderungen über die Zeit sind in beiden Analysen insgesamt ähnlich, obwohl die Sprechgeschwindigkeit bei der MAUS-basierten Analyse im Vergleich zur signalbasierten Analyse im Allgemeinen höher ist.

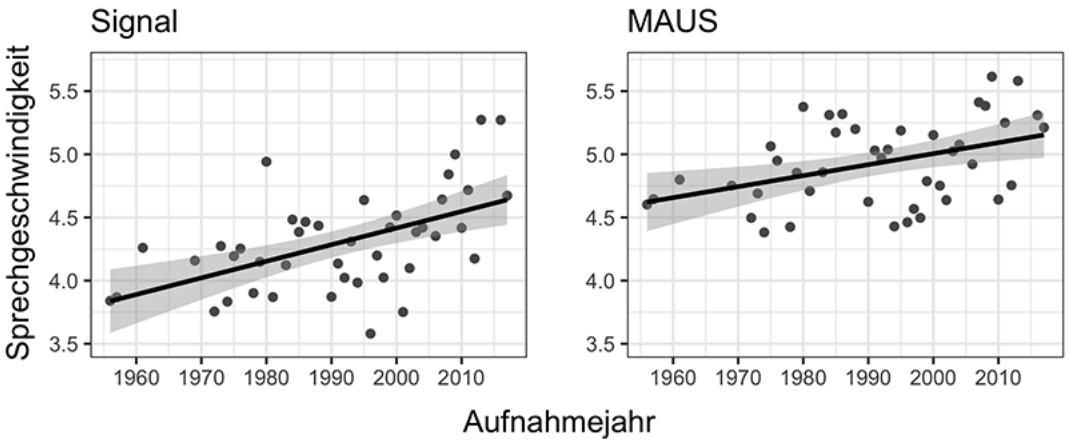

Abb. 4: Sprechgeschwindigkeit (s/s inklusive Pausen) in Abhängigkeit des Aufnahmejahres getrennt nach Messmethode (links: signalbasiert, rechts MAUS-basiert)

Die Anzahl der Pausen pro Minute ist hingegen beim signalbasierten Ansatz insgesamt geringer als beim MAUS-basierten Ansatz (siehe Abb. 5). Aber auch hier unterscheiden sich die Messungen nicht hinsichtlich des allgemeinen Trends hin zu weniger Pausen pro Minute in den jüngeren Aufnahmen (signalbasierte Analyse: $R^2 = 0{,}15$, $F[1{,}41] = 7{,}2$, $p < 0{,}05$; MAUS-basierte Analyse: $R^2 = 0{,}24$, $F[1{,}41] = 13{,}07$, $p < 0{,}001$).

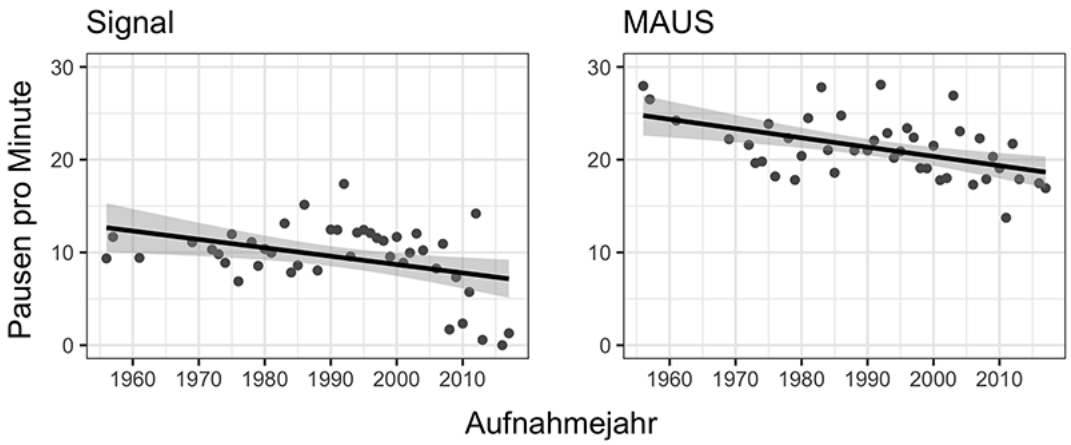

Abb. 5: Mittlere Anzahl der Pausen normalisiert für die mittlere Dauer der Kernnachrichten in Abhängigkeit des Aufnahmejahres getrennt nach Messmethode

Die jeweils ähnlichen Ergebnisse für beide Ansätze legitimieren die messmethodisch unabhängige Analyse der verbleibenden abhängigen Variablen, die entweder aus der signal- (Artikulationsrate, vgl. 2.3.2) oder der MAUS-basierten (Pausen- und IPU-Dauer, vgl. 2.3.3) Messung abgeleitet wurde.

Die linke Graphik in Abbildung 6 zeigt, dass es keine signifikante Korrelation zwischen Aufnahmejahr und Artikulationsrate im hier analysierten Korpus gibt (R^2 = 0,04, F[1,41] = 1,9, p = 1,18), die Artikulationsrate einzelner IPUs also im Gegensatz zur Sprechgeschwindigkeit nicht signifikant abgenommen hat.

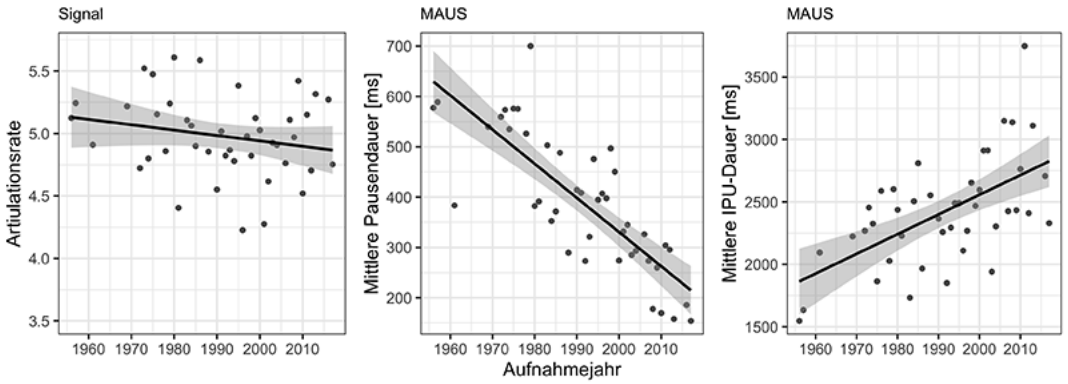

Abb. 6: Artikulationsrate (s/s exklusive Pausen; links), mittlere Pausen- (Mitte) und IPU-Dauer (rechts) in Abhängigkeit des Aufnahmejahres

Die mittlere Dauer der Pausen (Abb. 6 Mitte) hingegen hat über den 60-jährigen Zeitraum deutlich abgenommen (R^2 = 0,64, F[1,41] = 74,28, p < 0,001). Die Kürzung der Pausendauer geht einher mit einer über die Jahrzehnte signifikanten Zunahme der IPU-Dauer (Abb. 6 rechts; R^2 = 0,35, F[1,41] = 21,76, p < 0,001). Der beobachtete Trend in der Zunahme der Sprechgeschwindigkeit lässt sich somit auf weniger und kürzere Pausen zurückführen. Und obwohl die IPUs in den neueren Aufnahmen länger waren, hat sich die Artikulationsrate nicht erhöht, d. h. in jüngeren Sendungen wurden nicht mehr Silben pro Sekunde realisiert im Vergleich zu älteren.

4. Diskussion

Die Ziele der vorliegenden Studie lagen in der Vorstellung eines halbautomatisierten Verfahrens zur Aufbereitung eines Nachrichtenkorpus für Sprachwandelstudien im Allgemeinen und Lautwandelstudien im Besonderen sowie ersten Analysen zur Verteilung von Nachrichtenelementen innerhalb einzelner Sendungen und von Sprechgeschwindigkeit, Nachrichten- und Pausendauern in den Kernnachrichten.

Sowohl die Anordnung der Nachrichtenelemente als auch die Sprechgeschwindigkeit haben sich in dem hier untersuchten Korpus über den 60-jährigen Zeitraum verändert. Der Anteil der Kernnachrichten ist in den jüngeren Sendungen zurückgegangen und wird nicht länger *en bloc* präsentiert. Im Gegensatz zu französischen Radionachrichten (de Mareüil/Rilliard/Allauzen 2012), aber in Übereinstimmung mit früheren Berichten zum Italienischen (Savino/Wehrle/Grice 2024) zeigten die signal- und MAUS-basierten Analysen der Kernnachrichten, dass die Sprechgeschwindigkeit im Laufe der Zeit zunahm[7], nicht jedoch die Artikulationsgeschwindigkeit, was im hier untersuchten Korpus auf weniger und kürzere Pausen zurückzuführen ist. Inwiefern das Aufkommen privatrechtlicher Radiostationen und/oder die Anpassung an ein sich veränderndes Zielpublikum als mögliche Erklärungsansätze herangezogen werden können, wäre in weiteren Studien zu prüfen. Gleiches gilt für die Untersuchung einer möglichen Zunahme der Sprechgeschwindigkeit in der Standardsprache nicht-professioneller Sprecherinnen und Sprecher im selben Zeitraum, wobei dies aufgrund fehlender, vergleichbarer Daten kaum möglich ist. Gerade aufgrund ihrer Verfügbarkeit und der Repräsentativität für die Standardsprachen bieten sich die Daten für Sprachwandelstudien zur Standardaussprache an (vgl. 1.1 und Bose/Grawunder/Schwarz 2018). Dennoch ist anzumerken, dass die Stichprobe durch einzelne Sprechende, die jeweils ein bis wenige Jahr(e) repräsentieren, beeinflusst worden sein kann. Es handelt sich nicht um eine Langzeitstudie einzelner Nachrichtensprechenden. Die aktuelle Stichprobe von 43 Fernsehnachrichten mit 251 Kernnachrichtenelementen aus fast 60 Jahren muss also erweitert werden – möglichst mit vergleichbaren Daten anderer öffentlich-rechtlicher Sender – um die hier beschriebenen Trends zur Sprech- und Pausenrate vertiefend zu untersuchen. Auch eine weiterführende Analyse von Pausen, deren Bedeutung hier schon deutlich wurde, sowie weiterer prosodischer Merkmale (vgl. 1.2) ist wichtig und angedacht. Die Bewertung der Qualität der Pausenerkennung wird dabei zentral sein.

Die bei der Datenaufbereitung gewonnenen Erfahrungen haben zu einer Verbesserung des Octra-Backends geführt: Die lokale ASR ist nun als Hintergrundprozess innerhalb des Tools zugänglich, und das Schneiden des Audiosignals sowie die Erstellung paralleler Textdateien wurden als neue Funktionen integriert. Eine automatische Kennzeichnung der Nachrichtenelemente könnte von einer KI-basierten Lösung profitieren, sobald entsprechende Trainingsdaten mit manuellen Segmentierungen vorliegen. Ebenso wäre die automatische Erkennung, Reduzierung und Subtraktion von nichtsprachlichen Elementen (Atmung, Musikuntermalung) ein weiterer Schritt in der Datenvorverarbeitung, um bessere Ergebnisse bei der Erkennung von Pausen und Silben zu erzielen. Die Vorverarbeitung von Rundfunkaufnahmen für phonetisch-linguistische Korpora (die auch für andere Wissenschaftsbereiche relevant sind) ist aufwendig. Eine massive Reduzierung des manuellen Arbeitsaufwands, wie sie hier gezeigt wird, ist daher zu begrüßen.

7 Das erklärt möglicherweise auch die unterschiedlichen Sprechgeschwindigkeitswerte in Grawunder/Kettel (2015) und Iivonen/Niemi/Paananen (1995) (vgl. 1.1).

Literatur

Bain, Max/Huh, Jaesung/Han, Tengda/Zisserman, Andrew (2023): WhisperX: Time-accurate speech transcription of long-form audio. In: Proc. Interspeech 2023, 20–24 August 2023, Dublin, Ireland, S. 4489–4493.

Barbosa, Plinio A./Madureira, Sandra/de Mareüil, Philippe B. (2017): Cross-linguistic distinctions between professional and non-professional speaking styles. In: Proc. Interspeech 2017, August 20–24, 2017, Stockholm, Sweden, S. 3921–3925.

Bose, Ines/Schwiesau, Dietz (Hg.) (2011): Nachrichten Schreiben, sprechen, hören: Forschungen zur Hörverständlichkeit von Radionachrichten. Berlin: Frank & Timme.

Bose, Ines/Grawunder, Sven/Schwarze, Cordula (2018): Alles RElativ oder relaTIV? Sprachwandel, (Standard-)Aussprache und DaF-Unterricht. In: Moraldo, Sandro M. (Hg.): Sprachwandel: Perspektiven für den Unterricht Deutsch als Fremdsprache. (= Sprache – Literatur und Geschichte. Studien zur Linguistik/Germanistik 49). Heidelberg: Winter, S. 69–114.

Bülow, Lars/Scheutz, Hannes/Wallner, Dominik (2019): Variation and change of plural verbs in Salzburg's base dialects. In: Dammel, Antje/Schallert, Oliver (Hg.): Morphological variation. Theoretical and empirical perspectives. (= Studies in Language Companion Series 207). Amsterdam/Philadelphia: Benjamins, S. 95–134.

de Jong, Nivja H./Wempe, Ton (2009): Praat script to detect syllable nuclei and measure speech rate automatically. In: Behavior Research Methods 41, 2, S. 385–390.

De Mareüil, Philippe B./Rilliard, Albert/Allauzen, Alexandre (2012): A diachronic study of initial stress and other prosodic features in the French news announcer style: corpus-based measurements and perceptual experiments. In: Language and Speech 55, 2, S. 263–293.

Draxler, Christoph/Pömp, Julian (2024): Octra Backend – eine skalierbare Infrastruktur für Transkriptionsprojekte. In: Konferenz Elektronische Sprachsignalverarbeitung (ESSV) 2024, Regensburg, S. 102–107.

Draxler, Christoph/Kleber, Felicitas/Grawunder, Sven/Trouvain, Jürgen (2025): Teilautomatisierter Workflow zur Aufbereitung großer Audiodatenmengen für signalbasierte Analysen. In: Konferenz Elektronische Sprachsignalverarbeitung (ESSV) 2025, Halle/Saale, S. 138–145.

Dussel, Konrad (2022): Deutsche Rundfunkgeschichte. 4. Aufl. Köln: Halem.

European Commission: Directorate-General for Research and Innovation/Senftleben, Martin R. F. (2022): Study on EU copyright and related rights and access to and reuse of data. Luxemburg: Publications Office of the European Union. https://data.europa.eu/doi/10.2777/78973 (Stand: 10.12.2025).

Gordon, Elizabeth/Maclagan, Margaret/Hay, Jennifer (2007). The ONZE corpus. In: Beal, Joan C./Corrigan, Karen P./Moisl, Hermann L. (Hg.): Creating and digitizing language corpora. Bd. 2: Diachronic databases. London: Palgrave Macmillan, S. 82–104.

Grawunder, Sven (2011): Die Erforschung des Sprechens mittels Nachrichtenkorpora: Die Nachrichtenarche der ARD. In: Bose/Schwiesau (Hg.), S. 157–178.

Grawunder, Sven/Kettel, Sonja (2015): Anmoderieren / Überleiten / Antexten. In: SPIEL. Eine Zeitschrift für Medienkultur 1, 1, S. 151–170.

Hahn, Matthias/Siebenhaar, Beat (2019): Schwa unbreakable – Reduktion von Schwa im Gebrauchsstandard und die Sonderposition des ostoberdeutschen Sprachraums. In: Kürschner/Habermann/Müller (Hg.), S. 215–236.

Hecht, Robert/Riedler, Jürgen/Backfried, Gerhard (2002): German broadcast news transcription. In: 7th International Conference on Spoken Language Processing 2002, Denver, Colorado, USA, September 16–20, S. 1753–1756.

Iivonen, Antti/Niemi, Tuija/Paananen, Minna (1995): Comparison of prosodic characteristics in English, Finnish and German radio and TV newscasts. In: Proc. International Congress of Phonetic Sciences (ICPhS) 1995, Stockholm, S. 382–385.

Kisler, Thomas/Kleber, Felicitas (2019): Zur Validität automatisch segmentierter Daten: Eine akustische Analyse der mittelbairischen Lenisierung im Deutsch Heute-Korpus. In: Kürschner/Habermann/Müller (Hg.), S. 279–311.

Kisler, Thomas/Reichel, Uwe/Schiel, Florian (2017): Multilingual processing of speech via web services. In: Computer Speech & Language 45, S. 326–347.

Kleber, Felicitas (2016): Lautwandel. In: Domahs, Ulrike/Primus, Beate (Hg.): Handbuch Laut – Gebärde – Buchstabe. (= Handbücher Sprachwissen 2). Berlin/Boston: De Gruyter, S. 125–143.

Kleber, Felicitas (2025): Synchrone und diachrone Variation in der temporalen Struktur gesprochener Wörter in Varietäten der DACH-Region: Empirie, Simulation und ein interaktiv-phonetischer Lautwandelansatz. In: Proske, Nadine/Weber, Thilo/Dannerer, Monika/Deppermann, Arnulf (Hg.): Gesprochenes Deutsch: Struktur, Variation, Interaktion. (= Jahrbuch des Instituts für Deutsche Sprache 2024). Berlin/Boston: De Gruyter, S. 105–130.

Kleber, Felicitas/Schmid, Stephan (im Dr.): Towards a signal-based typology of consonant quantity in southern German varieties. In: Filipponio, Lorenzo/Dipino, Dalila/Garassino, Davide (Hg.): Exploring the relation between duration and length. Italo-Romance, Romance and beyond. (= Linguistik Aktuell). Amsterdam/Philadelphia: Benjamins.

Krech, Eva-Maria/Stock, Eberhard/Hirschfeld, Ursula/Anders, Lutz C. (2009): Deutsches Aussprachewörterbuch. Berlin/New York: De Gruyter.

Kürschner, Sebastian/Habermann, Mechthild/Müller, Peter O. (Hg.) (2019): Methodik moderner Dialektforschung: Erhebung, Aufbereitung und Auswertung von Daten am Beispiel des Oberdeutschen. (= Germanistische Linguistik 241–243). Hildesheim/Zürich/New York: Olms.

Labov, William/Rosenfelder, Ingrid/Fruehwald, Josef (2013): One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. In: Language 89, 1, S. 30–65.

Ladefoged, Peter (2001): Vowels and consonants – An introduction to the sounds of languages. Maldon, MA/Oxford: Blackwell.

Lameli, Alfred (2004): Standard und Substandard: Regionalismen im diachronen Längsschnitt. (= Zeitschrift für Dialektologie und Linguistik. Beihefte 128). Stuttgart: Steiner.

Lindblom, Björn (1990): Explaining phonetic variation: A sketch of the H&H theory. In: Hardcastle, William J./Marchal, Alain (Hg.): Speech production and speech modelling. Dordrecht: Springer, S. 403–439.

Margoni, Thomas/Kretschmer, Martin (2022): A deeper Llok into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology. In: GRUR International 71, 8, S. 685–701.

Mok, Peggy P. K./Fung, Holly S. H./Li, Jingwen (2014): A preliminary study on the prosody of broadcast news in Hong Kong Cantonese. In: Proc. Speech Prosody 2014, Dublin, S. 1072–1075.

Page, Richard (1999): The Germanic Verschärfung and prosodic change. In: Diachronica 16, 2, S. 297–334.

Pellegrino, François/Coupé, Christophe/Marsico, Egidio (2011): A cross-language perspective on speech information rate. In: Language 87, 3, S. 539–558.

Pömp, Julian/Draxler, Christoph (2017): Octra – A configurable browser-based editor for orthographic transcription. In: Proc. Phonetik und Phonologie 2017, Berlin, S. 145–148.

Rilliard, Albert/Doukhan, David/Uro, Rémi/Devauchelle, Simon (2023): Evolution of voices in French audiovisual media across genders and age in a diachronic perspective. In: Proc. International Congress of Phonetic Sciences (ICPhS) 2023, Prag, S. 753–757.

Savino, Michelina/Wehrle, Simon/Grice, Martine (2024): The prosody of Italian newsreading: a diachronic analysis. In: Proc. Speech Prosody, 2–5 July 2024, Leiden, The Netherlands, S. 836–840.

Schwiesau, Dietz/Grawunder, Sven/Bose, Ines (2011): Die Nachrichtenarche der ARD. In: Bose/Schwiesau (Hg.), S. 147–155.

Spiekermann, Helmut (2007): Standardsprache im DaF-Unterricht: Normstandard – nationale Standardvarietäten – regionale Standardvarietäten. In: Linguistik Online 32, 3, S. 119–137.

Šturm, Pavel/Volín, Jan (2023): Occurrence and duration of pauses in relation to speech tempo and structural organization in two speech genres. In: Languages 8: 23. https://doi.org/10.3390/languages8010023 (Stand: 22.1.2026).

Trouvain, Jürgen (2015): Notes on the development of speaking styles over decades — the case of live football commentaries. In: Proc. First International Workshop on the History of Speech Communication Research 2015 (HSCR), Dresden, Germany, September 4—5, S. 160–166.

Trouvain, Jügen/Werner, Raphael/Möbius, Bernd (2020): An acoustic analysis of inbreath noises in read and spontaneous speech. In: Proc. Speech Prosody 2020, 25—28 May, Tokyo, Japan, S. 789–793.

Vaessen, Nik (2018): JiWER: Similarity measures for automatic speech recognition evaluation. Python Package Index. https://pypi.org/project/jiwer/ (Stand: 10.12.2025).

Winkelmann, Raphael/Harrington, Jonathan/Jänsch, Klaus (2017): EMU-SDMS: Advanced speech database management and analysis in R. In: Computer Speech and Language 45, S. 392–410.

Zellou, Georgia/Tamminga, Meredith (2014): Nasal coarticulation changes over time in Philadelphia English. In: Journal of Phonetics 47, S. 18–35.

Kontaktinformation

Felicitas Kleber
Fachrichtung Sprachwissenschaft und Sprachtechnologie
Universität des Saarlandes
Campus C7.2
66123 Saarbrücken
E-Mail: kleber@lst.uni-saarland.de

Christoph Draxler
Institut für Phonetik und Sprachverarbeitung
Ludwig-Maximilians-Universität München
Schellingstraße 3
80799 München
E-Mail: draxler@phonetik.uni-muenchen.de

Jürgen Trouvain
Fachrichtung Sprachwissenschaft und Sprachtechnologie
Universität des Saarlandes
Campus C7.2
66123 Saarbrücken
E-Mail: trouvain@lst.uni-saarland.de

Sven Grawunder
Institut für Musik-, Medien- und Sprechwissenschaften
Martin-Luther-Universität Halle-Wittenberg
Emil-Abderhalden-Straße 26/27
06108 Halle (Saale)
E-Mail: sven.grawunder@sprechwiss.uni-halle.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.70

Charlotte Rein/Timo Schürmann

PALAVA

Eine Smartphone-App zur Erforschung der regionalen Alltagssprache in NRW

Abstract Since June 2023, the Landschaftsverbände Rheinland (LVR-Institut für Landeskunde und Regionalgeschichte) and Westfalen-Lippe (LWL-Kommission für Mundart- und Namenforschung) have been collecting data on everyday regional language in North Rhine-Westphalia using the PALAVA smartphone app. In the meantime, a customised analysis infrastructure running with *R* and *shiny* has been created. This enables secure input of the data processing and low-threshold exploration of the quantitative and spatial distribution of the data. An example analysis also shows that the data not only enables synchronous analyses, but also the observation of recent language change with the help of *apparent time*-analyses.

Keywords colloquial german, regional variation, apparent time change, data collection

1. Einleitung

Die Erforschung der regionalen Alltagssprache, also dem Bereich zwischen Basisdialekt und Standardsprache[1], hat in den letzten Jahrzehnten im deutschen Sprachraum stark zugenommen. Dies trifft auch auf Nordrhein-Westfalen zu. Die vorhandenen Untersuchungen konzentrieren sich dabei in der Regel auf einen Ortspunkt (z. B. zu Dortmund Salewski 1998, zu Erftstadt-Erp Rein 2020) oder einzelne Regionen (z. B. zum Westmünsterland Lanwer 2015, zu Teilen des Niederrheins und Westfalens Elmentaler/Rosenberg 2015, zum Niederrhein Elmentaler 2005) und erfassen oftmals nur phonologische Variation. Ein Desiderat stellt daher bislang die flächendeckende Erhebung vergleichbarer Daten für das gesamte Bundesland dar und somit auch die Möglichkeit, Aussagen über die räumliche Struktur der Sprachlage treffen zu können. Um diese Forschungslücke zu füllen, haben die Landschaftsverbände Rheinland (LVR-Institut für Landeskunde und Regionalgeschichte) und Westfalen-Lippe (LWL-Kommission für Mundart- und Namenforschung) eine Smartphone-App entwickelt, mit deren Hilfe umfangreiches Sprachmaterial erhoben werden kann, so dass eine Beschreibung von Struktur und Besonderheiten des regionalsprachlichen Spektrums möglich werden.

2. Sprachdatenerhebung mit Smartphone-App

Die Erhebung von regionaler Sprache mit Hilfe von Smartphone-Apps wird seit etwa zehn Jahren in verschiedenen Gebieten des deutschen Sprachraums und angrenzenden Ländern erprobt, insbesondere in der Schweiz (*Dialäkt Äpp* 2013, vgl. Kolly/Leemann 2015, *Voice Äpp* 2014, vgl. Hove et al. 2015, *Gschmöis* 2018, vgl. Hasse/Bachmann/Glaser 2021) und in

1 Zur Problematik dieses Begriffs bzw. des Konzepts vgl. Freywald et. al (2023, S. 49–70).

Luxemburg (*Schnëssen* 2018, vgl. Entringer et al. 2021).[2] Der Vorteil bei der Nutzung dieses Erhebungsinstruments ist der vergleichsweise geringe finanzielle und personelle Aufwand bei der Datensammlung: Nach dem Launch der App kann jede*r Interessierte eigenständig Antworten in die App einsprechen bzw. eingeben. Natürlich bedarf es einer durchdachten Presse- und Öffentlichkeitsarbeit, um die App bekannt zu machen, sowie einer konstanten Überwachung der Technik. Ein weiterer großer Gewinn gegenüber herkömmlichen Fragebögen, die handschriftlich oder digital ausgefüllt werden, ist die Möglichkeit, gesprochene Sprache zu erheben. Durch die gute Qualität, die Mikrophone handelsüblicher Smartphones inzwischen haben, entstehen bei der Erhebung mittels App Aufnahmen, die auch für phonologische Fragestellungen nutzbar gemacht werden können (vgl. z. B. Gilles 2019, 2023).

Natürlich hat die Verwendung einer Smartphone-App gegenüber anderen Erhebungsinstrumenten wie einem persönlichen Interview auch Nachteile. Nachfragen zu den gestellten Aufgaben sind beiderseits nicht möglich, was in der Regel zu einem gewissen Prozentsatz unbrauchbarer Antworten führt. Auch kann unterschiedliche Interpretation der Fragestellung nicht durch Rückfragen aufgeklärt werden. Dies erschwert dann den Vergleich der Daten untereinander. Auch hat sich gezeigt, dass Aufnahmen teilweise während starker Umgebungslautstärke gemacht werden, was zur Unverständlichkeit und Unbrauchbarkeit der Daten führen kann, insbesondere in Hinblick auf phonetische/phonologische Fragestellungen. Des Weiteren wäre theoretisch eine Mehrfachteilnahme derselben Person möglich, was tatsächlich jedoch selten der Fall ist (unter 3%, vgl. Reips 2002). Und es gibt natürlich immer wieder Menschen, die absichtlich falsche oder auch unangemessene Antworten geben. Wie aber bereits die Erfahrungen aus dem großen Online-Projekt „Atlas zur deutschen Alltagssprache“ (AdA) zeigen, werden all diese Probleme in der Regel durch die große Masse erhobener Daten ausgeglichen (vgl. Elspaß/Möller 2006, S. 147). Hinsichtlich der inhaltlichen Güte der App-Daten hat sich im Vergleich mit Material, das mit traditionellen Erhebungsmethoden erhoben worden ist, deutliche Ähnlichkeit gezeigt, so dass von einer Validität des neuen Datentyps ausgegangen werden kann (vgl. Leemann et al. 2014 und Leemann et al. 2016 zur /l/-Vokalisierung im Schweizerdeutschen sowie Leemann 2017 zur Artikulationsrate im Schweizerdeutschen).

2.1 Konzeption und Nutzung der Smartphone-App PALAVA

Die App PALAVA ist angelehnt an die bereits vorhandenen Sprach-Apps *Schnëssen*, *Gschmöis* sowie die französische App *Français de nos régions* gestaltet worden. Die inhaltliche Konzeption erfolgte durch die wissenschaftlichen Mitarbeiter*innen der beteiligten Institute, die technische Umsetzung durch die Firma i.Bros, die auch die drei zuvor genannten Apps programmiert hat.

2 Seit Erscheinen von PALAVA sind noch weitere Apps hinzugekommen: *OeDA!* (Dialekt-App für Österreich, Universität Salzburg), *DaBay* (Dialekt-App für Bayern, Universität München), *Plapper* (Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin).

Abb. 1: Startbildschirm der PALAVA-App

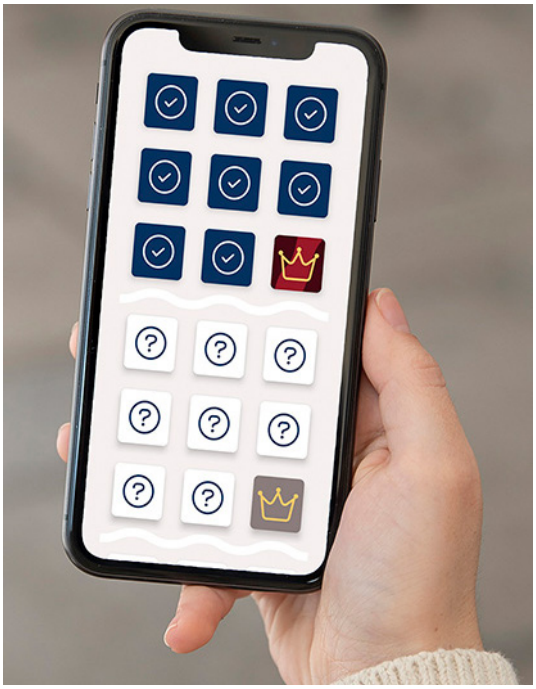

Abb. 2: Ansicht der Aufgabenübersicht

PALAVA gliedert sich in drei Module: „Mitmachen“, „Anhören“ und „Ergebnisse“ (vgl. Abb. 1).[3] Relevant für die Datenerhebung ist der Bereich „Mitmachen“. Nach einem Screen mit kurzen Erklärungen zur Teilnahme, werden einige demographische Daten abgefragt: Alter, Geschlecht, andere Muttersprachen als Deutsch und höchster Bildungsabschluss. Auf einer Karte kann der Ort, für den die Angaben gelten sollen (relevantester Ort für die sprachliche Prägung), markiert werden. In der Datenbank wird dieses Profil mit einem anonymisierten Identifikationsschlüssel versehen, so dass die Metadaten mit den Sprachdaten der Person verbunden werden können. Das Profil bleibt auf dem Smartphone gespeichert, so dass die Daten bei wiederholter Nutzung nicht erneut eingegeben werden müssen. Es ist aber möglich, mit einem Smartphone mehrere Profile anzulegen, um beispielsweise (älteren) Menschen, die kein eigenes Gerät haben, die Teilnahme zu ermöglichen. Im Anschluss erscheint die Aufgabenübersicht, die Fragen werden jeweils in Neunerblöcken präsentiert (vgl. Abb. 2).

Hierbei handelt es sich zum größten Teil um offene Fragen, in denen durch Fragesätze, Satzergänzungen, Bilder oder Videos Antworten, die als Sprachaufnahme eingesprochen werden, elizitiert werden sollen. Daneben gibt es einige geschlossene Fragen (ja/nein-Fragen, Single/Multiple Choice-Fragen), bei denen vorgegebene Antworten ausgewählt werden können. Alle Fragen werden schriftlich gestellt. Bei der letzten Frage jedes Fragesets handelt es sich um eine sogenannte Kronenaufgabe. Diese wird nur freigeschaltet, wenn die acht vorherigen Fragen beantwortet wurden. Dies soll zur konstanten Teilnahme und zur Beantwortung möglichst vieler Neunerblöcke motivieren (vgl. Hasse/Bachmann/Glaser 2021, S. 2). Seit der zweiten Fragerunde gibt es zudem Aufgaben, die eine längere zusammenhängende Erzählung evozieren sollen.[4] Die vielfältigen Fragen zielen auf Varia-

3 Vgl. auch Seiferheld (2024, S. 93–97).

4 Hierzu werden Erzählanreize formuliert, beispielsweise „Erzähl doch mal, wenn dich jemand schon einmal darauf angesprochen hat, dass man hört, dass du aus dieser Region kommst.“ oder „Erzähl doch mal von einem Erlebnis auf einer Zugreise in Deutschland.“.

tionsphänomene auf allen linguistischen Ebenen ab (Phonetik/Phonologie, Morphologie, Syntax, Pragmatik, siehe Abb. 3a), b) und c)).

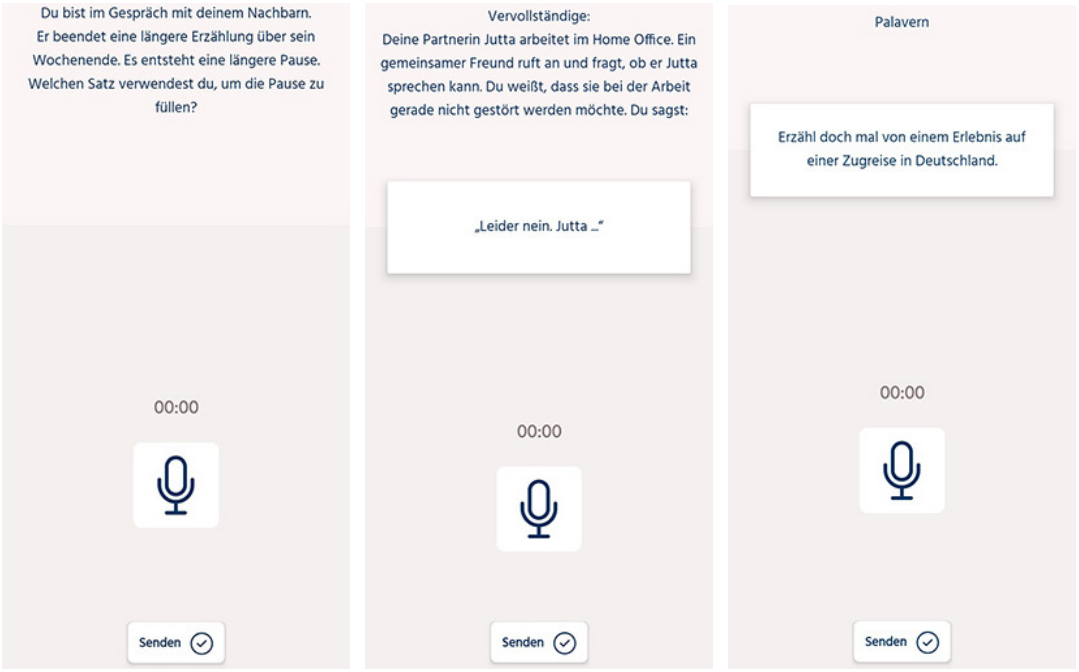

Abb. 3: a) Aufgabe zum Füllen von Gesprächspausen b) Aufgabe zum *am*-Progressiv c) Erzählaufgabe

Unter „Anhören" werden die Sprachaufnahmen (sofern die Teilnehmenden ihr Einverständnis gegeben haben) direkt nach der Eingabe auf einer Karte dargestellt[5] und für alle User*innen hörbar gemacht. Die Filterfunktion ermöglicht es, gezielt die verschiedenen Antworten zur gleichen Frage miteinander zu vergleichen. Damit wird ein Infotainment-Ansatz verfolgt: Die an der Sache interessierte teilnehmende Person soll einen direkten Einblick in die Daten bekommen und für ihre Datenspende mit Informationen über die Sprachvariation in der Region „belohnt" werden.

Der Bereich „Ergebnisse" präsentiert fertige Sprachkarten, die auf Grundlage der PALAVA-Daten angefertigt worden sind, und dient so der Wissenschaftskommunikation. Ein Kartenkommentar erläutert die regionale Verteilung sowie die (etymologische) Herkunft der verschiedenen Varianten. Monatlich wird eine neue Karte mit Kommentar eingestellt,[6] die User*innen werden mittels Push-Nachricht darüber informiert.[7] Die Betreuung der App sowie des dazugehörigen Instagramkanals sowie die Auswertung der Daten und die Erstellung der Karte (vgl. Abschn. 4) erfolgt durch die Mitarbeiter*innen des LVR-ILR und der LWL-KoMuNa.

Die App wurde am 09.06.2023 veröffentlicht und steht für die Betriebssysteme Android und iOS kostenfrei zur Verfügung. Bis Ende Juni 2025 haben sich über 15.600 Personen, wobei die älteste teilnehmende Person 1922 und die jüngste 2012 geboren wurde, die App heruntergeladen und insgesamt über 472.000 Antworten eingesprochen.[8] Die erste Fragerunde (Juni 2023 bis Oktober 2024) umfasste 108 Fragen, die zweite Fragerunde (seit Ende

5 Hierfür wird der Ortspunkt genutzt, den der/die User*in bei Abfrage der Metadaten als „relevantesten Ort für die sprachliche Prägung" angegeben hat.

6 Ein Beispiel für eine solche Karte ist Abbildung 8.

7 Bis Ende Mai 2025 sind 18 Karten und Kartenkommentare in der App veröffentlich worden.

8 Detailliertere Ausführungen zu den User*innen der App und ihrem Nutzungsverhalten finden sich in Rein/Schürmann (2024, S. 100–106).

Oktober 2024) 90 Fragen. Ein Großteil der Fragen wird von den Projektmitarbeitenden entwickelt. Daneben werden auch Fragen anderer Forschenden eingebunden, beispielsweise im Rahmen von Dissertationen oder Projekten.

2.2 Auswertungstool

Die Auswertung der zahlreichen Datenpunkte erfordert eine Infrastruktur, die kollaborative, zeitgleiche Zusammenarbeit an mehreren Standorten ermöglicht. Weit verbreitete und niedrigschwellige Lösungen wie GoogleSheets oder OnlyOffice (in Kombination mit Cloud-Systemen wie Nextcloud oder OwnCloud) sind voraussetzungsarm einsetzbar, sind aber datenschutzrechtlich problematisch oder mit Bezug auf Datenvalidität unzuverlässig. Gemeint ist damit, dass durch unkontrollierte Bearbeitung von mehreren Personen stets die Gefahr besteht, dass Dinge gelöscht werden oder in einem Maße uneinheitlich vermerkt werden, was die Auswertung erschwert. Aus diesen Gründen wurde eine eigene Datenbanklösung erstellt. Mithilfe einer Shiny-Anwendung[9] wird auf diese Datenbank zugegriffen und eine strukturierte Dateneingabe sowie niedrigschwellig erste Analysen zu Phänomenen und Sammlungen von Phänomenen ermöglicht. Aktuell wird dieses Tool nur projektintern genutzt. Zukünftig sollen Studierende und interessierte Bürger*innen ebenfalls eingebunden werden. Im Folgenden geben wir einen kurzen Einblick in die Prozesse nach der Datensammlung.

Das Auswertungstool strukturiert die verschiedenen Schritte jeweils danach, ob es auf die einzelnen Aufgaben (wie sie in der App erhoben wurden) oder auf Kollektionen, die aus mehreren Aufgaben gebildet werden können, abzielt. Erster Schritt ist in jedem Fall die Aufbereitung der Tondateien. Dies geschieht nur aufgabenbezogen. Als besonders zeitsparend hat sich die Einbindung von automatischer Transkription durch *Whisper* (Radford et al. 2022) erwiesen. Die Transkription steht als Vorschlag in der Eingabemaske und kann nach Überprüfung durch die annotierende Person übernommen werden (vgl. Abb 4, (1)). Weiterhin stehen bisher vergebene Transkriptionen oder Annotationen als Vorschlag in einem Auswahlmenü zur Verfügung (vgl. Abb. 4).

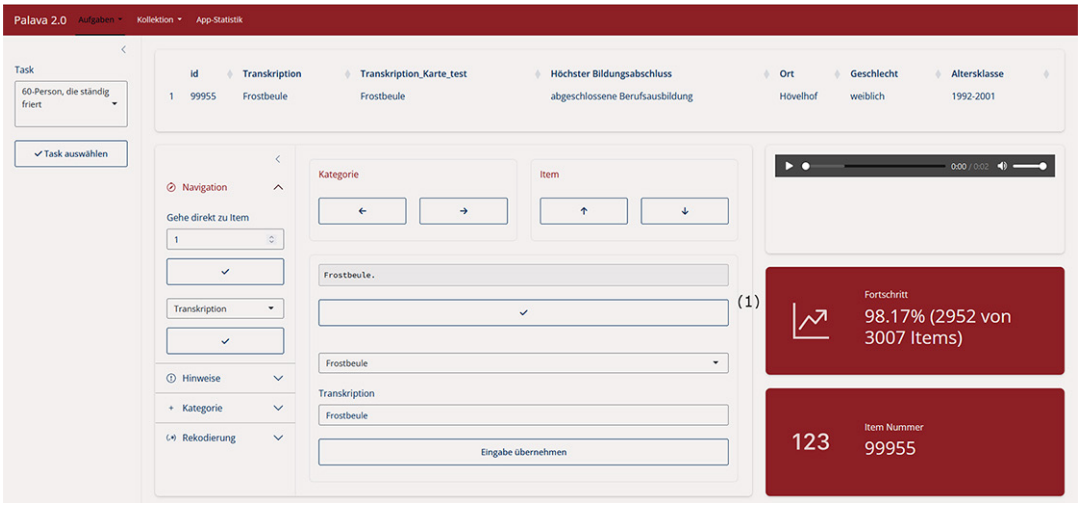

Abb. 4: Eingabemaske für die Datenaufbereitung (1): Automatische Transkription durch *Whisper*

Nach der ersten Aufbereitung kann dann eine Exploration der quantitativen Verhältnisse vorgenommen werden. Es können sowohl die ungruppierten Verteilungen als auch die

9 Shiny ist ein R-Paket, das es ermöglicht, grafische Oberflächen für R-Code zu erstellen (Chang et al. 2025).

Aufschlüsselung nach den erhobenen Sozialparametern Bildung, Alter und Geschlecht in Balkendiagrammen abgebildet werden. Dabei können relative und absolute Zahlen angezeigt werden.

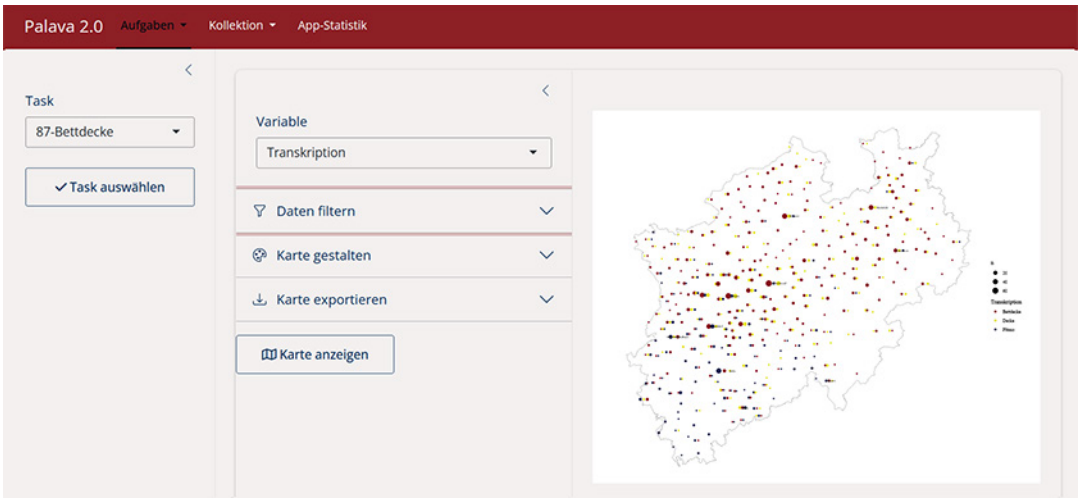

Abb. 5: Erstellung von Karten

Der wichtigste Schritt ist aber sicherlich die niedrigschwellige Kartierung. Hier gibt es verschiedene Möglichkeiten der Exploration. Es ist möglich, nur bestimmte Varianten anzuzeigen und die Daten nach den Sozialparametern zu filtern und so Muster klarer erkennen zu können. Darüber hinaus sind zwei unterschiedliche Kartentypen möglich. Erste Option ist eine Punktkarte, wobei die Größe des Punktes die relative oder absolute Häufigkeit einer Variante repräsentiert, die Farbe die jeweilige Variante (vgl. Abb. 5).

Zweite Option ist eine Flächenkarte, bei der entweder die politische Gemeindefläche oder ein Voronoi-Polygon eingefärbt wird (vgl. Abb. 6 und 7). Die Intensität der Färbung kodiert dabei die absolute oder relative Häufigkeit. Die Kartentypen heben unterschiedliche Faktoren hervor. Während die Punktkarten einerseits Verteilung von Varianten, aber auch Räume von großer Variation betonen, bieten die Flächenkarten schnelle Einblicke in zusammenhängende Verbreitungsräume. Entscheidender Vorteil dieser Lösung ist es, schnell und ohne Programmierkenntnisse einen Einblick in quantitative und räumliche Verteilung der Daten zu bekommen. In der Praxis ermöglicht dies einerseits die kollaborative Aufbereitung gemeinsam mit studentischen Mitarbeiter*innen sowie die schnelle Verfügbarkeit der Daten für Kooperationspartner*innen. Auch ist es möglich, die so erhobenen Daten in die universitäre Lehre einzubinden und so forschendes Lernen zu ermöglichen.

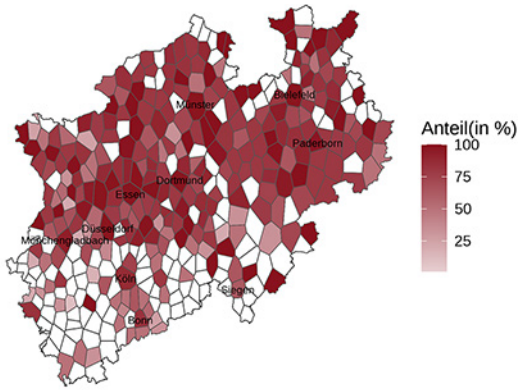

Abb. 6: Flächenkarte *Bettdecke*

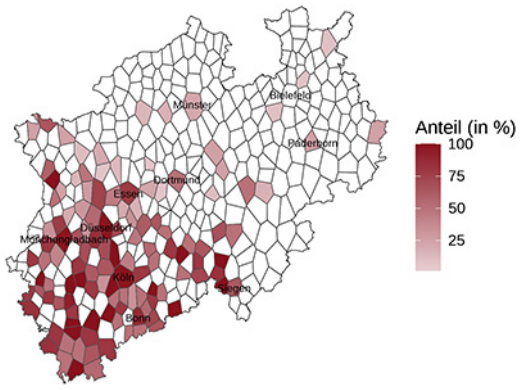

Abb. 7: Flächenkarte *Plümo*

3. Beispielauswertung: „Wie nennst du das Textil, womit man sich nachts zudeckt?“

Im Folgenden sollen anhand eines Beispiels die weiteren, inhaltlichen Auswertungsschritte demonstriert werden. Des Weiteren wird durch eine nach Alter differenzierte Betrachtung der Antworten gezeigt, welches Potenzial die PALAVA-Daten für die Erforschung rezenter Sprachwandelphänomene bietet.

In der ersten Fragerunde der PALAVA-App (Juni 2023–Oktober 2024) wurde den Teilnehmenden u. a. die Frage gestellt „Wie nennst du das Textil, womit man sich nachts zudeckt?“. Die Frage wurde schriftlich präsentiert, die Antworten durch die Teilnehmenden mündlich eingesprochen. Die Aufgabe zielt auf die Erhebung lexikalischer Variation ab, insbesondere soll überprüft werden, wie bekannt und verbreitet die Variante *Plümo* ist.[10] Das Wort wurde im 19. Jahrhundert aus dem Französischen in die rheinischen Dialekte entlehnt (frz. *plumeau* ‚Federbett‘) und wird von den Sprecher*innen oft als typisches Beispiel für französische Lehnwörter im Rheinland genannt, das auch noch in der regionalen Alltagssprache genutzt werde.[11] Der Auswertung liegen 2.479 Antworten zugrunde, die sich auf 24 Lexeme (lautliche Varianten wie *Bettzeuch* und *Bettzeuk* wurden zusammengezählt) verteilen.

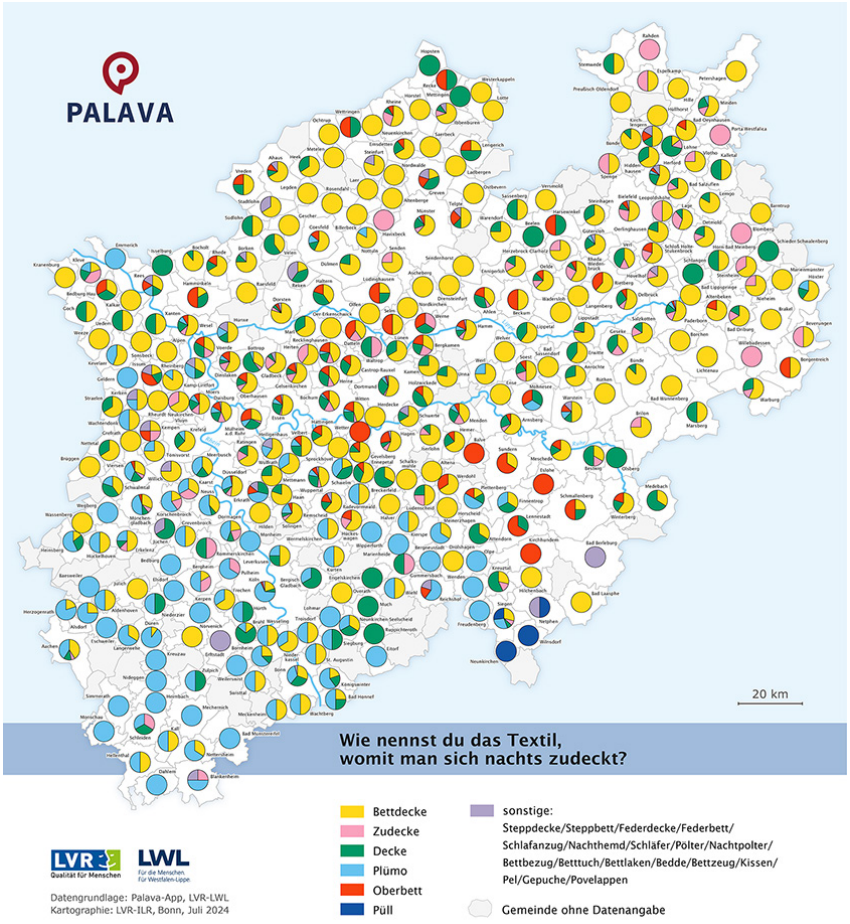

Abb. 8: PALAVA-Sprachkarte zur Verbreitung der verschiedenen Bezeichnungen für „Bettdecke“

10 Das Lexem kann sowohl mit kurzem als auch mit langem Vokal realisiert werden. Für diese Auswertung wurden die beiden Varianten nicht differenziert und in der Schreibung <Plümo> zusammengefasst.

11 Zum Einfluss des Französischen auf die rheinischen Dialekte vgl. Honnen (2018, S. 143–147).

Um die Sprachkarten auch für linguistische Laien gut lesbar zu gestalten, hat sich eine maximale Anzahl von zehn unterschiedlichen Varianten (d. h. Farben) als sinnvoll erwiesen. Die genannten Lexeme wurden daher zu sechs Gruppen zusammengefasst (*Bettdecke*, *Zudecke*, *Decke*, *Plümo*, *Oberbett*, *Püll*). Bei der Erstellung dieser Kategorien wurde zum einen darauf geachtet, dass die Varianten etymologisch zusammengehören, zum anderen auf eine vergleichbare Verteilung im Raum. Die Antworten, die sich diesen sechs Gruppen nicht zuordnen ließen und die jeweils weniger als fünfmal vorkamen, wurden unter „sonstige" zusammengefasst (vgl. Abb. 8). Die häufigste Antwort mit fast 50% ist *Bettdecke*.[12] Diese Variante ist, wie auch die verkürzte Form *Decke*[13], in ganz Nordrhein-Westfalen verbreitet, kommt im Rheinland allerdings weniger oft vor als in den anderen Regionen. Denn hier liegt mit dem Dialektwort *Plümo*[14] eine starke Konkurrenzform vor, die insbesondere im zentralen Rheinland und im Bergischen Land oft genannt wird, aber auch am Niederrhein vorkommt. Insbesondere im Ruhrgebiet zwischen Duisburg und Hamm sowie in Südwestfalen findet sich die Variante *Oberbett*[15]. Die *Zudecke*[16] wird im gesamten Erhebungsgebiet genannt, besonders häufig in Ostwestfalen. Ein kleinregional vorkommendes Lexem findet sich mit der Dialektvariante *Püll* im Siegerland.

Durch die große Altersspanne der Teilnehmenden bietet sich die Möglichkeit, die Daten im Sinne einer *apparent time*-Analyse getrennt nach Altersklassen zu betrachten (vgl. Abb. 9). Dabei zeigen sich deutliche Unterschiede. Die von den beiden älteren Gruppen noch häufig genannten Varianten *Oberbett* und *Zudecke* werden von den jüngeren Teilnehmenden kaum noch genannt. Auch die *Bettdecke* kommt von Altersklasse zu Altersklasse weniger oft vor. Diese Variantenreduktion erfolgt schrittweise zugunsten der verkürzten Form *Decke*, die im gesamten Erhebungsgebiet nun deutlich häufiger genannt wird. Die beiden dialektalen Varianten *Plümo* und *Püll* werden ebenfalls mit abnehmender Häufigkeit genannt, allerdings in geringerem Ausmaß als es bei *Bettdecke*, *Zudecke* und *Oberbett* der Fall ist.

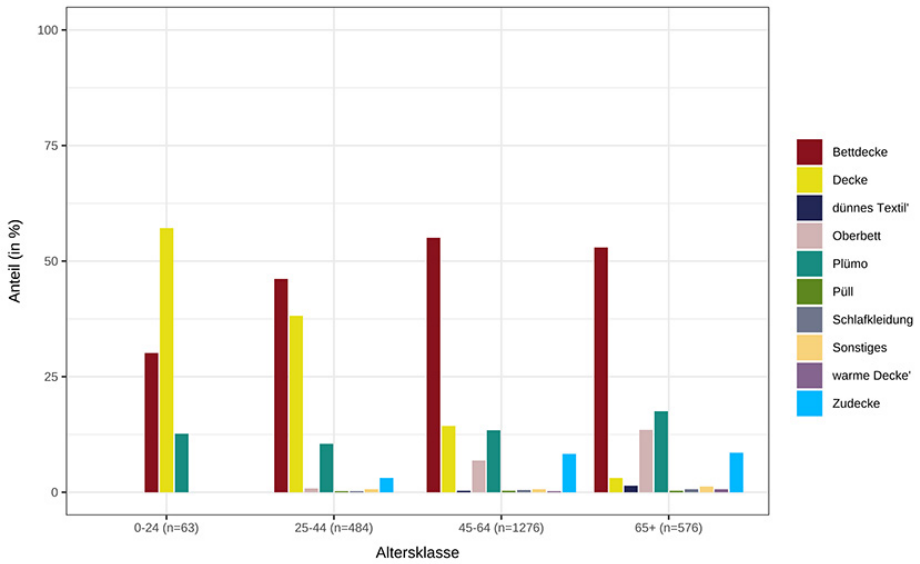

Abb. 9: Differenzierung der Antwort nach Alter

12 Zusammengefasst wurden hierunter die Varianten *Bettdecke*, *Bettdeck*, *Bettdeeg*.

13 Mit den Varianten *Decke*, *Deck*, *Degge*.

14 Mit den Varianten *Plümo* und *Plümmo*.

15 Mit den Varianten *Oberbett*, *Überbett*, *Övebedde*, *Deckbett*, *Zubett*.

16 Mit den Varianten *Zudecke* und *Zudeck*.

4. Fazit

Datenerhebungen via App sind – wie wir zeigen konnten – äußerst gewinnbringend und bringen vielfältige und zahlreiche Daten ein, bei relativ geringen Kosten und Personalaufwand. Die Auswertung benötigt allerdings technische Kenntnisse. Diese Einstiegshürde konnten wir durch die Erstellung einer eigenen Anwendung senken. So versprechen wir uns perspektivisch stärker auch Studierende und Externe sowie interessierte Bürger*innen in die Auswertung der Daten einzubinden. Gleichzeitig zeigt eine erste Analyse, dass uns die Daten neben synchroner Betrachtung der sprachlichen Variation in den Regionalsprachen Nordrhein-Westfalens auch Einblicke in rezente Entwicklungen erlauben.

Literatur

Atlas zur deutschen Alltagssprache (AdA) = Elspaß, Stephan/Möller, Robert (2003–): Atlas zur deutschen Alltagssprache (AdA). www.atlas-alltagssprache.de (Stand: 14.8.2025).

Chang, Winston/Cheng, Joe/Allaire, J. J./Sievert, Carson/Schloerke, Barret/Aden-Buie, Garrick/Xie, Yihui/Allen, Jeff/McPherson, Jonathan/Dipert, Alan/Borges, Barbara (2025): shiny: Web application framework for R. R package version 1.11.0. https://shiny.posit.co/ (Stand: 27.11.2025).

Elmentaler, Michael (2005): Die Rolle des überregionalen Sprachkontakts bei der Genese regionaler Umgangssprachen. In: Zeitschrift für Deutsche Philologie 124, 3, S. 395–415.

Elmentaler, Michael/Rosenberg, Peter (2015): Norddeutscher Sprachatlas (NOSA). Bd. 1: Regiolektale Sprachlagen. (= Deutsche Dialektgeographie 113.1). Hildesheim u. a.: Olms.

Entringer, Nathalie/Gilles, Peter/Martin, Sara/Purschke, Christoph (2021): Schnëssen. Surveying language dynamics in Luxembourgish with a mobile research app. In: Linguistics Vanguard 7, s1.

Elspaß, Stephan/Möller, Robert (2006): Internet-Exploration: Zu den Chancen, die eine Online-Erhebung regional gefärbter Alltagssprache bietet. In: Osnabrücker Beiträge zur Sprachtheorie 71: Dialekt im Wandel Perspektiven einer neuen Dialektologie, S. 143–158.

Freywald, Ulrike/Wiese, Heike/Boas, Hans C./Brizić, Katharina/Dammel, Antje/Elspaß, Stephan (2023): Deutsche Sprache der Gegenwart. Eine Einführung. Berlin: Metzler.

Gilles, Peter (2019): Using crowd-sourced data to analyse the ongoing merger of [ɕ] and [ʃ] in Luxembourgish. In: Calhoun, Sasha/Escudero, Paola/Tabain, Marija/Warren, Paul (Hg.): Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, 5–9 August 2019, Australia, S. 1590–1594. Canberra: Australasian Speech Science and Technology Association. https://assta.org/proceedings/ICPhS2019/papers/ICPhS_1639.pdf (Stand: 27.11.2025).

Gilles, Peter (2023): Regional variation, internal change and lanugage contact in Luxembourgish: results from an app-based language survey. In: Taal & Tongval 75, 1, S. 29–57.

Hasse, Anja/Bachmann, Sandro/Glaser, Elvira (2021): Gschmöis – Crowdsourcing grammatical data of Swiss German. In: Linguistics Vanguard 7, Heft s1.

Honnen, Peter (2018): Wo kommt dat her? Herkunftswörterbuch der Umgangssprache an Rhein und Ruhr. Köln: Greven.

Hove, Ingrid/Leemann, Adrian/Kolly, Marie-José/Dellwo, Volker/Goldmann, Jean-Philippe/Almajai, Ibrahim/Wanitsch, Daniel (2015): Voice Äpp: „My Voice – my dialect“. In: Leemann/Kolly/Schmid/Dellwo (Hg.), S. 287–300.

Kolly, Marie-José/Leemann, Adrian (2015): Dialäkt Äpp: Communicating dialectology to the public – crowdsourcing dialects from the public. In: Leemann/Kolly/Schmid/Dellwo (Hg.), S. 271–285.

Lanwer, Jens P. (2015): Regionale Alltagssprache. Theorie, Methodologie und Empirie einer gebrauchsbasierten Areallinguistik. (= Empirische Linguistik/Empirical Linguistics 4). Berlin/Boston: De Gruyter.

Leemann, Adrian (2017): Analyzing geospatial variation in articulation rate using crowd-sourced speed data. In: Journal of Linguistic Geography 4, 2, S. 76–96.

Leemann, Adrian/Kolly, Marie-José/Schmid, Stephan/Dellwo, Volker (Hg.) (2015): Trends in phonetics and phonology. Studies from german-speaking Europe. Berlin u.a: Lang.

Leemann, Adrian/Kolly, Marie-José/Purves, Ross/Britain, David/Glaser, Elvira (2016): Crowdsourcing Language Change with Smartphone Applications. In: PloS one 11, 1, S. 1–25. https://doi.org/10.1371/journal.pone.0143060.

Leemann, Adrian/Kolly, Marie-José/Werlen, Iwar/Britain, David/Studer-Joho, Dieter (2014): The diffusion of /l/-vocalization in Swiss German. In: Language Variation and Change 26, S. 191–218.

Radford, Alec/Kim, Jong W./Xu, Tao/Brockman, Greg/McLeavey, Christine/Sutskever, Ilya (2022): Robust speech recognition via large-scale weak supervision. DOI: 10.48550/ARXIV.2212.04356.

Rein, Charlotte (2020): Zurück nach Erp. Individueller und intergenerationeller Sprachwandel in einer ripuarischen Sprechergemeinschaft. (= Rheinisches Archiv 162). Wien/Köln/Weimar: Böhlau.

Rein, Charlotte/Schürmann, Timo (2024): Wer nutzt die PALAVA-App wie? In: Niederdeutsches Wort 64, S. 99–108.

Reips, Ulf-Dietrich (2002): Standards for internet-based experimenting. In: Experimental Psychology 49, S. 243–256.

Salewski, Kerstin (1998): Zur Homogenität des Substandards älterer Bergleute im Ruhrgebiet. (= Zeitschrift für Dialektologie und Linguistik – Beihefte 99). Stuttgart: Steiner.

Seiferheld, Maila (2023): Die PALAVA-App – Sprachdatensammlung mit dem Smartphone. In: Niederdeutsches Wort 63, S. 91–102.

Kontaktinformation

Dr. Charlotte Rein
LVR-Institut für Landeskunde und Regionalgeschichte
Endenicher Straße 133
53115 Bonn
E-Mail: Charlotte.Rein@lvr.de

Timo Schürmann
LWL-Kommission für Mundart- und Namenforschung
Schlossplatz 34
48143 Münster
E-Mail: Timo.Schuermann@lwl.org

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.72

Matthias Stemmler/Sonja Zeman

Annimate

Ein Tool zum Datenexport aus den historischen Referenzkorpora

Abstract *Annimate* (https://github.com/matthias-stemmler/annimate, Stand: 26.11.2025) was developed to facilitate the export of results from the reference corpora for the historical German language levels (e. g. ReA (750–1050), ReM (1050–1350), ReN (1200–1650) and ReF (1350–1650)), whose data is ANNIS-based. The tool provides a user interface that allows users to export search results from all these corpora directly, along with the desired additional information. This article demonstrates the functions of *Annimate* and its potential for investigating language change. As a case study, we present a corpus study on the diachrony of the interjection *ei/ey*. In addition, we demonstrate how corpora can be evaluated using *Annimate*. In conclusion, the program offers a solution to a variety of practical problems in synchronous and diachronic corpus linguistic studies, and can also contribute to the evaluation of existing corpus annotations.

Keywords reference corpora, data export, corpus linguistics, language change, interjections

1. Ausgangslage: Zum Wert der Referenzkorpora in der Germanistischen Linguistik

Für diachrone Untersuchungen von Sprachwandelphänomenen im Deutschen sind die Referenzkorpora für die historischen Sprachstufen zu einer grundlegenden Ressource geworden, um sprachhistorische wie diachrone Korpusstudien durchzuführen. Mit dem Referenzkorpus Altdeutsch (ReA; 750–1050), dem Referenzkorpus Mittelhochdeutsch (ReM; 1050–1350), dem Referenzkorpus Frühneuhochdeutsch (ReF; 1350–1650) und dem Referenzkorpus Mittelniederdeutsch/Niederrheinisch (ReN; 1200–1650) decken die Korpora sowohl eine große Zeitspanne als auch verschiedene Sprachräume ab.[1] Zeitlich ergänzt werden die Referenzkorpora inzwischen noch durch das syntaktisch annotierte GiesKaNe-Korpus (1650–1950),[2] das bislang 24 Texte enthält (GiesKaNe 0.3) und noch auf 72 Texte erweitert werden soll. Weitere Ressourcen für diachrone korpusbasierte Untersuchungen bilden registerspezifische Korpora zu bspw. Hexenverhörprotokollen, Predigten, wissenschaftlichen Texten etc. (vgl. die Übersicht unter https://corpus-tools.org/annis/corpora.html, Stand: 26.11.2025). Allen genannten Datenressourcen ist gemein, dass sie im ANNIS-Format vorliegen. ANNIS (‚ANNotation of Information Structure') ist ein open source sowie plattformübergreifendes Such- und Visualisierungstool (Krause/Zeldes 2016). Es wurde für komplexe, mehrschichtige linguistische Korpora mit verschiedenen Arten von Annotationen entwickelt und ermöglicht sowohl die Suche nach Daten durch ein web-

1 Zugang zu den Referenzkorpora über www.deutschdiachrondigital.de (Stand: 31.5.2025).

2 www.uni-giessen.de/de/fbz/fb05/germanistik/forschung/sprache/gieskane (Stand: 31.5.2025).

basiertes Such-Interface und die formale Abfragesprache ANNIS Query Language (AQL) als auch verschiedene Darstellungsmöglichkeiten wie beispielsweise in Tabellenansicht und als Graphen. Über ANNIS können damit in den Korpora komplexe Suchabfragen durchgeführt werden, die als empirische Datengrundlage für vielfältige linguistische Fragestellungen genutzt werden können.

Die Datenlage für die Untersuchung von Sprachwandelphänomenen im Deutschen hat sich mit der Erschließung der Referenzkorpora bedeutend verbessert. Der Bezug auf die Referenzkorpora hat Vorteile, die bereits von Wegera (1990) im Vorfeld der Entstehung des Referenzkorpus Mittelhochdeutsch (ReM) zusammengefasst wurden. Für dieses Korpus, das ca. 2 Mio. Wortformen umfasst, wurden Originalhandschriften zugrunde gelegt, da gerade für das Mittelhochdeutsche die häufig literaturwissenschaftlich orientierten Editionen für linguistische Untersuchungen nicht als verlässlich angesehen werden können. Zur Transparenz sind im Referenzkorpus die Texte sowohl als diplomatischer Lesetext, als modernisierter Lesetext als auch als normalisierter Lesetext hinterlegt. Zudem ist das Korpus weitgehend ausgewogen in Bezug auf die Faktoren Zeit, Raum und Textsorte zusammengestellt. Untersucht man ein bestimmtes Sprachwandelphänomen, lässt sich etwa nach bestimmten Zeitperioden und Schreiborten suchen bzw. filtern. Ebenso können die Ergebnisse zu den verschiedenen Textbereichen (Wissenschaft, Alltag, Religion, Poesie etc.), Textsorten (Arzneibuch, Minnesang, höfische Epik, Gebet etc.) und Textart (Vers, Prosa, Mischformen und Urkunde) zugeordnet werden. Zudem sind sie nach verschiedenen morphologischen und syntaktischen Parametern annotiert (vgl. für einen detaillierten Überblick Petran et al. 2016; Klein/Dipper 2016). Mit der Initiative „Deutsch Diachron Digital“ (www.deutschdiachrondigital.de/, Stand: 26.11.2025) besteht ein Verbund der Referenzkorpora für die verschiedenen Sprachstufen, der die technologische Kooperation und den Austausch in Bezug auf Nutzungsszenarien für Forschung und Lehre fördert.

Die Referenzkorpora bieten zweifellos große Vorteile bei diachronen Untersuchungen des Deutschen. Durch die Nutzung der Korpora für verschiedene sprachliche Phänomene ergibt sich zudem der Vorteil der Vergleichbarkeit der Ergebnisse. Gerade vor dem Hintergrund, dass die textsortenspezifischen Charakteristika einen großen Einfluss auf die Verwendung grammatischer Elemente haben (vgl. etwa zum Einfluss der Unterscheidung Prosa vs. Vers exemplarisch Prell 2005; zur Unterscheidung zwischen Nähe- vs. Distanzsprache Ágel/Hennig (Hg.) 2006; Elspaß 2005 und zum Einfluss regionaler Variation Elspaß 2010), ist die registerbezogene und dialektale Ausgewogenheit eine relevante Eigenschaft der Referenzkorpora. Einzeluntersuchungen, die sich auf nur eine Textsorte oder nur einen Text stützen, sind in ihrer Aussagekraft eingeschränkt, da erst zu untersuchen ist, ob die Ergebnisse auch für andere Texte gelten. Legen zwei Studien jeweils die gleiche Datengrundlage zugrunde, lassen sich die jeweiligen Ergebnisse zudem besser aufeinander beziehen.

Da die historischen Referenzkorpora aufgrund der jeweiligen Spezifika der zu annotierenden Sprachstufen in eigenständigen Projekten verwirklicht wurden, besteht in Bezug auf die Nutzung der Korpora allerdings das Problem, dass für jedes Korpus eigene Annotationsrichtlinien bestehen und die Termini nicht vereinheitlicht sind. Gerade mit Blick auf die Untersuchung diachroner Fragestellungen erschwert das die Nutzung der Korpora. Ein weiteres bekanntes arbeitspraktisches Problem stellt zudem der Export der Suchergebnisse dar. Zur Aufbereitung der Daten für eine anschließende qualitative Analyse ist es in der Regel wünschenswert, die Treffer sowohl in ihrem Kontext als auch mit den jeweiligen Annotationen (bspw. Lemma, Wortart, grammatische Flexion) zu extrahieren. Zusätzlich sollten auch die entsprechenden Metadaten zur Textquelle wie Dokumentname, Zeitraum, Sprachregion etc. erfasst werden. In ANNIS stehen dafür verschiedene Export-Funktionen

zur Verfügung, die dieses Ziel aber nicht ohne Weiteres ermöglichen. Der CSVExporter exportiert beispielsweise nur die gefundenen Treffer mit den jeweiligen Annotationen und ignoriert den Kontext. Eine einfache Suche nach dem Lexem „Himmel“ (AQL-Abfrage: `norm="himel"`) im ReM (Version 2.1) führt zu einer CSV-Liste, in der der Treffer (`1_span`) mit den jeweiligen Annotationen aber ohne Kontext ausgegeben wird, vgl. Abbildung 1:

1_id	1_span	1_anno_grammar::inflection	1_anno_grammar::inflectionClass	1_anno_grammar::lemma	1_anno_grammar::pos
salt:/11-12_1-obd-PV-X/M040#t262_m1	himil	Akk.Sg	st_Masc	himel	NA
salt:/11-12_1-obd-PV-X/M095#t158_m2	|hilmil	Dat.Sg	st_Masc	himel	NA
salt:/11-12_1-obd-PV-X/M137#t215_m1	himel	Nom.Sg	st_Masc	himel	NA

Abb. 1: Drei Suchergebnisse der mit dem CSVExporter exportierten Daten importiert in Excel (die Abb. zeigt aus Platzgründen nur eine Auswahl der Spalten an)

Der TextColumnExporter erlaubt den Export mit Kontext vor und nach dem Treffer. Diese Exportmöglichkeit gibt die Treffer aber ohne weitere Metadaten und Annotationen aus. Es können zwar noch „Annotation Keys“ eingegeben werden. Ohne weitere Kenntnisse über das jeweilige Korpus ist aber nicht offenkundig, welche Art von Annotation Keys eingegeben werden können. Eingaben wie „lemma“ oder „pos“ werden ignoriert. Bei „Parameters“ können Metadaten wie „title“ mit ausgegeben werden. Auch dafür ist es aber notwendig, die jeweiligen Parameter bereits zu kennen. Wird ein ungültiger Parameter (bspw. `meta-keys=doc`) eingegeben, führt der Export ohne weitere Fehlermeldung zu einer leeren Liste. Dabei ist zu berücksichtigen, dass die jeweiligen Annotationskategorien in den verschiedenen Referenzkorpora unterschiedlich sind und insofern für jede Sprachstufe separat die jeweiligen Handbücher und Annotations-Guidelines konsultiert werden müssen.

match_number	speaker	left_context	match_column	right_context
1	sText1	enperinnit,, unta alumbe $in ungewiteri michila .(.) er geladata den	himil	hin uf unta die erdan geunter$ceiden(.) [...] zerucke .(.) oba
2	sText1	menin$che ,, vn\- er an de\-me uierzgo$te\-me tage . zi|	|hilmil	vóir(:) dannan gelo\vbi ich ín chunftich z| |ir# teilin ubir
3	sText1	*(S*)c\-a *(M*)aRia <.>(.) *(D*)o gehit ime $o werde . der	himel	zu\o der erde ,, da der e$il unde daz rint

Abb. 2: Drei Suchergebnisse der mit dem TextColumnExporter exportierten Daten importiert in Excel (alle Spalten)

In Abbildung 2 zeigt sich als weiteres Problem, dass der Text verschiedene Sonderzeichen enthält (vgl. *an de\-me uierzgo$te\-me tage* statt *an dēme uierzgoſtēme tage*). Das liegt daran, dass sowohl der CSVExporter als auch der TextColumnExporter nicht auf die Ebene der handschriftlichen Tokenisierung (tok_dipl) oder der den modernen Erwartungen entsprechenden Tokenisierung (tok_anno) zugreifen, sondern auf eine dritte Default-Ebene. Dieses Problem lässt sich mit dem GridExporter lösen, bei dem die entsprechende Ebene (tok_dipl bzw. tok_anno) explizit im Feld „Annotation Keys“ angegeben werden kann. In den Daten ist dann aber nicht mehr der eigentliche Treffer ersichtlich. Das bedeutet, dass sich für den jeweiligen Treffer nicht gezielt weitere Annotationen ausgeben lassen, vgl. Abbildung 3, die die Ausgabe mit dem GridExporter zeigt.

1	tok_dipl	enperinnit unta alumbe ſin ungewiteri michila . er geladata den himil hin uf unta die erdan geunterſceiden [...] zerucke . oba
2	tok_dipl	meninſche . vñ er an dēme uierzgoſtēme tage . zihilmil vóir dannan gelǒbi ich ín chunftich zir teilin ubir
3	tok_dipl	Sča MaRia . Do gehit ime ſo werde . der himel zů der erde . da der eſil unde daz rint

Abb. 3: Drei Suchergebnisse der mit dem GridExporter exportierten Daten importiert in Excel (alle Spalten)

Die dargestellten Probleme wurden auch bereits in Hartmann (2022) zusammengestellt und adressiert. In einem umfangreichen Tutorial wird dort ein mehrschrittiges Verfahren unter der Verwendung mehrerer Exportfunktionen und einem Excel-Makro vorgeschlagen. Der Datenexport ist so aber nur über Umwege möglich. Hinzu kommt, dass für den

Export der Metadaten die entsprechenden Bezeichnungen in den Korpora unterschiedlich sind und für die jeweiligen Abfragen erst eruiert werden müssen.

2. Annimate

Das von Matthias Stemmler in Kooperation mit dem Lehrstuhl für Deutsche Sprachwissenschaft der Universität Augsburg entwickelte Tool *Annimate* (URL: https://github.com/matthias-stemmler/annimate, Stand: 26.11.2025) begegnet diesen Problemen, indem es eine komfortable Benutzer-Oberfläche bereitstellt, über die Daten zu Suchergebnissen in allen ANNIS-Korpora direkt und mit den jeweils gewünschten Zusatzinformationen exportiert werden können. Auf der Basis der Suchabfrage erstellt das Desktop-Programm wahlweise eine CSV- oder Excel-Datei mit einer Zeile pro Beleg, die die jeweiligen Treffer (bzw. Match-Knoten) in ihrem Kontext im KWIC-Format (KeyWord In Context) sowie zusätzliche Annotationen und Metadaten auf Teilkorpus- und Dokumentebene anzeigt. Die Funktionalität ähnelt einer Kombination aus dem TextColumnExporter und dem CSVExporter in ANNIS, bietet jedoch eine benutzerfreundlichere Oberfläche (vgl. den Screenshot in Abb. 4).

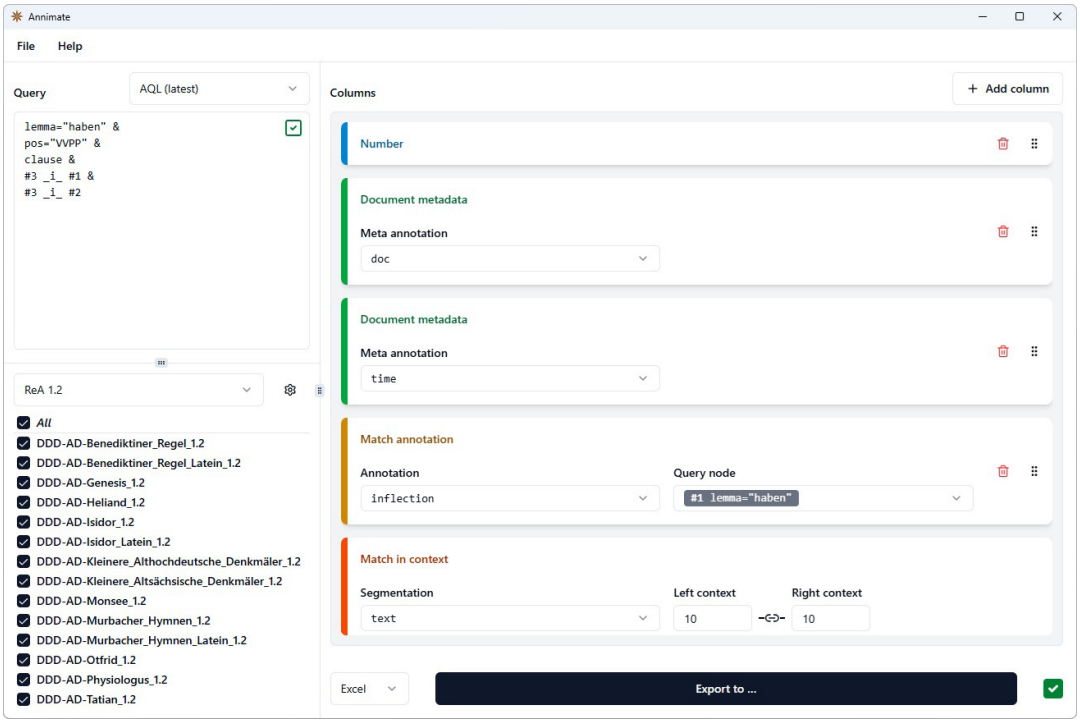

Abb. 4: Screenshot: Benutzeroberfläche mit einer Beispielabfrage in *Annimate* für das Referenzkorpus Altdeutsch

Im Rahmen dieses Artikels ist es nicht möglich, die gesamte Funktionalität des frei verfügbaren Programms zu erläutern. Wir verweisen diesbezüglich auf das ausführliche und detaillierte Benutzerhandbuch (https://matthias-stemmler.github.io/annimate/user-guide/, Stand: 26.11.2025). Im Folgenden kann nur die allgemeine Funktionsweise erklärt und eine Übersicht der wichtigsten Features gegeben werden.

Um das Programm nutzen zu können, muss *Annimate* zunächst über die GitHub-Seite heruntergeladen und installiert werden. Hierfür stehen Optionen für Linux, Windows und Mac zur Verfügung. Vor der ersten Verwendung müssen zudem die gewünschten Daten aus den ANNIS-basierten Korpora in das Programm importiert werden. Um etwa Abfragen

im Referenzkorpus Altdeutsch durchführen zu können, muss das Korpus in seiner aktuellen Version zunächst heruntergeladen werden, bspw. über die Webseite www.laudatio-repository.org/ (Stand: 26.11.2025), auf der verschiedene annotierte Korpora zur Verfügung gestellt sind. Daten im Format graphANNIS oder relANNIS können direkt in *Annimate* importiert werden. Dieser Schritt muss nur einmalig durchgeführt werden. Bei erneuten Korpusabfragen kann das Korpus direkt in *Annimate* aufgerufen werden. Mit *Annimate* können Korpusabfragen daher auch ohne Internetzugang durchgeführt werden. Es sollte aber gelegentlich überprüft werden, ob das Korpus nach einem eventuellen Release in einer aktuelleren Version zur Verfügung steht. Um mit verschiedenen Korpora, wie beispielsweise dem ReM und dem ReA, in derselben *Annimate*-Installation arbeiten zu können, können deren Teilkorpora entsprechend gruppiert werden. Damit kann über die gleiche Benutzeroberfläche auf die verschiedenen Korpora zugegriffen werden. Die Suche kann dann mit der gleichen Abfragesprache wie auf der Benutzeroberfläche der Referenzkorpora durchgeführt werden. Arbeitstechnisch ist es dabei von Vorteil, die Suchabfrage zunächst auf der Webseite des entsprechenden Korpus durchzuführen und *Annimate* mit der entsprechenden Suchabfrage für den Export zu verwenden. Das stellt durch den Vergleich der Zahl der gefundenen Treffer auch sicher, dass die jeweils aktuelle Version des jeweiligen Korpus importiert ist.

Für den Export mit *Annimate* wird das gewünschte Korpus ausgewählt, wobei es auch möglich ist, Teilkorpora (im Fall des ReA beispielsweise nur das Teilkorpus Tatian) zu durchsuchen. Analog zur webbasierten Benutzeroberfläche der Referenzkorpora wird die Suchanfrage in einem Feld links oben eingegeben, deren Gültigkeit durch das Programm geprüft und angezeigt wird. Bei einer gültigen Abfrage wird ein grünes Häkchen angezeigt, bei einer ungültigen Abfrage ein rotes Kreuz und eine Beschreibung des Fehlers. Auf der rechten Seite der Benutzeroberfläche können dann die jeweiligen Informationen zusammengestellt werden, die zu den Suchergebnissen ausgegeben werden sollen. Die Informationen sind in die folgenden Kategorien unterteilt, vgl. Tabelle 1.

Number	Zuweisung einer Referenznummer für jeden Treffer
Match in context	KWIC/Anpassung der Kontextlänge (= Anzahl der Tokens, die links bzw. rechts des Treffers ausgegeben werden sollen). Bei komplexeren Suchabfragen muss hier noch angegeben werden, für welchen Treffer der Kontext ausgegeben werden soll.
Match annotation	Information aus den Annotationen. Beispielsweise können morphologisch-syntaktische Merkmale oder das jeweilige Lemma ausgegeben werden.
Document metadata	Metadaten zum Dokument (bspw. Zeitraum, Region, etc.)
Corpus metadata	Metadaten zum Teilkorpus

Tab. 1: Übersicht über die jeweiligen Kategorien, zu denen über *Annimate* Informationen zu den jeweiligen Treffern ausgewählt werden können

Ein Vorteil des Programms ist es, dass insbesondere für die Kategorien „Match annotation“ und „Document metadata“ aus einem Vorauswahl-Menü der in ANNIS suchbaren (Meta-) Annotationskategorien ausgewählt werden kann. Bei den morphologischen Annotationen muss demnach nicht zuerst auf der Basis der Annotationen des jeweiligen Korpus und des

Benutzerhandbuchs die jeweils annotierte Kategorie (bspw. `lemma`, `inflection`, `language` etc.) eruiert werden, sondern diese kann direkt aus dem Auswahl-Menü angeklickt werden. Ab der neueren Version 1.5.0 von *Annimate* wird neben dem Exportbutton eine Information darüber angezeigt, ob die Angaben für die Parameterauswahl vollständig sind bzw. welche Angaben fehlen.

Mit Ausnahme der Kategorie „Match in context“ erstellt *Annimate* für jede der ausgewählten Kategorien eine Spalte in der Ausgabedatei. Für „Match in context“ werden für jeden Treffer zusätzlich zwei Spalten (eine für den linken, eine für den rechten Kontext) ausgegeben. Nach Angabe der benötigten Informationen können die Ergebnisse der Suchabfrage mit den erwünschten Informationen wahlweise im CSV- oder Excel-Format exportiert werden. Neben den Suchergebnissen (vgl. Abb. 5) werden auf einem zweiten Excel-Sheet die Informationen zur Abfrage dokumentiert.

	A	B	C	D	E	F	G	H
1	Number	Document	Document time	H1 inflection	Context 1 (text)	Match 1 (text)	Context 2 (text)	Match 2 (text)
2	1	AGB_AlemGlaubenundBeichte	12-13	IND_PRES_SG_1	minen werchen : den lon den fiurht ich ser , wan ich diche	gesundot		han
3	2	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	SUBJ_PRES_SG_3	an disiu wort denchet , wie vorhtlich dei sîn . der sich	versumit		habe
4	3	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	SUBJ_PRES_SG_3	. der sich versumit habe ennenhär durch sine tracheit , daz er sin niut	giernet		habe
5	4	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	IND_PRES_PL_2	, er chome ze râwe . EXORTATIO AD CONFESSIONEM . Nu	habet	ir iuch	gevestenet
6	5	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	IND_PRES_PL_2	erchennen alles iwers unrehts . iuch sol vil harte riwen swaz ir wider gotis hulden	habet		getân
7	6	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	IND_PRES_PL_3	sele wisen in die gnade unde in die vroude , die si selbe	besezzen		habent
8	7	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	IND_PRES_PL_2	. Nach so getaner bihte unde nach dem geheizze , den ir unserm herren got	geheizzen		habet
9	8	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	IND_PRES_PL_3	. Vil guoten liute , so getaniu bihte hilfet einigenote die ir bihte tougliche	habent		getan
10	9	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	IND_PRES_PL_3	bihte tougliche habent getan unde die ouch tougeliche suntint . die auer offenlich	habent		gesuntit
11	10	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	IND_PRES_PL_3	manslahte , uberhuor , sippehuor . swelhe die sint , die houpthafte sunte	habent		getân
12	11	BGB3_BenediktbeurerGlaubenundBeichteIII	12-13	IND_PRES_PL_3	manslahte , uberhuor , sippehuor . swelhe die sint , die houpthafte sunte	habent	getân unde noch der christenheit niht ze wizzene sint	getan
13	12	BR1_BaslerRezept1	8	SUBJ_PRES_SG_3	ni neouuiht ni uuirce , nipu z de gisehe , de imo daz tranc gebe enti simplum	piuuartan		habe
14	13	BambGB1_Bamberger_Glaube_und_Beichte	12	SUBJ_PRES_SG_3	aller menniſglich uúre ſich ſelban göte réda gében ſól , ſuie s er	gilébet		habe
15	14	DD_DeDefinitione	11.2	IND_PRES_SG_2	animal rationale et mortale . Taz ist imo gemacze , mit tiu	habist	tu in	genoman
16	15	GGB1_SangallerBeichte_I	11-12	IND_PRES_SG_1	unde fone sancte Petre unde fone allen gotes heiligon , so filo hich keuualdes	háben		anfangen
17	16	GGB3_SangallerGlaubenBeichteIII	11-12	IND_PAST_PL_3	begrabin wart unde er uon der helle lóiste alle , die sinin willin	háton		gitan
18	17	GGB3_SangallerGlaubenBeichteIII	11-12	IND_PRES_SG_1	nach minin werchin . Den lon vurht ich sere , wand ich	gesundot		han
19	18	MM_MementoMori	11-12	IND_PRES_PL_3	leil , si ne dúhté sie nie sô minnesam , si	hebent	si ie duh	verlâzen
20	19	MM_MementoMori	11-12	IND_PRES_PL_3	sie des gedâhton , war sie ze iungest varn solton ! nû	habint	siu iz	bevunden
21	20	MM_MementoMori	11-12	IND_PRES_PL_2	ze wesinne , taz ir wârint als ein man : taz	hânt	ir uber	gangen
22	21	MM_MementoMori	11-12	SUBJ_PAST_PL_2	als ein man : taz hânt ir uber gangen ,	habetint	ir anders niewit	getan
23	22	MM_MementoMori	11-12	IND_PRES_SG_3	die hella ; dâ muoz er iemer inne wesen : got selben	hât	er hin	gegeben
24	23	MM_MementoMori	11-12	IND_PRES_SG_3	tes ne mag imo der scaz ze guote werden .	Habit	er sinin richtuom sô	geleit
25	24	MM_MementoMori	11-12	IND_PRES_SG_3	er hie verleibet , taz wirt imo ubilo geteilit .	habit	er iet hina	gegebin
26	25	MM_MementoMori	11-12	IND_PRES_SG_3	er wol mac : hie noh chumit der tac :	habit	er is tenne niwit	getân
27	26	MM_MementoMori	11-12	IND_PRES_SG_2	vil ubelor mundus , wic betrugist tû uns sus ! dû	habist	uns	gerichin
28	27	MM_MementoMori	11-12	IND_PRES_SG_3	unde lîb , alsô lango sô wir hie lebin , got	habit	uns selbwala	gegibin
29	28	M_Muspilli	9.2	IND_PAST_SG_3	Dar scal er vora demo rihhe az rahhu stantan , pi daz er in uuerolti eo	kiuuerkot		hapeta
30	29	M_Muspilli	9.2	IND_PRES_SG_3	enti mit fastun dio virina kipuazti . Denne der paldet der	gipuazzit		hapet

Abb. 5: Screenshot: Beispiel für die Ausgabe der Treffer und Zusatzinformationen im Excel-Format

Als weitere arbeitspraktische Funktion ermöglicht *Annimate* zudem die Speicherung der Suchabfragen in Projekten. Auf diese Weise kann die Suchabfrage mit der entsprechenden Konfiguration zu einem späteren Zeitpunkt wiederholt bzw. modifiziert durchgeführt werden. Auch macht es diese Funktion möglich, die Suchabfragen innerhalb einer Arbeitsgruppe zu teilen und zu reproduzieren.

Insgesamt ist *Annimate* mit seinen verschiedenen Funktionen – Überprüfung der Suchabfrage mit Fehleranzeige, das personalisierte Zusammenstellen und Speichern von Teilkorpora, die automatische Dokumentation der Suchabfrage, die Ausgabe der Belege mit variabel anpassbarer Kontext-Länge, ein Vorauswahl-Menü der in ANNIS suchbaren (Meta-) Annotationskategorien und die Speicherung der Such-Konfiguration – ein Programm, das den Export von Ergebnissen aus ANNIS-basierten Daten entscheidend vereinfacht. Im folgenden Abschnitt soll das an einer kleinen diachronen Studie zur Verwendung der Interjektion *ei*/*ey* im Deutschen demonstriert werden.

3. Fallstudie

Das Potenzial des Programms demonstrieren wir im Folgenden anhand einer kleinen Korpusstudie zur Diachronie der Interjektion *ei*/*ey*. Als Merkmal der gesprochenen Sprache und als Subjektivitätsmarker, die evaluative oder emotive Äußerungen einer Sprechinstanz indizieren, sind Interjektionen allgemein ein interessanter Untersuchungsgegenstand zur Erforschung historischer Nähesprachlichkeit (vgl. im Detail Zeman im Ersch.). Wir greifen hier exemplarisch die Interjektion *ei*/*ey* heraus, da diesbezüglich in der Forschungsliteratur bereits einige Beobachtungen zu ihrer diachronen Entwicklung vorliegen (vgl. Kilian 2012; Müller 2022; Imo/Müller 2023), die an den historischen Korpora überprüft werden können.

ei/*ey* ist laut dem Grimmschen Wörterbuch in althochdeutscher Zeit nicht belegt, kommt aber ab dem Mhd. häufig vor[3] und dient dort dem Mittelhochdeutschen Handwörterbuch von Matthias Lexer zufolge als Ausruf der Verwunderung, Freude und Klage.[4] Das Grimmsche Wörterbuch attestiert zudem die Bedeutungen des Staunens und des Liebkosens. In Bezug auf ihren Gebrauch ist bekannt, dass die Interjektion häufig mit dem Vokativ und dem Imperativ (*ei mutter seht doch her!*) sowie in Fragen (*ei was?*), Ausrufen mit *w*-Phrasen (*ei wie gern!*), Wunschsätzen und Verwünschungen verwendet wird.[5] Spätestens im 18. Jahrhundert zeichnet sich dann ein Sprachwandel ab, insofern als die Interjektion Kilian (2012) zufolge in diesem Zeitraum als gesprächsschritteinleitende Partikel (*Ei, was redst du doch*) verwendet wird, die häufig „ablehnende bzw. widersprechende Funktion des Gesprächsschritts ankündigt" (Kilian 2012, S. 117).[6]

Mit dem Fokus auf dem Tool *Annimate* kann es in der Kürze des vorliegenden Artikels nicht das Ziel sein, den Sprachwandel der Interjektion in allen Details nachzuzeichnen (vgl. dazu insbesondere die umfangreiche Studie von Imo/Müller 2023). Im Folgenden geht es dagegen um die Frage, ob die Interjektion auch in den historischen Sprachstufen als „gesprächsschritteinleitende" bzw. rede-einleitende Partikel verwendet wird.

Zu diesem Zweck vergleichen wir die Vorkommen von *ei*/*ey* im mittelhochdeutschen Referenzkorpus (ReM, 1050–1350) und dem GiesKaNe-Korpus (1650–1950). Die Belege im Mittelhochdeutschen Referenzkorpus elizitieren wir mit *Annimate* über die Suchabfrage `pos="ITJ"`. Für unsere Fragestellung hätte es ausgereicht, nach dem entsprechenden Lemma zu suchen. Uns ist es aber wichtig, *ei*/*ey* im Vergleich zu anderen Interjektionen einschätzen zu können. Daher präferieren wir die Suche nach allen Belegen, die als Interjektionen annotiert sind.

Dazu wird die Suchabfrage `pos="ITJ"` auf der Benutzeroberfläche in *Annimate* eingegeben und das ReM-Korpus in der Version 2.1 ausgewählt, das wir bereits vorher importiert haben. Wir klicken auf „All", um alle Teilkorpora des ReM bei der Suche zu berücksichtigen. Die Suchabfrage wird dann automatisch überprüft. Wird sie als gültig angezeigt, dann lassen sich die entsprechenden Parameter auswählen, die am Ende in der ausgegebenen Export-Liste erscheinen sollen. In diesem Fall wählen wir die Parameter mit folgenden Werten, vgl. Tabelle 2:

3 „ei", Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/25. www.woerterbuchnetz.de/DWB?lemid=E00867, (Stand: 31.5.2025).

4 „ei, interj.", Mittelhochdeutsches Handwörterbuch von Matthias Lexer, digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/25. www.woerterbuchnetz.de/Lexer?lemid=E00323, (Stand: 31.5.2025).

5 „ei", Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/25. www.woerterbuchnetz.de/DWB?lemid=E00867, (Stand: 31.5.2025).

6 Die häufig als Merkmal der Jugendsprache angeführte Partikel *ey* teilt die Funktionen des attention gettings, der Redeeinleitung und der emotionalen Färbung der Anrede (vgl. die Funktionsbeschreibung in Bahlo et al. 2019, S. 62). Sie geht dem Duden zufolge jedoch auf eine englische Entlehnung zurück.

Number	
Document metadata	Meta annotation = `doc`
Document metadata	Meta annotation = `time`
Document metadata	Meta annotation = `language-type`
Match in context	Segmentation = tok_anno; Left/Right context = 20
Match annotation	Annotation = norm; Query node = `#1 pos="ITJ"`

Tab. 2: Übersicht über die ausgewählten Werte in *Annimate*

Die Ergebnisse der Suchabfrage lassen wir uns als Excel-Liste ausgeben. Von den gefundenen Interjektionen beschränken wir unsere Analyse dann auf diejenigen Interjektionen, die in der „match"-Spalte als „ei", „ay" oder „ey" aufgeführt sind, sowie diejenigen, die als norm=„ei" annotiert sind. Nach einer zusätzlichen manuellen Überprüfung ergeben sich insgesamt 200 Treffer.[7]

Für die qualitative Klassifikation ergänzen wir die Excel-Liste noch um folgende Spalten, in denen wir manuell die Werte für folgende Parameter annotieren.

Verwendung	+ Vokativ, + Imperativ, + *w*-Phrase, sonstige
Rede-Einleitung	RE (rede-initial) – NRE (nicht rede-initial)

Tab. 3: Übersicht über die Annotations-Parameter, die manuell klassifiziert werden

Die quantitative Auswertung der Belege offenbart in Übereinstimmung mit der Beschreibung im Grimmschen Wörterbuch folgende Verwendungsweisen.

Die häufigste Verwendung (53,5%) sind Exklamationen mit anschließendem Vokativ, dem teilweise noch ein Imperativ folgt. Der Kontext ist durch sprecher-orientierte Verweise und Bezüge auf die erste und zweite Person charakterisiert, vgl. (1).

(1) ***ey selig wip*** *des* ***volgit*** *mime rade* [M331 – Hessische Reimpredigten, 14. Jh.]
‚Ei, selige Frau, folgt meinem Rat'

Eine weitere häufige Verwendung (26%) sind Exklamationen im Kontext von *w*-Phrasen.

(2) ***Ei wie*** *gerne weres tu gestorven. […]* ***ei wie*** *gerne heddes tu din leven. vmbe dines svones leven gegeven.* [M335 – Rheinisches Marienlob, 13. Jh.]
‚Ei wie gerne wärest du gestorben. […] Ei wie gerne hättest du dein Leben um deines Sohnes Leben gegeben.'

Unter die Rubrik „sonstige" fallen Verwendungen in hypothetischen Sätzen (*ei wüßte ich…*), in Wunschsätzen (*ei, möge es ihm wohl ergehen*), in Konditionalsätzen und weiteren Satzarten.

Mit wenigen Ausnahmen steht *ey* satz-initial, aber nicht unbedingt rede-initial. Das zeigt sich in folgendem Beleg aus einer längeren direkten Rede. *Ey* markiert nicht den Beginn

7 Neben *ei*/*ey* findet sich im ReM etwa auch die Interjektion *eia*/*eya*, die aber funktional nicht ganz synonym zu sein scheint und die wir daher nicht berücksichtigen. Auch das fragliche Verhältnis zwischen *ei*/*ey* und *hei*/*hey* (Imo/Müller 2023) blenden wir im begrenzten Umfang dieses Beitrags aus.

der direkten Rede. Vielmehr ist die Verwendung deutlich durch die Exklamation in Kombination mit dem Vokativ *herre* motiviert.

(3) *Nv han ich disem lieben man Ni kein vngemach getan war vm zvrnet er wid‘ mich* ***ey herre*** *weder bin ich Im zv iunc od‘ zv alt od‘ bin ich als vngestalt daz er min nicht zv wiebe wil* [M311 – Heinrich von Freiberg: Tristan, 14. Jh.]
‚Nun habe ich diesem lieben Mann niemals ein Ungemach angetan. Warum zürnt er gegen mich? Ei Herr, weder bin ich ihm zu jung oder zu alt oder bin ich so ungestalt, dass er mich nicht zur Frau nehmen will.‘

Die Setzung von *ey* kann allerdings auch mit einem neuen Gedanken verknüpft sein. Es zeichnet sich damit bereits das Potenzial der funktionalen Verschiebung zu einem „redeschritteinleitenden“ Element ab. Gerade in narrativen Texten wird *ey* als rede-einleitendes Element nach Verba dicendi verwendet (4) oder als Einleitung eines neuen Erzählabschnitts durch den Erzähler (5).

(4) *Zv dem ritter sprach er do* ***Ey*** *dv vngetruwer man Waz has tu leides mir getan* [M326 – Passional, 14. Jh.]
‚Zu dem Ritter sprach er da: „Ei, du untreuer Mann. Was hast du mir Leid angetan.“‘

(5) ***ey*** *waz tut nv her tristan da kerte er ab‘ drate*
[M311 – Heinrich von Freiberg: Tristan, 14. Jh.]
‚Ei, was tut nun Herr Tristan? Da kehrte er schnell um…‘

Die Verwendungen in der Funktion als Rede-Einleitung machen insgesamt aber nur einen geringen Anteil von 16% aus (32/200).

***Ei/ey* im ReM**	200	davon rede-einleitend: 32/16%
Vokativ (teils mit Imp.)	107	53,5%
Imperativ (ohne Vok.)	12	6%
***w*-Phrasen**	54	27%
Andere	27	13,5%

Tab. 4: Übersicht über die Funktionen der Interjektion *ei/ey* im ReM

Im Vergleich zum ReM ergibt sich im GiesKaNe-Korpus[8] eine andere funktionale Verteilung für die Interjektion *ey*. Um zunächst wieder die entsprechenden Belege zu elizitieren, formulieren wir die entsprechende Suchabfrage `wa="itj"`.[9] Die Suche liefert insgesamt 193 Treffer. Aufgrund von fünf Belegen mit der doppelten Interjektion *ey ey*, die wir jeweils als einen Treffer werten, ergibt sich eine Gesamtanzahl von 188 Belegen. Quantitativ ergibt sich, dass *ey* bzw. *ey ey* mit 21 von 188 Belegen neben *ach* (63/188) und *o/oh* (54/188) zu den drei frequentesten Interjektionen im Korpus zählt. Dieses Ergebnis korreliert mit der Verteilung der Interjektionen in den Dramen von Andreas Gryphius, wie sie von Müller (2022) untersucht wurden. Auch hier sind *ach*, *ey*, und *o* mit Abstand die frequentesten Interjektionen. Das deutet darauf hin, dass die literarischen Texte der Zeit durchaus den Gebrauch der Interjektionen in den Gebrauchstexten widerspiegeln. Zu berücksichtigen ist, dass die Treffer-Anzahl von 21 *ei/ey*-Belegen sowohl im Vergleich zum ReM (n=200) als auch zur

8 Wir danken Mathilde Hennig und Volker Emmrich für die Bereitstellung der Korpusdaten im ANNIS-Format.

9 Zur Vergleichbarkeit der Daten mit dem ReM präferieren wir auch hier eine Suchabfrage nach allen Vorkommen der Interjektionen.

Studie von Müller (2022) (n=107) gering ist, was einerseits an der Auswahl der Textsorten und berücksichtigten Regionen liegen kann, andererseits an der Größe des noch weiter im Aufbau befindlichen Korpus. Das GiesKaNe-Korpus umfasst zum derzeitigen Stand (Juli 2025) ca. 300.000 Tokens, das ReM dagegen 2 Mio. Wortformen.

Die qualitative Klassifikation der Belege zeigt, dass *ey* bzw. *ey ey* zu 90% in der direkten Rede verwendet wird (19 von 21 Belegen). Die meisten Belege (17/21) kommen im Text Nehrlich[10] vor, der sich durch einen hohen Anteil an direkter Rede (23%, vgl. Zeman im Ersch.) auszeichnet. *Ey* hat in diesem Text eine appellative, aber auch eine diskursstrukturierende Funktion, indem es als Rede-Einleitung verwendet wird, vgl. (6). Diese Funktion ist besonders in dialogischen Passagen mit Sprecherwechseln zu erkennen, in denen *ey* auch als Signal für ‚turn-taking' verwendet wird, vgl. (7) und (8), in denen *ey* jeweils einen Sprecherwechsel anzeigt.

(6) […] *und fing an mit bewegligen wortten* ***Eÿ eÿ****, du lieber Hanß Ludwig was machstu mit dem buch* [Nehrlich, 18. Jh.]

(7) *und eröffnete mir: Meine zwey Geissen seyen auf und da von. „****Ey*** *so fahre denn alles hin"! sagt' ich, „wenn's so seyn muß"* [Nehrlich, 18. Jh.]

(8) *ich sagte Herr Suprint ich weiß nicht, was Cloßen sind, sagen sie mir doch teutsch,* ***eÿ*** *ob ihr s wist oder nicht das hilft euch nichts.* [Nehrlich, 18. Jh.]

Tabelle 5 zeigt die unterschiedliche funktionale Verteilung.

***Ei/ey* in GiesKaNe**	21		davon rede-einleitend: 19/90%
Vokativ	9	43%	
Imperativ	1	5%	
***w*-Fragen**	4	19%	
Andere	7	33%	

Tab. 5: Übersicht über die Funktionen der Interjektion *ei/ey* im GiesKaNe-Korpus

Die geringe Zahl von 21 Belegen im Gesamtkorpus ist zwar nur eingeschränkt aussagekräftig. Auch ist die Funktion der Rede-Einleitung hauptsächlich in einem Text der GiesKaNe belegt. Allgemeine Aussagen zum Sprachwandel der Form können daher kaum auf der vorliegenden Datenbasis getroffen werden. Vor dem Hintergrund bereits vorliegender Studien (vgl. Kilian 2012; Müller 2022; Imo/Müller 2023) zeigt die Verteilung aber durchaus ein deutliches Muster. Die häufige Verwendung im Nehrlich-Text deutet darauf hin, dass die Funktion der Partikel als Rede-Einleitung etabliert ist: In 19 von 21 Belegen wird *ey* bzw. *ey ey* in dieser Funktion verwendet.[11]

Weitere empirische Untersuchungen könnten die Hypothese eines Funktionswandels von der emotiven Funktion zu einem rede-einleitenden Element weiter überprüfen. Das Muster der Verwendung im Mittelhochdeutschen deutet zudem darauf hin, wie der Funktions-

10 Nehrlich, Hans Ludwig. 1723. Erlebnisse eines frommen Handwerkers im späten 17. Jahrhundert. Herausgegeben von Rainer Lächele. Tübingen: Niemeyer 1997.

11 Dieser Befund passt auch zur Studie von Müller (2022, S. 300), die *ey* eine „sehr eindeutige Komödienspezifik" attestiert. Aufgrund ihres Bezugs zur dialogischen Kommunikation wäre es naheliegend, dass *ey* in den Dramen von Gryphius als Indikator gesprochener Sprache verwendet wird (vgl. auch Zeman im Ersch.).

wandel motiviert sein könnte: Die enge Verbindung mit dem Vokativ bedingt dort, dass *ei*/*ey* gelegentlich turn-initial und rede-einleitend gebraucht wird. Auch im GiesKaNe-Korpus wird *ey* häufig in der Umgebung des Vokativs verwendet. Der Vokativ steht aber nicht mehr direkt hinter der Partikel, was darauf hindeutet, dass *ey* nicht mehr vornehmlich als Anrede-Formel mit dem Vokativ verbunden ist, sondern die Rede-Sequenz eröffnet, in der dann auch ein Vokativ erscheinen kann.

(9) *eÿ eÿ sehe er doch Herr gefatter, was der Man in das buch geschriben.* [Nehrlich, 18. Jh.]

Auch ist der Anteil von rede-einleitenden Verwendungen mit 90% deutlich höher als im Referenzkorpus mit 16%. Wenngleich berücksichtigt werden muss, dass sich die Textzusammenstellungen im GiesKaNe-Korpus und im ReM sowohl qualitativ als auch quantitativ nicht problemlos vergleichen lassen, deutet sich hier eine klare funktionale Verschiebung an, die mit Hilfe von *Annimate* leicht sichtbar gemacht werden kann.

Als interessante Nebenergebnisse für das ReM lassen sich zudem folgende Befunde festhalten, die wir uns mit *Annimate* über den Parameter „Document metadata“ für die Werte „time“ und „language-type“ ausgeben haben lassen: Im Vergleich der Jahrhunderte ergibt sich, dass der Anteil der Verwendungen von *ei*/*ey* als Rede-Einleitung zunimmt (von 4% im 13. Jh. zu 35% im 14. Jh. – für das 11. Jh. wird kein entsprechender Treffer gefunden, für das 12. Jh. lediglich einer). Zudem zeichnet sich ab, dass die Interjektion zu 89% im mitteldeutschen Sprachraum verwendet wird. Dagegen sind nur 22 der 200 Belege im oberdeutschen Sprachraum belegt.

Diese Korpusbeobachtungen können natürlich nicht ohne eine sorgfältige Diskussion der Repräsentativität und Vergleichbarkeit der Daten interpretiert werden. Deutlich ist in jedem Fall geworden, dass sich aus den mit *Annimate* erzielten Suchergebnissen nicht nur auf komfortable Weise Korpus-Untersuchungen durchführen lassen, die zu aufschlussreichen Ergebnissen bei der Untersuchung von Sprachwandel führen können, sondern sich aus den gewonnenen Ergebnissen wiederum weiterführende Anschlussfragen und neue Hypothesen ableiten lassen.

4. Evaluierung von Korpora

Annimate ist ein Tool für den Export von Belegen. Exportiert werden können demnach auch nur Informationen, die in den entsprechenden Korpora vorhanden sind. Das bedeutet, dass auch nicht-korrekte Annotationen, wie sie in allen größeren Korpora zu unterschiedlichen Graden vorhanden sind, zu den jeweiligen Treffern ausgegeben werden. *Annimate* kann aber auch dazu eingesetzt werden, Korpora zu evaluieren und damit die Qualität der Daten zu verbessern. Wir demonstrieren dieses Potenzial anhand von Daten des GiesKaNe-Korpus.

In der Fallstudie zur Interjektion *ey* in Abschnitt 3 wurde überprüft, ob *ey* rede-initial verwendet wird. Für das mittelhochdeutsche Referenzkorpus lässt sich diese Information nicht aus den Korpusdaten gewinnen. Die Angabe ‚rede-initial – nicht-rede-initial‘ wurde daher in der mit *Annimate* ausgegebenen Excel-Liste manuell annotiert. Im GiesKaNe-Korpus werden Zitate dagegen auf der textgrammatischen Makro-Ebene auf der Annotationsebene „Sonstige Strukturen“ (`strukt`) annotiert. Anhand der Annotation lassen sich rede-einleitende Interjektionen damit durch eine jeweilige Suchabfrage elizitieren. Voraussetzung dafür ist jedoch die korrekte Annotation. Um die Annotationen für `strukt` systematisch zu überprüfen, lassen sich mit *Annimate* die gesamten Tokens eines Textes und die jeweils zu überprüfende Annotation im Excel-Format ausgeben. In unserem Fall führen

wir die Suche `tok @* doc="nehrlich"` aus und wählen in *Annimate* die folgenden Parameter, vgl. Tabelle 6.

Number	
Match annotation	Annotation = `tok`; Query node = `#1 tok`
Match annotation	Annotation = `strukt`; Query node = `#1 tok`

Tab. 6: Übersicht über die ausgewählten Werte in *Annimate*

Annimate gibt dann eine Datei aus, die eine Spalte mit den gesamten Tokens des Textes Nehrlich sowie eine Spalte mit den Annotationen enthält, die für die Annotationsebene `strukt` hinterlegt sind (z. B. Zitate, Parenthesen, Vokative). Diese Spalten können dann systematisch im jeweiligen Textzusammenhang überprüft werden.

5. Fazit

Annimate vereinfacht den Datenexport aus den historischen Referenzkorpora signifikant und ist damit ein intuitives und zeitsparendes Tool zur Datenaufbereitung, das nicht nur bei jeder Art von sprachhistorischen sowie diachronen Untersuchungen hilfreich ist, die auf Daten von historischen Korpora im ANNIS-Format basieren, sondern auch einen vielfältigen Einsatz in der Lehre ermöglicht.

Lassen sich diese Studien auch ohne *Annimate* durchführen? Zweifellos. Der bislang schwierige Export führt allerdings dazu, dass manchmal nicht auf die Referenzkorpora zurückgegriffen wird, sondern beispielsweise eigene Textkorpora erstellt und den Untersuchungen zugrunde gelegt werden. Das wird für einige Fragestellungen der historischen Sprachwissenschaft auch weiterhin unausweichlich sein. Für andere Fragestellungen stellt gerade aber auch die Vergleichbarkeit der Ergebnisse auf der Datenbasis der Referenzkorpora einen Vorteil dar, der in der historischen Sprachwissenschaft entsprechend genutzt werden können sollte. Hier stellt die Verwendung von *Annimate* nicht nur für die individuelle Arbeitspraxis, sondern auch für die Vergleichbarkeit der Studien einen großen Mehrwert dar.

Annimate macht zudem den Export so schnell und einfach, dass sich das Programm auch in der Lehre einsetzen lässt. Durch die Praktikabilität kann das Programm auch für Korpusabfragen in studentischen Arbeiten (nicht nur) zur historischen Sprachwissenschaft verwendet werden. Wir hoffen daher, dass das Tool nicht nur ein arbeitspraktisches Problem löst, sondern auch die Attraktivität diachroner Studien in den historischen Referenzkorpora im Vergleich zu weiteren ANNIS-basierten Korpora erhöhen und damit positive Auswirkungen auf die historische Sprachwissenschaftsforschung insgesamt haben wird. Ein weiteres großes Potenzial sehen wir zudem im Einsatz von *Annimate* bei der Evaluierung der Korpora. *Annimate* kann auch dazu beitragen, die Qualität der Korpora zu erhöhen. Gerade weil die Referenzkorpora inzwischen einen Status als Standard-Ressourcen innerhalb der Sprachwissenschaft des Deutschen innehaben, ist es unabdingbar, die Qualität der Annotationen zu überprüfen und zu verbessern. *Annimate* bietet auch hierfür ein großes Potenzial.

Ressourcen

Annimate zum Download auf GitHub: https://github.com/matthias-stemmler/annimate (Stand: 1.6.2025)

Repository zum Download von Korpora: www.laudatio-repository.org (Stand: 1.6.2025).

Stemmler, Matthias (o. D.): Annimate User Guide. https://matthias-stemmler.github.io/annimate/user-guide/ (Stand: 1.6.2025).

Literatur

Ágel, Vilmos/Hennig, Mathilde (Hg.) (2006): Grammatik aus Nähe und Distanz. Theorie und Praxis am Beispiel von Nähetexten 1650–2000. Tübingen: Niemeyer.

Bahlo, Nils U./Becker, Tabea/Kalkavan-Aydın, Zeynep/Lotze, Netaya/Marx, Konstanze/Schwarz, Christian/Şimşek, Yazgül (2019): Jugendsprache. Eine Einführung. Stuttgart: Metzler.

Elspaß, Stephan (2005): Sprachgeschichte von unten. Untersuchungen zum geschriebenen Alltagsdeutsch im 19. Jahrhundert. (= Reihe Germanistische Linguistik 263). Tübingen: Niemeyer.

Elspaß, Stephan (2010): Zum Verhältnis von ‚Nähegrammatik' und Regionalsprachlichkeit in historischen Texten. In: Ágel, Vilmos/Hennig, Mathilde (Hg.): Nähe und Distanz im Kontext variationslinguistischer Forschung. (= Linguistik – Impulse & Tendenzen 35). Berlin/New York: De Gruyter, S. 63–84.

Imo, Wolfgang/Müller, Melissa (2023): „Von „Ey Pickelhåring/das ist wider Ehr und Redligkeit" zu „ey TImo; lass_ma RISCHtisch laut (.) öh schrEIen" – *ey* und *ei* gestern und heute. In: Imo, Wolfgang/Wesche, Jörg (Hg.): Interaktionale Sprache im Dramenwerk von Andreas Gryphius. Literatur- und sprachwissenschaftliche Studien. (= Sprache – Literatur und Geschichte 53). Heidelberg: Winter, S. 99–186.

Kilian, Jörg (2012): »Man spricht hier in Meißen oft: Je nu!« Historische Gesprächswörter vom 17.–21. Jahrhundert. In: Leupold, Gabriele/Passet, Eveline (Hg.): Im Bergwerk der Sprache: Eine Geschichte des Deutschen in Episoden. Göttingen: Wallstein, S. 102–123.

Klein, Thomas/Dipper, Stefanie (2016): Handbuch zum Referenzkorpus Mittelhochdeutsch. In: Bochumer Linguistische Arbeitsberichte 19. https://linguistics.rub.de/forschung/arbeitsberichte/19.pdf (Stand: 31.5.2025).

Krause, Thomas/Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: Digital Scholarship in the Humanities 2016, 31, S. 118–139. http://dsh.oxfordjournals.org/content/31/1/118 (Stand: 31.5.2025).

Müller, Melissa (2022): Partikeln im Frühneuhochdeutschen. Eine korpuslinguistisch-interaktionale Untersuchung des Dramenwerks von A. Gryphius. Diss. Hamburg: Universität Hamburg. https://ediss.sub.uni-hamburg.de/bitstream/ediss/10841/1/2024-04-02%20Diss%20M%C3%BCller%20Melissa.pdf (Stand: 31.5.2025).

Petran, Florian/Bollmann, Marcel/Dipper, Stefanie/Klein, Thomas (2016): ReM: A reference corpus of Middle High German – corpus compilation, annotation, and access. In: Journal for Language Technology and Computational Linguistics 31, 2, S. 1–15. https://jlcl.org/article/view/208/206 (Stand: 25.3.2025).

Prell, Heinz-Peter (2005): Zur Neubearbeitung der mittelhochdeutschen Syntax von Ingeborg Schröbler. Mit einem Vergleich von Vers- und Prosatexten am Beispiel der Nominalphrase. In: Simmler, Franz (Hg.): Syntax. Althochdeutsch – Mittelhochdeutsch. Eine Gegenüberstellung von Metrik und Prosa. Akten zum Internationalen Kongress an der Freien Universität Berlin 26. bis 29. Mai 2004. (= Berliner sprachwissenschaftliceh Studien 7). Berlin: Weidler, S. 209–222.

Wegera, Klaus-Peter (1990): Mittelhochdeutsche Grammatik und Sprachgeschichte. In: Besch, Werner (Hg.): Deutsche Sprachgeschichte. Grundlagen, Methoden, Perspektiven. Festschrift für Johannes Erben zum 65. Geburtstag. Frankfurt a. M.: Lang, S. 103–113.

Zeman, Sonja (im Ersch.): Explorationen in GiesKaNe: Grammatik des Neuhochdeutschen zwischen Nähe, Distanz und Narration. In: Ágel, Vilmos/Linnenkohl, Marcel/Schäfer, Karolin/Sievers, Laura (Hg.): Grammatik des Neuhochdeutschen zwischen Gegenwart und Geschichte. Berlin/New York: De Gruyter.

Kontaktinformation

Matthias Stemmler
E-Mail: matthias.stemmler@gmail.com

Sonja Zeman
Lehrstuhl für Deutsche Sprachwissenschaft
Universität Augsburg
Universitätsstraße 10
86159 Augsburg
E-Mail: sonja.zeman@uni-a.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.75

Laura Duve/Valeria Schick

Wandel im Pronomengebrauch vom Frühneuhochdeutschen zum frühen Neuhochdeutschen

Korpusaufbereitung und Annotation mit *INCEpTION* und Auswertung und Visualisierung mit *ANNIS*

Abstract This paper presents the workflow created for the corpus preparation and annotation by two projects of the DFG-funded research group *Practices of Personal Reference*. Both projects examine diachronic developments in the use of pronouns deployed for personal reference, with one project compiling a corpus of German dramas from the 17th–19th centuries and the other collecting a corpus of German medical texts from the 15th–17th centuries. Operating with the tools *INCEpTION* and *ANNIS*, the data will be annotated and analyzed following project-specific linguistic categories. The methodological approach will be illustrated by an exemplary study of so-called *Heische*-formulae which contain the impersonal pronoun *man* and contextualize various acts of request. The comparison of different text types, represented in the projects, allows to draw conclusions about functional and contextual differences in the pronoun usage across time and intends to show a way in which cross-project research aims can be realized successfully.

Keywords OCR, *INCEpTION*, *ANNIS*, historic corpora, annotation

1. Einleitung

In diesem Beitrag beschreiben wir die Erstellung und Aufbereitung zweier historischer Korpora von der Datensammlung bis hin zur Veröffentlichung der annotierten Daten. Die Korpora entstammen zwei historisch ausgerichteten Teilprojekten der DFG-geförderten Forschungsgruppe *Praktiken der Personenreferenz* (Projektnummer: 457855466).[1] Beide Projekte setzen sich mit Wandelprozessen im personenreferierenden Pronomengebrauch auseinander.

Das Teilprojekt II *Das Pronomen ‚man' in der Diachronie des Deutschen* (im Folgenden *man*-Projekt) untersucht die Entstehung und Verfestigung von Referenzspektren und Funktionen des impersonalen Pronomens *man* sowie deren Herausbildung in verschiedenen frühneuzeitlichen Gattungen als Produkt wiederkehrender kommunikativer Bedarfe. Für diesen Gattungsvergleich wird u. a. ein Korpus von Pesttraktaten und Bäderkunden erstellt. Bei beiden Gattungen handelt es sich um medizinisch-wissenschaftliche Texte, die sich sowohl an Laien als auch an medizinisches Personal richten. Erstere beschreiben Symptomatik und Behandlungsmöglichkeiten der Pest (vgl. Eckkrammer 2015, S. 38), während zweitere sich mit Heilbädern und ihrer Anwendung beschäftigen (vgl. Fürbeth 2012, S. 194). Da

1 Informationen zur Forschungsgruppe und den beteiligten Teilprojekten finden sich unter: https://forschungsgruppe-pronomen.de (Stand: 3.12.2025).

beide Gattungen Instruktionen beinhalten und diese Kontexte eine wichtige funktionale Domäne von *man* darstellen, eignen sie sich besonders gut für die Ziele des Projekts.

Das Teilprojekt IV *Personenreferierende Pronomen in Dramen – interaktionale und dramaturgische Funktionen sowie historischer Wandel von Barock über Aufklärung zu Sturm und Drang und Klassik* (im Folgenden Dramen-Projekt) analysiert aus einer historisch-interaktionalen Perspektive Veränderungen im Gebrauch von Personal-, Indefinit- und Demonstrativpronomen im Zeitraum vom 17.–19. Jahrhundert. Von Interesse ist dabei z. B., welche sprachlichen Muster die fokussierten Pronomengruppen in dieser Zeit hervorgebracht haben und welche interaktionalen und dramaturgischen Funktionen damit erfüllt wurden. Als Datengrundlage der Projektarbeit dienen Dramentexte (Komödien und Tragödien) von vier einschlägigen Autoren aus der Epochenspanne Barock bis Klassik.

Der Fokus dieses Beitrags liegt einerseits auf der Darstellung des in den Projekten erstellten Workflows und der verwendeten Tools. Dazu existieren insbesondere für historische Textgattungen bisher wenige umfassende Beschreibungen, da sich daraus spezifische Anforderungen an nahezu alle Schritte der Korpuserstellung und -aufbereitung ergeben. Andererseits möchten wir auch auf Fallstricke und datenspezifische Probleme hinweisen, die sich beim Verfolgen pragmatischer Fragestellungen in einem derartigen Rahmen ergeben können. Insgesamt zeigt unser Vorgehen einen probaten Weg auf, wie korpusbasierte Forschungsvorhaben mit historischen Daten projektübergreifend und in größeren Teams zielführend realisiert werden können. Die von uns beschriebenen Schritte und Tools können sowohl im Ganzen als auch in einzelnen Punkten adaptiert und an eigene Forschungsfragen und Daten angepasst werden.

Zum Abschluss wollen wir den Nutzen unseres Vorgehens für pragmatische Fragestellungen exemplarisch an sogenannten Heische-Formeln mit *man* illustrieren. Diese weisen die Grundstruktur *man + Verb in 3. Pers. Sg. Konj. I* auf und kontextualisieren verschiedene Formen von direktiven Handlungen (vgl. Duve/Schick eingereicht). Abgesehen von Beschreibungen in Grammatiken gibt es bisher wenige empirische Untersuchungen zu diesem Phänomen (vgl. z. B. Jäger 1971), wobei v. a. seine historische Entwicklung auf formaler und funktionaler Ebene noch wenig Beachtung gefunden hat (vgl. aber Duve/Schick eingereicht). Ausgehend von der genannten Untersuchung werden wir daher in Kapitel 6 eine Fallstudie zu diesem Thema vorstellen. Zunächst wird jedoch der Prozess der Datensammlung und -vorbereitung mithilfe verschiedener Datenbanken und automatischer Texterkennung beschrieben. In Kapitel 3 wird die Korpusanreicherung durch automatisches Tagging skizziert. Daraufhin wird auf das Annotationsvorgehen eingegangen, das projektübergreifend durch das Annotationstool *INCEpTION* (vgl. Klie et al. 2018) realisiert wird. Kapitel 5 erläutert die Auswertung beider Datensets in dem Korpusanalysetool *ANNIS* (vgl. Krause/Zeldes 2016).[2] Bereits in den ersten fünf Kapiteln nutzen wir Heische-Formeln mit *man* als Beispiel zur Veranschaulichung der jeweiligen Arbeitsphasen und resümieren im Anschluss an Kapitel 6 die wichtigsten Erkenntnisse.

2 Die Tools *INCEpTION* und *ANNIS* können kostenfrei unter https://inception-project.github.io/ (Stand: 3.12.2025) sowie https://corpus-tools.org/annis/download (Stand: 3.12.2025) heruntergeladen werden.

2. Datensammlung

Zu Beginn der Projektarbeit bestand der erste Schritt in der Erschließung der für den Korpusaufbau relevanten Textgrundlagen als durchsuchbare TXT-Dateien. Die Datenbasis beider Projekte ist in Tabelle 1 dargestellt.[3]

Projekt	**Zeitraum**	**Anzahl der Texte je Gattung**	**Anzahl der Token im kürzesten/ längsten (Sprech-)Text**	**Anzahl der Token insgesamt und im Durchschnitt**
Teilprojekt II *Das Pronomen ‚man' in der Diachronie des Deutschen*	15.–17. Jh.	Pesttraktate (12)	n = 1.867/5.928	n = 77.590 Ø = 6.466
		Bäderkunden (7)	n = 7.975/16.180	n = 76.604 Ø = 10.943
Teilprojekt IV *Personenreferierende Pronomen in Dramen*	17.–19. Jh.	Dramen (31)	n = 4.153/54.126	n = 754.783 Ø = 24.348

Tab. 1: Datenbasis der beiden Teilprojekte nach Zeitraum, Gattung und Tokenanzahl

Der Tabelle ist zu entnehmen, dass die beiden präsentierten Korpora mit je 19 und 31 Texten eine unterschiedlich große Datenmenge umfassen. Die Pesttraktate sind im Durchschnitt am kürzesten, die Dramen zeichnen sich dagegen durch die längsten Texte aus. Die Länge der einzelnen Texte variiert insgesamt stark: von 1.867 Token (im *man*-Projekt) bis 54.126 Token (im Dramen-Projekt). Die Korpuszusammensetzung wird im Folgenden in Verbindung mit der Beschreibung der Textsammlung genauer erläutert.

Für das Dramen-Projekt lagen die Dateien bereits aus unterschiedlichen Quellen vor: So konnte für die Dramen von A. Gryphius, der die Epoche des Barocks repräsentiert, auf das aus zehn Stücken bestehende Korpus zurückgegriffen werden, das 2017–2020 im Rahmen des DFG-Projekts *Interaktionale Sprache bei Andreas Gryphius – datenbankbasiertes Arbeiten zum Dramenwerk aus linguistisch-literaturwissenschaftlicher Perspektive* erstellt und im JSON- bzw. CLS-Format veröffentlicht wurde (vgl. Imo et al. 2020; im Folgenden Gryphius-Projekt).[4] Für die Texte der drei übrigen Autoren wurde das online frei zugängliche *German Drama Corpus* (*DraCor*, vgl. Fischer et al. 2019) herangezogen, welches über 700 deutsche Dramen vom 15.–20. Jahrhundert im TXT- bzw. TEI-Format sowie bestimmte Metainfor-

3 Eine detaillierte Übersicht der konkreten Texte findet sich in Duve (2024) und Schick (2024). Im *man*-Projekt werden neben den hier genannten Textgattungen weitere zum Vergleich hinzugezogen. Diese sind z. T. bereits als Korpora vorhanden (neue Zeitungen, Hexenverhörprotokolle) oder liegen in Form einzelner TXT-Dateien vor, die in Excel annotiert und nicht als eigenständiges Korpus aufbereitet werden (Komödien, Musterdialoge in Fremdsprachenlehrwerken). Als Token gelten in beiden Projekten alle deutschen sowie fremdsprachlichen Wörter und Satzzeichen und als ‚Sprechtext' ist die Figurenrede (ohne Regiekommentare) zu verstehen.

4 Auf das Gryphius-Korpus kann über das *TALAR – Tübingen Archive of Language Resources* zugegriffen werden: https://fdat.uni-tuebingen.de/records/hshq9-w0g79 (Stand: 3.12.2025).

mationen zu den entsprechenden Autoren und Werken enthält.[5] Daraus konnten stellvertretend für die Epochen der Aufklärung, des Sturm und Drang sowie der Klassik insgesamt 21 Dramen der Autoren G. E. Lessing (6 Dramen), F. Schiller (8 Dramen) und J. W. v. Goethe (7 Dramen) ausgewählt werden. Die Auswahl fiel auf die bekanntesten Werke der Autoren, die einerseits die jeweilige Epoche sprachhistorisch und -stilistisch am besten widerspiegeln und andererseits eine adäquate Annäherung an die Rekonstruktion des mündlichen Sprachgebrauchs dieser Zeit zulassen. Die quantitative Ausgewogenheit der Teilkorpora war dabei aufgrund der vorrangig qualitativen Untersuchungsziele des Projekts kein gewichtiger Faktor.

Das Korpus des *man*-Projekts ist in zeitliche Abschnitte von 50 Jahren unterteilt und darüber hinaus nach räumlichen Parametern strukturiert. Es wurde pro Zeitabschnitt jeweils ein Text aus der ostmittel- und ostoberdeutschen Region einbezogen, da beide Regionen die nhd. Schriftsprache maßgeblich prägten. Die Texte sollten außerdem möglichst ähnlich im Umfang sein, daher wurde für den quantitativen Ausgleich teilweise ein zweiter Text ergänzt. Alle Schriftstücke lagen zunächst nur als PDF-Bilddigitalisate vor, welche über die Datenbank *Deutsche Fach- und Wissenschaftssprache bis 1700* (vgl. Klein et al. 2017) erschlossen wurden.[6] Diese Datenbank bündelt Metadaten und Informationen über verfügbare Digitalisate von Texten, „die für die frühe Geschichte der deutschen Fach- und Wissenschaftssprachen von Bedeutung sind“ (ebd.). Darunter fallen auch Pesttraktate und Bäderkunden, die dem Sachbereich ‚Medizin‘ zugeordnet werden. Die Datenbank verweist auf die jeweiligen Bibliotheken, bei denen die Digitalisate der Drucke frei verfügbar vorliegen. Über diesen Weg wurden die zugehörigen PDF-Dateien heruntergeladen.

Daraufhin wurde im *man*-Projekt mit der Software *Transkribus* (vgl. Kahle et al. 2017) eine automatische Texterkennung (OCR) durchgeführt.[7] *Transkribus* ist eine KI-gestützte Plattform, die Tools zur Digitalisierung, Transkription und Bearbeitung von historischen Drucken oder Handschriften bereitstellt (vgl. ebd., S. 463). Der Zugang ist bis zu einem bestimmten Textkontingent frei, für die Verarbeitung größerer Textmengen und für bestimmte Texterkennungsmodelle und Tools muss eine Nutzungslizenz erworben werden. Ein Kernangebot von *Transkribus* ist die Bereitstellung von Texterkennungsmodellen, die für unterschiedliche Textarten (Sprachen, Zeiträume, Druckarten) trainiert wurden. Neben den von *Transkribus* erstellten Modellen, sind auch Modelle anderer NutzerInnen verwendbar. Auch das Trainieren eigener Modelle ist möglich, lohnte sich aber für dieses Projekt angesichts der kleinen Textmenge nicht. Ein Problem bei der Suche nach einem passenden Modell war, dass die meisten spezifischeren, an Konventionen und Druckpraktiken der frühen Neuzeit angepassten Modelle Ligaturen, diakritische Zeichen oder Abkürzungen wie Nasalstriche auflösen, die aber gerade für historische Korpora relevante Informationen darstellen und daher auch in diesem Projekt erhalten bleiben sollten. Daher wurde das Modell *Transkribus Print M1* genutzt, welches an Drucken in verschiedenen europäischen Sprachen erprobt wurde (vgl. Transkribus Print Mehrsprachig 2023). Bei diesem Modell wird auch das Layout des jeweiligen Digitalisats automatisch erkannt und kann ggf. angepasst werden, um Zeilen- und Seitenumbrüche originalgetreu wiederzugeben (vgl. Abb. 1).

5 Das *DraCor* ist über https://dracor.org/ger (Stand: 3.12.2025) zu erreichen.

6 Die Datenbank findet sich unter: https://fachtexte.kallimachos.de/ (Stand: 3.12.2025).

7 Die *Transkribus*-Plattform ist unter folgender URL zu erreichen: www.transkribus.org/ (Stand: 3.12.2025).

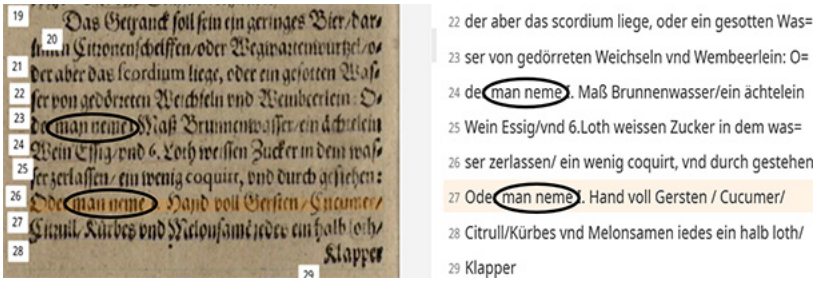

Abb. 1: Texterkennung eines Pesttraktats von C. H. Ayrer (1607) – Ansicht in *Transkribus* © Münchener Digitalisierungszentrum

Da das Modell eine Fehlerrate von 2,2% der Gesamtzeichen aufweist, mussten die entstandenen Transkripte noch einmal Korrektur gelesen werden. Die tatsächliche Fehlerrate variierte sehr stark in Bezug auf die Qualität und das Layout der Drucke in den Digitalisaten. Demzufolge ist bei der Planung eines solchen Vorhabens nur bedingt im Voraus bestimmbar, wie viel Zeit die Revision eines Datums in Anspruch nehmen wird. Von Vorteil ist jedoch, dass die Textkorrektur direkt in *Transkribus* durchgeführt werden kann. Sie wird dadurch erleichtert, dass z. B. Parallelstellen durch eine farbige Markierung gut erkennbar sind, sodass bei der Revision effizient vorgegangen werden kann (vgl. Abb. 1). Zudem ist es möglich, Dokumente per Link zu teilen und so gemeinsam Korrekturen vorzunehmen. Am Ende der Korrektur können die Texte als TXT-Dateien exportiert werden.

3. Bearbeitung der Daten

Die Daten beider Korpora wurden im Anschluss an die Vorverarbeitung in Anlehnung an den im Gryphius-Projekt konzipierten Workflow für die Annotation vorbereitet (vgl. Imo et al. 2020). Jeder Volltext wurde zunächst einzeln in ein im Rahmen des Gryphius-Projekts entwickeltes webbasiertes Tokenisierungstool eingefügt.[8] Das Tool segmentiert den Text automatisch in einzelne Token und gibt ihn im Text Corpus Format (TCF)[9] heraus, das die folgende Annotation in *INCEpTION* ermöglicht. Dafür erhalten alle Interpunktions- bzw. Sonderzeichen und Wörter eine eigene Token-ID (vgl. Abb. 2). Am Beispiel des Tokens mit der ID *"w394b">6.Loth<* ist jedoch zu erkennen, dass einige Zeichenkombinationen, z. B. die ohne zwischenstehende Leerzeichen, vom Programm fälschlicherweise als ein Token verarbeitet werden können, was je nach Forschungszweck einer manuellen Korrektur bedarf.[10] Bei diesem Zwischenschritt muss v. a. für ältere, weniger an einer Norm orientierte Texte genug Zeit eingeplant werden, da sich anderenfalls Probleme bei den Annotationen ergeben und z. B. im Falle der Lemmatisierung die Lemmata nicht tokengenau vergeben werden können.

8 Das frei verfügbare Tokenisierungstool ist unter folgender URL abrufbar: https://gryphius.sprache-interaktion.de/workflow/tokenizer_v4.html (Stand: 3.12.2025).

9 Eine Beschreibung des TCF erfolgt unter: https://weblicht.sfs.uni-tuebingen.de/englisch/tutorials/html/index.html (Stand: 3.12.2025).

10 Eine prozentuale Fehlerrate des Tools ist in der Dokumentation nicht aufgeführt und wurde in den Projekten nicht separat bestimmt.

Abb. 2: Heische-Formeln im Pesttraktat von C. H. Ayrer (1607) – Ansicht im Tokenizer

Außerdem werden auf Basis der Zeilenvorgaben des jeweiligen Originaltextes die entsprechenden Tokenkombinationen in Sätze eingeteilt. Das bedeutet, dass einer Zeile eine *sentence*-ID zugeordnet wird. Zu beachten ist dabei, dass dadurch sowohl ein einzelner Satz als auch mehrere Sätze unter eine *sentence*-ID gefasst werden können. Dies hat einerseits den Vorteil, dass bspw. in Dramentexten literarische Stilmittel wie Versprünge (Enjambements), die sich z. T. über mehrere Verse (Zeilen) erstrecken, erhalten bleiben und in der Annotationsansicht von *INCEpTION* gleichermaßen wiedergegeben werden. Andererseits müsste an dieser Stelle wie bei der Tokensegmentierung eine manuelle Nachkorrektur erfolgen, wenn die konkrete Satzanzahl relevant ist oder syntaktische Fragestellungen im Fokus stehen.

Das herausgegebene Tokenisierungsergebnis muss schließlich kopiert und in einem Texteditor (z. B. *Notepad++*, vgl. Ho 2003)[11] als Datei gespeichert werden, damit es im Webprogramm des Deutschen Textarchivs zur linguistischen Analyse historischer Texte *Cascaded Analysis Broker* (*DTA CAB*, vgl. Jurish 2012)[12] weiterbearbeitet werden kann. Dieses Programm ist wiederum für die orthografische Normkorrektur, die Lemmatisierung und das Part-of-Speech-Tagging (Wortartenzuweisung, im Folgenden POS-Tagging) der einzelnen Token, also die Ergänzung der eingespeisten Textvorlage um erste morphologische und morphosyntaktische Annotationsparameter, verantwortlich.[13] Die Verarbeitung kann dann je nach Textmenge einige Zeit (i. d. R. ein paar Minuten) in Anspruch nehmen. Nach aktuellem Stand bearbeitet der *DTA CAB* Dateien nur bis zu einer Größe von 1 MB. Bei größe-

11 Das Programm ist als Freeware unter https://notepad-plus-plus.org/ zu finden (Stand: 3.12.2025).

12 Auf den *DTA CAB* kann unter folgender URL zugegriffen werden: https://deutschestextarchiv.de/public/cab/ (Stand: 3.12.2025).

13 Sofern die Tokenisierung eines Textes im Vorfeld durchgeführt wurde, ist diese Option in der Maske des DTA-Webservice vor der Weiterverarbeitung zu entfernen, um Datenkontaminationen zu vermeiden.

ren Datenmengen kann allerdings der Programmdokumentation zufolge das DTA für weitere Informationen kontaktiert werden.[14] Abschließend muss das Ergebnis erneut als TCF-Datei gespeichert werden, welche die Grundlage der nächsten Arbeitsphase bildet.

4. Annotation

Nach der automatisierten Vorverarbeitung der Korpusdaten wurden diese in das Annotationsprogramm *INCEpTION* eingespeist. Für beide Projekte wurde in Absprache mit der IT der Universitäten jeweils eine eigene Webinstanz aufgesetzt, die das rechnerunabhängige Zugreifen auf die Daten und so die parallele Datenbearbeitung durch mehrere annotierende Personen über separat erstellte NutzerInnenaccounts zulässt. Es ist aber prinzipiell auch möglich, die Software lokal auf dem eigenen Rechner zu installieren.

In der jeweiligen Webinstanz wurde zunächst im Admin-Account ein Annotationsprojekt erstellt, in das die zuvor aufbereiteten TCF-Dateien hochgeladen wurden. Daraufhin wurden dort projektintern ausgearbeitete Layer (Annotationsebenen) angelegt, die anschließend mit den zugehörigen Tagsets (Bündel von Annotationskürzeln) verknüpft wurden.[15] Ein Layer kann dabei grundsätzlich auf unterschiedliche formale oder funktionale Codierungsziele ausgerichtet sein. Das entsprechende Tagset erlaubt – je nach Bedarf – als freies Tagset das eigenständige Befüllen des Annotationsfeldes (bspw. im Falle einer erforderlichen orthografischen Korrektur), während es als geschlossenes Tagset (bspw. auf der POS-Ebene) eine Vorauswahl an Merkmalsausprägungen der Annotationskategorie enthält. Für die formalen Layer sind in *INCEpTION* voreingestellte Tagsets vorhanden. So sind etwa für das POS-Tagging die Kodierungen des STTS (Stuttgart-Tübingen-Tagset, vgl. Schiller/Teufel/Stöckert 1995) hinterlegt, und es ist bereits eingestellt, dass die Annotation tokengenau erfolgt. Das vorbereitete Annotationsprojekt wurde anschließend für alle NutzerInnenaccounts zur Bearbeitung freigegeben.

Da die Texte beider Projekte eine automatische Vorannotation im *DTA CAB* durchlaufen haben, umfasste der erste Bearbeitungsschritt die parallele manuelle Korrektur jedes Textes auf den Ebenen i) orthografische Norm und ii) Lemma auf der Basis zuvor entwickelter projektspezifischer Guidelines.[16] Zeitgleich zur darauffolgenden POS-Korrekturphase wurde vor dem Hintergrund erster laufender Analysen mit der linguistischen Annotation ausgewählter Phänomene begonnen. So wurden im Dramen-Projekt für die ersten Texte Personalpronomen nach grammatischen Kategorien wie Kasus, Numerus, Genus etc. sowie nach semantisch-funktionalen Parametern wie Referenztyp oder Vorliegen einer Höflichkeitsform annotiert (vgl. das Token *man* in Abb. 3). Auf das Spektrum der Possessiv- und Demonstrativpronomen soll in einer späteren Projektphase ebenfalls genauer fokussiert werden. Im *man*-Projekt wurden Elemente der temporalen und lokalen Deixis annotiert,

14 Die Programmdokumentation des *DTA CAB* steht unter https://kaskade.dwds.de/~moocow/software/DTA-CAB/doc/html/DTA.CAB.WebServiceHowto.html (Stand: 3.12.2025) zur Verfügung. Im Dramen-Projekt wurden einige umfangreichere Werke im Texteditor Notepad++ nach einem eigenständig erarbeiteten Vorgehen händisch in mehrere Teildateien segmentiert und nach den beschriebenen Arbeitsschritten wieder zusammengesetzt. Dies kann hier allerdings aus Umfangs- und Komplexitätsgründen nicht im Detail aufgeschlüsselt werden.

15 Die genaue Vorgehensbeschreibung findet sich im User-Guide unter: https://inception-project.github.io/releases/33.5/docs/user-guide.html#_introduction (Stand: 3.12.2025).

16 Sowohl im Dramen- als auch im *man*-Projekt wurde sich zu Korrekturbeginn an den Guidelines zur formalen Annotation des o. g. Gryphius-Projekts (vgl. Imo et al. 2020) orientiert, da die Daten mit denselben Tools bearbeitet wurden. Es ist jedoch i. d. R. notwendig, vorhandene Guidelines mit Blick auf die eigenen Untersuchungsziele und die Datenspezifika aus der Praxis heraus anzupassen, was in beiden Teilprojekten innerhalb der jeweiligen Annotationsschleifen umgesetzt wurde.

da diese Hinweise auf Referenztypen liefern können. Zudem wurde in beiden Projekten eine Pilot-Codierung von Verben (Verbklasse, Bedeutungsgruppe u. Ä.) basierend auf Guidelines vorgenommen, die innerhalb der o. g. Forschungsgruppe *Praktiken der Personenreferenz* projektübergreifend konzipiert wurden (vgl. das Token anzůnde in Abb. 3). Im Zuge dessen können in beiden Datensets weiterhin nicht nur tokenspezifische Informationen annotiert werden, sondern auch Relationen zwischen einzelnen Token, die den Zusammenhang zwischen ihnen aufzeigen, wie die Verknüpfung zwischen *man* und *anzůnde*, die das Label <Heische Formel> trägt (vgl. Abb. 3). Die Annotation von Relationen ist eine Besonderheit in *INCEpTION*, die in anderen Annotationstools nicht immer vorhanden ist.

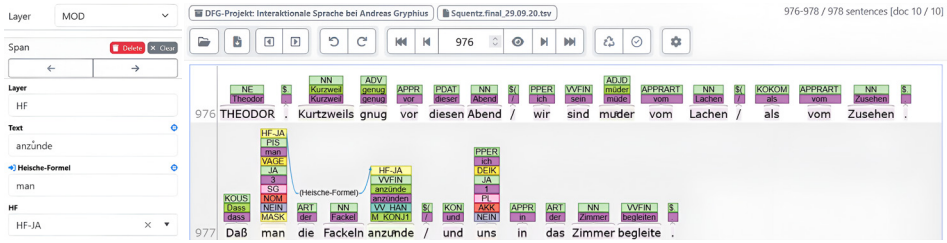

Abb. 3: Annotation einer Heische-Formel im Drama Peter Squentz von A. Gryphius (1657) – Ansicht in *INCEpTION*

Sowohl die Korrektur der formalen Merkmale als auch die Annotation linguistischer Kategorien wird jeweils von zwei AnnotatorInnen unabhängig voneinander im eigenen Account vorgenommen und daraufhin im Kurationsprozess miteinander verglichen.[17] Das Programm zeigt im Admin-Account innerhalb dieses Bearbeitungsschrittes durch farbige Markierungen der Annotationsfelder (Nicht-)Übereinstimmungen an. Diese Option wurde zum einen dafür genutzt, nach ersten Annotationen die Guidelines anzupassen und zu verfeinern, um die Fehlerrate zu verringern. Zum anderen werden die jeweiligen Annotationsversionen auf diesem Weg vom Admin regelmäßig korrigiert und wieder zu einer Version zusammengeführt, in der die nächste Annotationsetappe erfolgt. Die kuratierten Dateien sollten dann im TSV-Format exportiert werden, da nur in diesem Format alle eigenständig hinzugefügten Annotationen erhalten bleiben.

5. Auswertung

In einem letzten Schritt werden die annotierten Texte in das Korpustool *ANNIS* (vgl. Krause/Zeldes 2016) geladen, das die Suche nach sowie den Export und die Visualisierung von Belegen sprachlicher Phänomene ermöglicht. Dafür werden die entsprechenden Dateien aus *INCEpTION* im TSV-Format exportiert und mithilfe des Tools *Pepper* (vgl. Zipser/Romary 2010) in das von *ANNIS* benötigte GraphML-Format konvertiert.[18] *Pepper* ist ein *Java*-basiertes Programm, welches in der Command-Line eines Rechners bedient werden kann. Es ermöglicht die Umwandlung von linguistischen Korpora unterschiedlichster Ausgangsformate in andere Dateiformate. Es wurde jedoch v. a. mit dem Ziel entwickelt, die Konvertierung in das für *ANNIS* spezifische GraphML-Format zu ermöglichen. Während des Konvertierungsprozesses können auch einige Anpassungen der Daten wie Tagging-

17 Ein Inter-Annotator Agreement (vgl. Lemnitzer/Zinsmeister 2015, S. 61) wurde für die formalen Annotationsebenen nicht bestimmt. Für die funktionalen und damit fehleranfälligeren Kategorien wird dies angestrebt, allerdings wird die Annotation dieser Kategorien erst zum Ende der Projektlaufzeit abgeschlossen sein.

18 Das Tool *Pepper* ist unter https://corpus-tools.org/pepper/ (Stand: 3.12.2025) zu finden.

korrekturen bzw. -ergänzungen oder die Zusammenführung von Datensätzen vorgenommen werden.

ANNIS eignet sich als Infrastruktur zur Visualisierung und Auswertung von historischen Korpora, da Sonderzeichen wie Ligaturen oder diakritische Zeichen problemlos angezeigt werden. Für pragmatische Fragestellungen ist die Suchfunktion in *ANNIS* von Vorteil, da über die Suchsyntax (*ANNIS Query Language*)[19] neben einzelnen Token auch nach komplexen Phänomenen sowie auf unterschiedlichen linguistischen Ebenen gesucht werden kann. So können etwa V2- und VL-Heische-Formeln mit *man* ermittelt werden, indem nach dem Pronomen auf der Lemma-Ebene und nach finiten Verben (Voll-, Modal- und Auxiliarverb) im Konjunktiv I im Kontext von sechs Token mit der folgenden Abfrage gesucht wird:

LEMMA="man" ^1,6 POS=/(VVFIN|VMFIN|VAFIN) / _=_ MOD=/M_KONJ1/

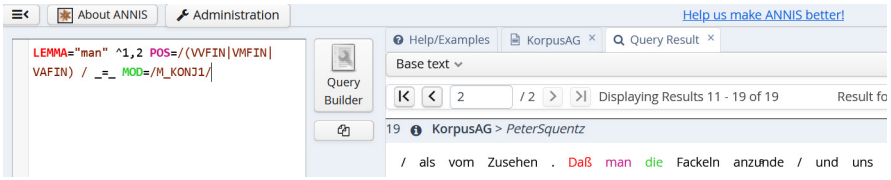

Abb. 4: Suche nach einer Heische-Formel in der Komödie *Peter Squentz* von A. Gryphius (1657) – Ansicht in *ANNIS*

Der relativ weit gesetzte Tokenabstand zwischen *man* und Verb dient der Ergebniserfassung der Fälle mit VL-Stellung, in denen komplexe NPs enthalten sind. Eine andere Möglichkeit stellt die Suche nach Relationen dar, wie z. B. nach der in *INCEpTION* veranschaulichten Relation < Heische-Formel >, die durch *ANNIS* angemessen dargestellt werden können. Die Belege können daraufhin im CSV-Format als KWICs oder mit den jeweils relevanten Annotationen exportiert werden. Dabei bietet das lokal installierbare Tool *Annimate* (vgl. Stemmler/Zemann i. d. B.) eine vereinfachte Infrastruktur für den Export.[20] Nach Projektabschluss werden die Korpora beider Teilprojekte voraussichtlich über eine öffentlich zugängliche *ANNIS*-Instanz für die Forschungsgemeinschaft verfügbar gemacht.[21]

6. Fallstudie: Heische-Formeln mit *man*

Im heutigen Sprachgebrauch finden sich sogenannte Heische-Formeln wie

(1) Man erinnere sich, daß…

oder

(2) Man nehme X (vgl. Zifonun/Hoffmann/Strecker 1997, S. 665; auch Duve/Schick eingereicht)

kaum noch wieder. Sie sind oft an spezifische (historische) Textgattungen und damit auch an spezifische Handlungsanforderungen gebunden (vgl. (1) für Zeitungen, (2) für Rezepte) und setzen sich im Gebrauch mit dem impersonalen Pronomen *man* in ihrer formalen

19 Genaue Informationen dazu können unter https://korpling.github.io/ANNIS/4/user-guide/aql/ (Stand: 3.12.2025) abgerufen werden.

20 *Annimate* kann unter https://github.com/matthias-stemmler/annimate (Stand: 3.12.2025) kostenfrei heruntergeladen werden.

21 Als Alternative dazu werden die Korpusdateien in unterschiedlichen Formaten zur Verfügung gestellt, sodass sie über eine lokal installierte Version von *ANNIS* oder mit anderen Tools analysiert werden können.

Grundstruktur aus *man + Verb in 3. Pers. Sg. Konj. I* zusammen.[22] Auf funktionaler Ebene werden Heische-Formeln Aufforderungsakten zugeordnet und stehen darin Imperativkonstruktionen nahe (vgl. Jäger 1970, S. 108), wobei die genaue Ausgestaltung der direktiven Kraft durch das weite Referenz- bzw. Funktionsspektrum des Pronomens und des Konjunktivs I sehr variabel und kontextsensitiv ist (vgl. Duve/Schick eingereicht). Vor diesem Hintergrund ist es interessant, zu überprüfen, ob die Struktur in früheren Sprachstufen frequenter war und/oder in einem breiteren Funktionsspektrum genutzt wurde. In diesem Zusammenhang hat eine erste empirische Untersuchung dieses Phänomens in Dramen des 16.–19. Jahrhundert gezeigt, dass die Konstruktion in dieser Gattung als Positionierungsressource abhängig von der sozialen Konstellation der SprecherInnen genutzt wird, um direktive Sprechakte zu formulieren, ohne u. a. soziale Hierarchien zu verletzen (vgl. Duve/Schick eingereicht). In der vorliegenden Fallstudie wird – wenn auch nur exemplarisch – der Blick auf die beiden weiteren Textgattungen des *man*-Projekts ausgeweitet.

Dafür wurde zunächst in beiden Korpora (vgl. Tab. 1) wie oben beschrieben nach Heische-Formeln mit *man* gesucht. Da für die Analyse des funktionalen Spektrums dieser Heische-Formeln die Belege in ihrem weiteren situativen Kontext aus qualitativer Sicht betrachtet werden müssen, wurde für den Export der Belegkollektion in *ANNIS* die größtmögliche Tokenzahl für den Kontext angegeben. Dass dieser in *ANNIS* besonders weit eingestellt werden kann, ist ein weiterer Vorteil des Tools für pragmatische Fragestellungen. Unentbehrlich blieb dabei allerdings die parallele Arbeit mit den Volltexten, da die inhaltliche Einbettung der Belege für ihre Interpretation zentral ist. Bei der interaktional ausgerichteten Analyse eines Dramenbelegs ist etwa das vorherige und nachfolgende Geschehen relevant, um die fokussierte Figurenkonstellation zu verstehen. Bei Rezeptbeschreibungen in den medizinischen Textsorten ist wiederum die Einordnung in Kapitel von Bedeutung, um zu rekonstruieren, ob die Heische-Formel sich an Kranke oder praktizierende Ärzte richtet.

Da die in Kapitel 3 dargestellte Annotation von Relationen in den Projekten noch aussteht, wurde in *ANNIS* nach dem Lemma *man* in Kombination mit einem Verb im Konjunktiv I im Abstand von sechs Token gesucht (vgl. Kap. 5). Nach Durchsicht und Eliminierung von Fehltreffern dienen 182 Belege (Pesttraktate n = 49, Bäderkunden n = 40, Dramen n = 93) als Datenbasis dieser Fallstudie. Die quantitative Übersicht in Abbildung 5 zeigt die relative Häufigkeit von *man* in Heische-Formeln pro 10.000 Token des Subkorpus der jeweiligen Gattung. Zwischen den Gattungen ist dabei ein deutlicher Unterschied erkennbar. Die Dramen weisen weniger Belege auf als die anderen beiden Gattungen[23] und es ist zu sehen, dass die Dichte an Heische-Formeln in den Pesttraktaten höher ist als in den Bäderkunden. Bei den Dramen zeigt sich zudem eine große Varianz zwischen den Autoren, wobei Heische-Formeln mit *man* in den Werken von Gryphius mit 3 Belegen/10.000 Token am häufigsten vorkommen (vgl. Abb. 6). Da sich die beiden Korpora nur in der zweiten Hälfte des 17. Jahrhundert überschneiden (vgl. Tab. 1), kann keine Aussage über den diachronen Verlauf getroffen werden. Auch innerhalb der Gattungen sind keine

22 Auch wenn diese Formeln mit unterschiedlichen Personal- und Indefinitpronomen realisiert werden können (vgl. Zifonun/Hoffmann/Strecker 1997, S. 663 f.), legen wir den Fokus vor dem Hintergrund unserer Projektarbeit auf den übergreifend relevanten Einsatz des impersonalen Pronomens *man* im Rahmen dieser spezifischen Struktur.

23 Die Anzahl der Belege gilt ausschließlich für den Sprechtext der Figuren, der für das interaktional ausgerichtete Analysevorgehen des Dramen-Projekts grundlegend ist. Sie würde jedoch mit hoher Wahrscheinlichkeit auch bei Einbezug von Paratexten wie Vor- bzw. Nachwort nicht signifikant höher sein, da innerhalb dieser Textpassagen direktive Handlungsformate nur in Ausnahmefällen wie der Adressierung der LeserInnen zu erwarten sind.

eindeutigen zeitlichen Entwicklungen im Einsatz der Heische-Formeln zu beobachten (vgl. beispielhaft Abb. 6 in Bezug auf die Dramen).

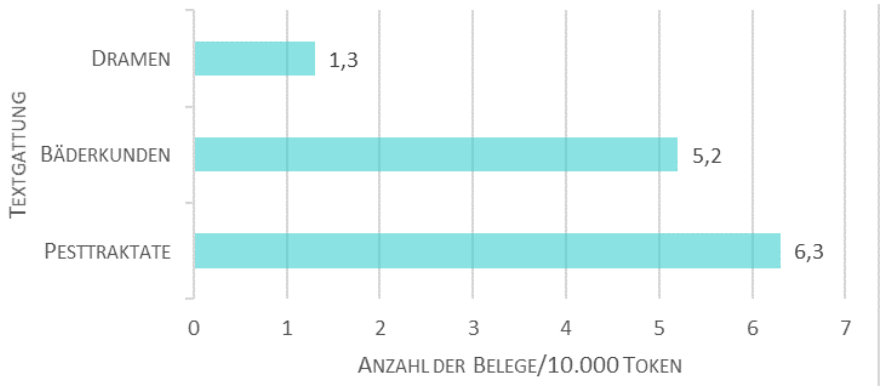

Abb. 5: Belege für *man* als Teil einer Heische-Formel pro 10.000 Token in den drei vertretenen Gattungen (n = 182)

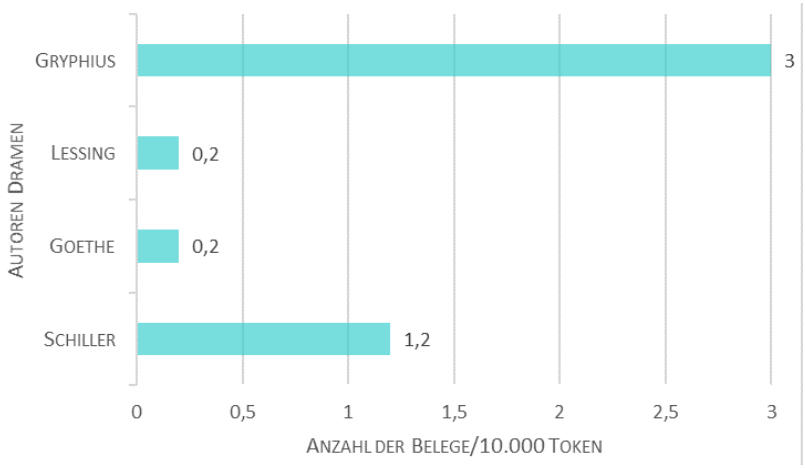

Abb. 6: Belege von *man* als Teil einer Heische-Formel pro 10.000 Token in den Dramen der jeweiligen Autoren (n = 93)

Im Folgenden wird exemplarisch anhand von zwei Textpassagen aufgezeigt, welche Funktionen Heische-Formeln mit *man* im Sprechtext der Dramen auf der einen und den medizinischen Gattungen auf der anderen Seite aufweisen. Bei Beispiel 3 handelt es sich um die Schlussszene der barocken Komödie *Peter Squentz* (1657) von A. Gryphius:

(3) **A. Gryphius – *Peter Squentz***
P. SQ. […] Gute Nacht Herr Kǒnig. Gute Nacht Frau Kǒnigin: gute Nacht Juncker / gute Nacht Jungfer / gute Nacht ihr Herren alle mit einander […].
THEODOR. Kurtzweils gnug vor diesen Abend / wir sind můder vom Lachen / als vom Zusehen. Daß man die Fackeln anzůnde / und uns in das Zimmer begleite.

In dieser Passage verabschiedet sich der Schreiber Peter Squentz, der für die Regieführung bei einem Theaterstück am königlichen Hofe verantwortlich war, nach der erfolgten Vorstellung von den ZuschauerInnen, zu denen u. a. der König Theodorus gehört. Daraufhin verkündet der König, dass die Adelsgesellschaft sich nach dem amüsanten Spektakel in ihre Gemächer zurückziehen möchte. In diesem Zuge äußert Theodorus die Heische-Formel *Daß man die Fackeln anzůnde / und uns in das Zimmer begleite*, welche im *man* + VL-Format innerhalb eines unabhängigen *dass*-Satzes gebildet wird (vgl. zu Heische-Formel-

Formaten in Dramentexten Duve/Schick eingereicht). Die darüber transportierte bindende Aufforderung gilt vor dem Hintergrund der sozialen Rolle, die der König repräsentiert, einer oder mehreren Gefolgsperson/en, deren Anwesenheit zwar nicht explizit erwähnt wird, jedoch kontextuell durch die Referenzleistung des impersonalen Pronomens *man* erschließbar ist. Durch die Verwendung der Heische-Formel mit offener Referenzleistung wird also deutlich, dass der König von der Erfüllung seines Willens ausgeht und dabei für ihn die konkreten Dienstleistenden keine Rolle spielen.

Wie die Studie von Duve/Schick (eingereicht) im Detail gezeigt hat, kommen Heische-Formeln in den innerhalb der beiden Teilprojekte untersuchten Dramentexten in je zwei V2- und VL-Realisierungsformaten vor, die zu unterschiedlichen Anteilen fünf direktive Handlungen, nämlich Irrelevanzkonditionale, Ratschläge, Ausrufe sowie (nicht-)bindende Aufforderungen, konstituieren (vgl. für erstere z. B. Schiller (1787): *Man nenn es Grille – Eitelkeit: gleichviel.*).[24] Mittels dieser Handlungen werden sowohl von Figuren eines höheren als auch niedrigeren sozialen Status verschiedene Willensbekundungen offengelegt. Dabei variieren die situativen Kontexte, in denen diese Willensbekundungen eingebettet sind, stärker als in den sich kontextuell ähnelnden Bäderkunden und Pesttraktaten, was am folgenden Beispiel im Ansatz verdeutlicht wird.

Der zweite ausgewählte Auszug entstammt einem Pesttraktat von C. H. Ayrer (1607) und ist Teil eines Kapitels, welches auf Getränke und Speisen eingeht, die für an der Pest Erkrankte empfohlen werden. Hier werden zwei Heische-Formeln im Rahmen einer Anleitung für die Zubereitung von gut bekömmlichen Getränken verwendet:

(4) **C. H. Ayrer – *Regiment und Ordnung vor den Seuchen der Pestilenz und Ruhr zu verwahren***
Oder man neme 1. Maß Brunnenwasser/ ein åchtelein Wein Essig/ vnd 6. Loth weissen Zucker in dem wasser zerlassen/ ein wenig coquirt, vnd durch gestehen: Oder man neme 1. Hand voll Gersten/ Cucumer/ Citrull/ Kůrbes vnd Melonsamē iedes ein halb loth/ Klapper oder klitzen Rosen/ Erbsichbeerlein jedes 1. Hand voll/ vnd lasse es in 2. Maß brunnē wasser sielen/ Tz letzt kann man ein wenig Zimmer daran thun/

Auch hier fungieren die V2-Heische-Formeln *man neme* und (*man*) *lasse sielen* als Aufforderungen, die allerdings sowohl an Kranke als auch an Wundärzte, die Kranke zu dieser Zeit versorgten (vgl. Eckkrammer 2015, S. 30), gerichtet sind. Das Pronomen *man* ermöglicht durch sein offenes Referenzpotenzial, dass die LeserInnen sich nach Bedarf von der Aufforderung angesprochen fühlen können. Bei der Erläuterung einer Zubereitungsoption im letzten Satz wird zusätzlich die Formulierung *kann man* verwendet, wobei das Modalverb *können* als Optionalitätsindikator aufzeigt, dass die vorausgehenden Heische-Formeln im Kontrast dazu eindeutig relevante Anweisungen kontextualisieren. Bemerkenswert ist außerdem, dass die sich hier wiederholende Konstruktion *man neme* als musterhafte Formel textübergreifend v. a. in den Pesttraktaten von Ayrer (1607), Strobelberger (1630) und Schorer (1667) vermehrt auftaucht (je 58,3%, 16,6% und 19% der Belege von Heische-Formeln mit *man* des jeweiligen Textes).

Insgesamt kommen alle Heische-Formeln (n = 89) sowohl in den Bäderkunden als auch in den Pesttraktaten in einem ähnlichen Kontext von Anleitungen bzw. Anweisungen vor. Diese können zum einen dazu dienen, konkrete Handlungsschritte zur Zubereitung von Medizin oder Bädern zu beschreiben oder Handlungsoptionen – etwa im Rahmen der Nut-

24 Für diese Studie wurden insgesamt 108 Belege aus 51 Dramentexten des 16.–19. Jahrhunderts ausgewertet. Die beiden anderen Textgattungen, die im *man*-Projekt analysiert werden, wurden dabei nicht einbezogen.

zung von Heilbädern – zu präferieren (vgl. Ryff 1549: *das man sich nit ubersettige mit speyß dermassen/ das der Magen darvon beschweret werde oder ein unwillen*). Die leicht höhere Belegzahl in den Pesttraktaten könnte sich daraus ergeben, dass dort anleitende Textteile einen größeren Stellenwert einnehmen, während in den Bäderkunden auch längere beschreibende oder theoretische Passagen vorkommen (vgl. Dammel/Duve eingereicht).

Das Funktionsspektrum in den beiden medizinischen Textgattungen ist demnach auf allgemeine Handlungsanweisungen beschränkt und ist damit eingegrenzter als in den untersuchten Dramentexten. Dies lässt sich v. a. auf die variierenden Drameninhalte sowie Beteiligungskonstellationen der Figuren zurückführen, die die Realisierung direktiver Sprechakte mit unterschiedlicher Verbindlichkeit relevant setzen (vgl. Duve/Schick eingereicht). Die höhere Frequenz von Heische-Formeln mit *man* in den anleitenden Texten legt die Tendenz nahe, dass sich für diese Gattungen schon früh typische und heute noch existente (wenn auch weniger gebräuchliche) Muster wie *man neme* (vgl. Ayrer 1607) herausgebildet haben. Zur Überprüfung dieser ersten Hypothesen wäre eine Ergänzung unserer Daten um die in den abgedeckten Zeiträumen fehlenden Gattungen (soweit vorliegend) notwendig. Dies wurde, wie oben erwähnt, anhand der im Rahmen des *man*-Projekts genutzten Dramentexte für das 16. Jahrhundert und die erste Hälfte des 17. Jahrhunderts zum Zwecke der Studie von Duve/Schick (eingereicht) bereits umgesetzt. In diesen Zeitabschnitten beträgt die relative Häufigkeit der Heische-Formeln nur 0,7 Belege pro 10.000 Token, was der Frequenz in den hier betrachteten Dramentexten nahekommt (vgl. Abb. 5, 6).

7. Fazit

Die Darlegung des Workflows zur programmgestützten Erstellung, Aufbereitung und Auswertung historischer Korpora sollte aufzeigen, wie diachrone, projektübergreifende Forschungsvorhaben ungeachtet unterschiedlicher Datengrundlagen ertragreich realisiert werden können. Dabei lag das Augenmerk darauf, inwieweit die von uns verwendeten Tools für die Arbeit (1) im Team, (2) mit pragmatischen Fragestellungen und (3) an historischen Daten besonders zielführend sind.

(1) Für eine kollektive Bearbeitung der Korpusdaten ist das von uns gewählte Programm *INCEpTION* insofern sehr gut geeignet, als es in Form einer Webinstanz den Zugriff auf die zu annotierenden Texte von verschiedenen Rechnern aus ermöglicht und es dem Administrator außerdem erlaubt, die zur Codierung benötigten Tagsets in einem Projekt einmalig anzulegen und sie allen Annotierenden gleichzeitig zugänglich zu machen. Des Weiteren erleichtert das Tool die Bündelung sowie Überprüfung mehrerer Annotationsergebnisse durch speziell dafür konzipierte Funktionen. (2) Ferner ist in *INCEpTION* bei der Auseinandersetzung mit pragmatischen Fragestellungen, wie den hier diskutierten Funktionen von Heische-Formeln mit *man*, die Möglichkeit der Erstellung von spezifischen Tagsets von Bedeutung, die konkret auf das eigene Forschungsinteresse zugeschnitten werden und mehrere linguistischen Ebenen umfassen können. In *ANNIS* sind wiederum die Auswertungs- und Darstellungsoptionen einfacher, aber auch komplexer Annotationen hilfreich, wie die für die Heische-Formeln codierte Beziehung zwischen pronominalen und verbalen Annotationsparametern. (3) Als Untersuchungsgrundlage stellen historische Daten aus den in Kapitel 2 und 3 beschriebenen Gründen in der Korpusaufbereitung meist eine Herausforderung dar. Es ist daher von Vorteil, zu ihrer Erschließung bestenfalls auf speziell für diesen Datentyp entwickelte und/oder daran erprobte Programme zurückzugreifen. Für die Vorhaben der hier diskutierten Projekte waren z. B. die Tools *Transkribus* und der *DTA CAB* unentbehrlich, um die jeweiligen Datensets für die Annotation vorzubereiten. Darüber hinaus war hinsichtlich der Textsegmentierung das für das Gryphius-Projekt imple-

mentierte Tokenisierungstool von Nutzen, da anderweitige Tokenizer ihre Analyse primär an orthografischen Zeichen ausrichten, die bspw. in frühneuzeitlichen Texten nur selten verwendet wurden und die demnach ungeeignete Anhaltspunkte für unsere Segmentierungsbedarfe geboten haben.

Anhand der Fallstudie zu Heische-Formeln mit *man* haben wir zudem illustriert, wie die projektübergreifend vorgenommenen Annotationen für konkrete Analysen fruchtbar gemacht werden können. Auffällig war dabei, dass Heische-Formeln v. a. in Pesttraktaten festzustellen waren, im Sprechtext der Dramen dagegen traten sie deutlich seltener auf. Aus qualitativer Sicht konnte ausgehend von den beiden exemplarisch diskutierten Auszügen aus einem Drama und einem Pesttraktat der Frühen Neuzeit das Funktionsspektrum der Konstruktion umrissen werden. Mit der Passage aus einer barocken Komödie wurde die Realisierung einer verbindlichen Aufforderung durch eine Heische-Formel mit *man* im Rahmen einer hierarchisch organisierten Interaktionssituation in den Blick genommen. Der Auszug aus einem Pesttraktat demonstrierte eine an die Leserschaft gerichtete Zubereitungsanleitung bzw. -anweisung, die über die darin eingesetzten Heische-Formeln kontextualisiert wurde und deren Ausführungsverbindlichkeit in erster Linie den (in-)direkt von der Krankheit betroffenen Personen gilt. Aufgrund des ähnlichen Gebrauchskontexts von Heische-Formeln innerhalb der Pesttraktate und Bäderkunden, mit denen bestimmte Adressatengruppen (Kranke, behandelnde Ärzte) allgemein im Handeln angeleitet werden, ist das Funktionsspektrum des Musters dort weniger ausdifferenziert als in Dramentexten mit variierenden Figurenkonstellationen und Interaktionssituationen im jeweiligen Geschehen. Dennoch lassen sich beide präsentierten Subfunktionen dem Dachbegriff direktiver Handlungen zuordnen, sodass sich in den untersuchten Datensets neben den gattungsbedingten Unterschieden durchaus auch Gemeinsamkeiten feststellen lassen.

Für weiterführende Forschungsvorhaben und die Ergänzung unserer Ergebnisse wäre z. B. ein Vergleich mit anderen Gebrauchstexten wünschenswert, denn Glaser (2002) weist im Hinblick auf die Textgattung ‚Kochrezepte' auf einen Anstieg im Einsatz von Heische-Formeln im 17. Jahrhundert hin, was u. a. in die Interpretation unserer quantitativen Beobachtungen neue Perspektiven einbringen würde.

Literatur

Dammel, Antje/Duve, Laura (eingereicht): Impersonale Pronomen als interpersonale Pronomen: Zur Perspektivierungsleistung von *man* im Zusammenspiel mit anderen Referenzformen in frühneuzeitlichen Bäderführern. In: Zeitschrift für germanistische Linguistik 26, 1 (Personenreferenz und Perspektive).

Duve, Laura (2024): Korpus. Teilprojekt 2: Referenzielle Praxis im Wandel: Das Pronomen *man* in der Diachronie des Deutschen. Forschungsgruppe Praktiken der Personenreferenz. https://tp2.forschungsgruppe-pronomen.de/korpus/ (Stand: 3.12.2025).

Duve, Laura/Schick, Valeria (eingereicht): „Man schaue und wundere sich." – Heische Formeln mit *man* als Positionierungsressourcen in Dramen des 16. bis 19. Jahrhunderts. In: Dammel, Antje/Imo, Wolfgang/Lanwer, Jens P. (Hg.): Pronomengebrauch und *stance taking*. Tübingen: Narr Francke Attempto.

Eckkrammer, Eva M. (2015): Medizinische Textsorten vom Mittelalter bis zum Internet. In: Busch, Albert/Spranz-Fogasy, Thomas (Hg.): Handbuch Sprache in der Medizin. (= Handbücher Sprachwissen 11). Berlin/München/Boston: De Gruyter, S. 26–46.

Fischer, Frank/Börner, Ingo/Göbel, Mathias/Hechtl, Angelika/Kittel, Christopher/Milling, Carsten/Trilcke, Peer (2019): Programmable corpora: Introducing DraCor, an infrastructure for the research

on European drama. In: Proceedings of Digital Humanities 2019: „Complexities“ (DH 2019). Utrecht, Utrecht University, S. 1–6.

Fürbeth, Frank (2012): Bäderdiskurse in den deutschsprachigen balneologischen Bestsellern des 16. Jahrhunderts (Paracelsus, Etschenreutter, Tabernaemontanus). In: Boisseuil, Didier/Wulfram, Hartmut (Hg.): Die Renaissance der Heilquellen in Italien und Europa von 1200 bis 1600. Geschichte, Kultur und Vorstellungswelt. Frankfurt a. M. u. a.: Lang, S. 193–212.

Glaser, Elvira (2002): Textmuster deutschsprachiger Kochrezepte im 19. und 20. Jh. In: Germanistik und Romanistik: Wissenschaft zwischen Ost und West. Materialien der internationalen wissenschaftlichen Konferenz „West-Ost: Bildung und Wissenschaft an der Schwelle des 21. Jahrhunderts“. Chabarovsk, 2002, S. 109–124.

Ho, Don (2003): Notepad++. https://notepad-plus-plus.org/ (Stand: 3.12.2025).

Imo, Wolfgang/Wesche Jörg/Müller, Melissa/Eggert, Lisa (2020): DGF-Projekt Interaktionale Sprache bei Andreas Gryphius – datenbankbasiertes Arbeiten zum Dramenwerk aus linguistisch-literaturwissenschaftlicher Perspektive. https://gryphius.sprache-interaktion.de/, https://gryphiusprojekt.wordpress.com/, https://gryphius.sprache-interaktion.de/workflow/ (Stand: 3.12.2025).

Jäger, Siegfried (1970): Indirekte Rede und Heischesatz. Gleiche formale Strukturen in ungleichen semantischen Bereichen. In: Moser, Hugo (Hg.): Studien zur Syntax des heutigen Deutsch. Paul Grebe zum 60. Geburtstag. (= Sprache der Gegenwart 6). Düsseldorf: Schwann, S. 103–117.

Jurish, Bryan (2012): Finite-state canonicalization techniques for historical German. Endliche Techniken zur Kanonikalisierung historischen deutschen Textes. Diss. Potsdam: Universität Potsdam.

Kahle, Philip/Colutto, Sebastian/Hackl, Günther/Mühlberger, Günter (2017): Transkribus – A service platform for transcription, recognition and retrieval of historical documents. 14th IAPR International conference on document analysis and recognition (ICDAR). Kyoto, Japan: IEEE: S. 19–24.

Klein, Wolf P./Breyl, Michael/Ahlers, Solveig/Kromer, Isabel/Schmid, Saskia (2017): Projektbeschreibung. https://fachtexte.kallimachos.de/Projektbeschreibung (Stand: 3.12.2025).

Klie, Jan-Christoph/Bugert, Michael/Boullosa, Beto/Eckart de Castilho, Richard/Gurevych, Iryna (2018): The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. In: Zhao, Dongyan (Hg.): Proceedings of system demonstrations of the 27th international conference on computational linguistics (COLING 2018), Santa Fe, New Mexico. Association for Computational Linguistics, S. 5–9.

Krause, Thomas/Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: Digital Scholarship in the Humanities 31, 1, S. 118–139.

Lemnitzer, Lothar/Zinsmeister, Heike (2015): Korpuslinguistik. Eine Einführung. 3., überarb. und erw. Aufl. Tübingen: Narr.

Schick, Valeria (2024): Unser Korpus. Teilprojekt 4: Personenreferierende Pronomen in Dramen: interaktionale und dramaturgische Funktionen sowie historischer Wandel von Barock über Aufklärung zu Sturm und Drang und Klassik. Forschungsgruppe Praktiken der Personenreferenz. https://tp4.forschungsgruppe-pronomen.de/unser-korpus/ (Stand: 3.12.25).

Schiller, Anne/Teufel, Simone/Stöckert, Simone (1995): Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Stuttgart: Institut für maschinelle Sprachverarbeitung. www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1995.pdf (Stand: 3.12.2025).

SfS Tübingen (o. D.): TCF format. https://weblicht.sfs.uni-tuebingen.de/englisch/tutorials/html/index.html (Stand: 3.12.2025).

Transkribus Print Mehrsprachig (2023): https://readcoop.eu/de/modelle/transkribus-print-multilanguage-dutch-german-english-finnish-french-swedish-etc/ (Stand: 3.12.2025).

Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno (1997): Grammatik der deutschen Sprache. 3 Bde. (= Schriften des Instituts für Deutsche Sprache 7). Berlin/New York: De Gruyter.

Zipser, Florian/Romary, Laurent (2010): A model oriented approach to the mapping of annotation formats using standards. In: Proceedings of the Workshop on Language Resource and Language Technology Standards (LREC 2010), May 2010, La Valette, Malta.

Kontaktinformation

Laura Duve
Universität Münster
Fachbereich 9 Philologie, Germanistisches Institut
Robert-Koche-Straße 29
48143 Münster
E-Mail: laura.duve@uni-muenster.de

Valeria Schick
Universität Hamburg
Fakultät für Geisteswissenschaften, Fachbereich SLM I, Institut für Germanistik
Von-Melle-Park 6
20146 Hamburg
E-Mail: valeria.schick@uni-hamburg.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.76

Thomas Burch/Jost Gippert/Sarah Ihden/Sarah Kwekkeboom/Ralf Plate/Lena Schnee/Ingrid Schröder/Roland Schuhmann/Johanna Wittmack

Neue Wege in der Analyse der historischen Wortbildung des Deutschen

Aufbau und Nutzungsmöglichkeiten der digitalen Forschungsumgebung ‚Wortfamilien diachron' (WoDia)

Abstract This paper outlines the long-term project 'Wortfamilien diachron' (= Word Families in Diachrony; WoDia), funded by Deutsche Forschungsgemeinschaft (DFG) and executed at the Universities of Frankfurt a. M., Hamburg, Rostock, and Trier. Its primary goal is the development of an online database-driven research environment for the historical word formation of German, and it is intended for implementation in both research and teaching. The dataset comprises lemmas from the Old High German, Middle High German, Old Saxon and Middle Low German dictionaries. WoDia integrates these lemmas into an overarching word family structure by putting each word in its position within its word family. The individual word formation elements and the word formation hierarchy are mapped in a structural formula. The four vocabularies are interconnected and linked to the online dictionary resources and to the reference corpora of historical German. This approach used in the research environment WoDia enables the user to analyse processes of change within the historical vocabularies of German, particularly transformations in word family structures and the usage of word formation elements. The methodical approach employed in constructing the database is presented in this paper, and concrete possibilities of application are illustrated by means of three exemplary case studies.

Keywords word formation, language change, word families, morphology, database

1. Einleitung

In den letzten Jahren hat die digitale Aufbereitung von Sprachdaten erheblich zugenommen, was sowohl der Forschung als auch der Lehre neue Möglichkeiten eröffnet. Für die historische Sprachwissenschaft sind in diesem Zusammenhang unter anderem die grammatisch annotierten Referenzkorpora zur deutschen Sprachgeschichte im Verbund ‚Deutsch Diachron Digital' mit den Korpora Altdeutsch, Mittelhochdeutsch, Frühneuhochdeutsch, Mittelniederdeutsch/Niederrheinisch und Deutsche Inschriften (vgl. Ihden et al. 2023)[1] von besonderer Bedeutung. Im Bereich der Lexikographie ist vor allem die Digitalisierung zahlreicher historischer Wörterbücher hervorzuheben.[2] Für die Forschung zur historischen Wortbildung steht jedoch bislang noch kein geeignetes Instrument zur Verfügung, denn die Annotationen in den Korpora konzentrieren sich auf Wortart, Flexion und ansatzweise

1 Vgl. auch die Webseite des Verbundes ‚Deutsch Diachron Digital' unter www.deutschdiachrondigital.de/ (Stand: 1.7.2025).

2 Die Wörterbücher sind im Trierer Wörterbuchnetz online publiziert und digital verfügbar unter https://woerterbuchnetz.de (Stand: 1.7.2025).

auf die Satzstruktur, während Wortbildungsstrukturen nicht ausgezeichnet sind. Das WoDia-Projekt schließt diese Lücke, indem eine digitale Forschungsumgebung geschaffen wird, die den historischen Wortschatz des Deutschen varietäten- und sprachstufenübergreifend auf der Wortbildungsebene analysiert, ihn in einer Wortfamilienstruktur abbildet und miteinander verbindet sowie eine Verknüpfung mit den Lemmata der Online-Wörterbücher und der Referenzkorpora vornimmt.

2. Ziele des Projekts

Im WoDia-Projekt wird eine umfassende empirische Grundlage für die Forschung zur historischen Wortbildung des Deutschen bereitgestellt. Die Basis bilden die Referenzwörterbücher dieser Wortschätze:

- Für Althochdeutsch das Wörterbuch von Jochen Splett (1993) (AWB^{Spl}) und das noch in Ausarbeitung befindliche Leipziger ‚Althochdeutsche Wörterbuch' (AWB).
- Für Mittelhochdeutsch Lexers Handwörterbuch (1872–1878) (Lexer) und das in Ausarbeitung befindliche neue Mittelhochdeutsche Wörterbuch (MWB).
- Für Altsächsisch das Handwörterbuch von Heinrich Tiefenbach (2010) (ASWB) und wiederum das AWB, das auch den gesamten altsächsischen Wortschatz der kleineren Texte und Glossen mitbearbeitet.
- Für Mittelniederdeutsch das in Bearbeitung befindliche Mittelniederdeutsche Handwörterbuch (MNWB) sowie ergänzend das Wörterbuch von Lübben/Walther (1888) (LW).

Mit diesen Wörterbüchern ist die gesamte ältere deutsche Sprachgeschichte bis ins 15. Jahrhundert (Lexer für das Mittelhochdeutsche) bzw. bis zum Anfang des 17. Jahrhunderts (MNWB für das Mittelniederdeutsche) abgedeckt. Für das Mittelhochdeutsche wird dabei jeweils notiert, ob ein Lemma bereits vor 1350 (rd. 50.000 Wörter) oder erst nach 1350 (rd. 31.000 Wörter) bezeugt ist. So ist der Anteil des Wortschatzes aus der älteren frühneuhochdeutschen Überlieferung unterschieden von dem mittelhochdeutschen Wortschatz im engeren Sinne der heutigen Epochengliederung.

Die Referenzwörterbücher werden als Materialbasis für WoDia gewählt, da der Umfang ihrer Wortschätze den der entsprechenden Referenzkorpora ReA (Referenzkorpus Altdeutsch = Althochdeutsch + Altsächsisch), ReM (Referenzkorpus Mittelhochdeutsch) und ReN (Referenzkorpus Mittelniederdeutsch/Niederrheinisch) erheblich übersteigt. Die Referenzkorpora zielen nicht primär auf eine lexikalische, sondern stärker auf eine grammatische, vor allem flexionsmorphologische Erschließung der jeweiligen Sprachstufe. Tabelle 1 zeigt die unterschiedlichen Umfänge.

Sprache	Wörterbuch	Referenzkorpus
Althochdeutsch	28.500 (AWB^{Spl})	10.900 (ReA)
Altsächsisch	5.700 (ASWB)	4.100 (ReA)
Mittelhochdeutsch	81.000 (Lexer, bis 15. Jh.) davon 50.000* bis 1350 (MWB)	22.300 (ReM, bis 1350)
Mittelniederdeutsch	80.000* (MNWB)	17.800 (ReN)

Tab. 1: Anzahl der Lemmata in Wörterbüchern und Referenzkorpora; * = hochgerechnete Endumfänge in noch nicht abgeschlossenen Wörterbuch-Projekten

Die vier Wortschätze werden im Projekt miteinander verknüpft, wortbildungsmorphologisch analysiert, in eine übergreifende Wortfamilienstruktur überführt und in einer Forschungsumgebung mit benutzerfreundlichem Frontend für eine Vielzahl von Fragestellungen der historischen Wortbildungslehre bereitgestellt. Die Lemmata der Datenbank sind über ihre IDs mit den Online-Angeboten der Referenzwörterbücher verlinkt, sodass zu jedem Wort seine lexikographische Darstellung verglichen werden kann, im Falle der großen Belegzitatwörterbücher (AWB, MWB) auch ihr Gebrauch in den Belegzitaten. Die Online-Angebote für das AWB^{Spl}, das ASWB, das MNWB und für LW werden vom Projekt selbst eingerichtet und im Trierer Wörterbuchnetz publiziert, in dem bereits die übrigen Referenzwörterbücher zugänglich sind. Über die Lemma-IDs erfolgt auch eine Verlinkung in die Referenzkorpora zur deutschen Sprachgeschichte, sodass ihre flexionsmorphologische Annotation um die wortbildungsmorphologische Analyse in WoDia ergänzt werden kann. Das ReM ist bereits mit der Mittelhochdeutsch-Lemmaliste (Lexer + MWB) verknüpft, für das ReA und das ReN ist die Verlinkung in die Referenzwörterbücher noch einzubringen.

Das auf neun Jahre (2024–2033) angelegte Projekt wird in drei Arbeitsgruppen an vier Standorten durchgeführt, die eng miteinander kooperieren. Mittelniederdeutsch wird an den Universitäten in Hamburg und Rostock bearbeitet, die Arbeiten zum Althochdeutschen, Altsächsischen und Mittelhochdeutschen werden an der Universität Frankfurt a. M. durchgeführt. Die Einrichtung der zentralen Datenbank, die Online-Publikation der Referenzwörterbücher und die Organisation des digitalen Workflows liegt in den Händen der Arbeitsgruppe am Trierer Kompetenzzentrum (Trier Center for Digital Humanities).[3]

3 In Frankfurt a. M. setzt sich das Projektteam aus Jost Gippert, Ralf Plate (Projektleitung), Sarah Kwekkeboom, Roland Schuhmann (wissenschaftliche/r Mitarbeiter/in), Alexandre Jouan und Janina Damberg (Hilfskräfte) zusammen (vgl. die Webseite am Frankfurter Standort unter www.uni-frankfurt.de/153903162/Wortfamilien_diachron, Stand: 1.7.2025), in Hamburg aus Ingrid Schröder (Projektleitung), Lena Schnee (wissenschaftliche Mitarbeiterin), Ronia Seiri und Sarah Möller (Hilfskräfte) (vgl. die Webseite am Hamburger Standort unter www.slm.uni-hamburg.de/niederdeutsch/forschung/projekte/wortfamilien-diachron.html, Stand: 1.7.2025), in Rostock aus Sarah Ihden (Projektleitung), Johanna Wittmack (wissenschaftliche Mitarbeiterin) und Lara Stapel (Hilfskraft) (vgl. die Webseite am Rostocker Standort unter https://www.germanistik.uni-rostock.de/forschung/sprachwissenschaft/wodia, Stand: 16.2.2026) und in Trier aus Thomas Burch (Projektleitung), Julia Fischer, Felix Thielen (wissenschaftliche/r Mitarbeiter/in) sowie Marina Spielberg und Bianca Ottenberg (Hilfskräfte) (vgl. die Webseite am Trierer Standort unter https://tcdh.uni-trier.de/de/projekt/wortfamilien-diachron-wodia, Stand: 1.7.2025).

3. Arbeitspakete und Methoden

Die Bearbeitung der digitalen Daten in WoDia basiert auf einem mehrstufigen Verfahren zur Analyse und Strukturierung der historischen Wortschätze. Diese Vorgehensweise wurde in einem Pilotprojekt an der Universität Frankfurt a. M. in den Jahren 2016–2019 erprobt (vgl. Plate 2020, 2022).[4] Das Ziel bestand darin, die althochdeutsche Wortfamiliengliederung nach dem Modell von Splett (AWB^{Spl}) auf das Mittelhochdeutsche zu übertragen. Besonders hervorzuheben sind die Vorarbeiten von Thomas Klein (2018), auf denen die WoDia-Systemarchitektur aufbaut.

Die Analysen in WoDia folgen einem systematisch modularisierten Ablauf, der es ermöglicht, große Wortschatzmengen effizient zu bearbeiten. Für jede Sprachvarietät (Althochdeutsch, Mittelhochdeutsch, Altsächsisch, Mittelniederdeutsch) werden bestimmte Arbeitsschritte durchlaufen, wobei aufgrund der oben erwähnten Vorarbeiten ein unterschiedlicher Bearbeitungsstand in den Sprachstufen vorliegt. Aus diesem Grund sind die Arbeitspakete nicht zwingend nacheinander angeordnet, sondern können je nach Sprachstufe teilweise auch parallel liegen. Dieser Abschnitt bietet einen Überblick über die Arbeitspakete in WoDia und stellt die technischen Verfahren vor, mit denen die teil-automatisierte Analyse der historischen Wortschätze ermöglicht werden soll.

3.1 Erstellung der Lemmalisten

Die Datenbasis bilden die auf den jeweiligen Referenzwörterbüchern basierenden Lemmalisten. Für Althochdeutsch und Mittelhochdeutsch liegen diese bereits vor und sind miteinander verknüpft, für Altsächsisch und Mittelniederdeutsch müssen sie zunächst aus den Digitalisaten der Referenzwörterbücher zusammengestellt werden. Insbesondere für den umfangreichen mittelniederdeutschen Wortschatz stellt dies keine triviale Aufgabe dar. Aufgrund der im MNWB teilweise uneinheitlichen formalen Auszeichnung beispielsweise der Verweislemmata oder Nestartikel[5] lässt sich nicht für alle Lemmata eindeutig formal festlegen, was ein Lemma ist und wie/ob es in die Liste aufgenommen werden soll, sodass dieser Schritt nicht voll automatisierbar ist. Für jedes Lemma werden aus den mittelniederdeutschen Wörterbüchern (MNWB + LW) zusätzlich die grammatischen Angaben sowie – soweit verfügbar – die Bedeutungsangaben extrahiert (vgl. exemplarisch Tab. 2 für das Mittelniederdeutsche).

Lemma	Grammatische Angabe	Bedeutungsangabe
begrāven	stv.	begraben
begrāvinge	f.	Bestattung, Begräbnis
begrēvenisse	f.	Bestattung, Begräbnis

Tab. 2: Auszug aus der mittelniederdeutschen Lemmaliste

4 Das Projekt ‚Wortfamilien. Computergestützte Etablierung epochenübergreifender Wortfamilienstrukturen' war Teil der mit Mitteln des BMBF geförderten Initiative für ein eHumanities-Zentrum für historische Lexikographie (ZHistLex) und wurde von Jost Gippert und Ralf Plate geleitet.

5 Nestartikel im MNWB sind in den Artikel eines durch Fettdruck gekennzeichneten Hauptlemmas (z. B. *veftich*) eingebundene Wörterbuchartikel weiterer, aus dem Hauptlemma gebildeter Lemmata (z. B. *veftichmāker*); letztere werden typographisch durch Kursivierung abgehoben.

3.2 Segmentierung der Lemmata und Ermittlung der Wortfamilienstämme

Ein erster Schritt auf dem Weg zur Zuordnung der Wortfamilien besteht in der Segmentierung der Lemmata in ihre Konstituenten. Diese wird mithilfe von zuvor erstellten Affixlisten durchgeführt. Auf diese Weise können die Affixe angesteuert und abgetrennt und so die Stämme der jeweiligen Lemmata freigelegt werden. Zusätzlich werden mögliche Kompositionsglieder anhand von Simplexlisten identifiziert, die zuvor durch ein Skript erstellt wurden (vgl. Abschn. 3.4). Dabei können auch Fugenelemente erkannt und abgetrennt werden.[6] Durch die Segmentierung, d. h. das Abtrennen der Affixe (wie *be-*, *-inge* und *-nisse* in Tab. 3) bzw. das Trennen von Simplizia, wird der Stamm zunächst in der Form ermittelt, in der er im entsprechenden Lemma realisiert ist. So zeigt z. B. in mnd. *begrēvenisse* (vgl. Tab. 3) der freigelegte Stamm *-grēv-* den durch das Suffix *-nisse* ausgelösten Umlaut. Dieser Stamm muss von Umlaut, Ablaut und anderen Lautwechselphänomenen bereinigt werden, um den Wortfamilienstamm, die Grundform des Stammparadigmas eines Lexems (vgl. Duden-Grammatik 2022, § 1050; Duden-Grammatik 2016, S. 664 f.; Eisenberg 2020, S. 30, 231), zu ermitteln. So erhält man hier im Ergebnis die Grundform *grāv*, den verbalen Stamm der Wortfamilie *GRĀVEN*, der die Bildungen damit zugeordnet sind. Alle Lemmata, die derselben Wortfamilie angehören, hier *GRĀVEN*, werden zu einem späteren Zeitpunkt nach dem Muster im AWB^{Spl} in ihrer Wortfamilie gruppiert (vgl. Abb. 1 in Abschn. 4.2.1).

Lemma	Segmentierung	Stammform	Wortfamilien-stamm	Wortfamilie
begrāven	be-grāv-en	grāv	grāv	GRĀVEN
begrāvinge	be-grāv-inge	grāv	grāv	GRĀVEN
begrēvenisse	be-grēv-e-nisse	grēv	grāv	GRĀVEN

Tab. 3: Segmentierung und Stammermittlung anhand mittelniederdeutscher Beispiele

In WoDia erfolgt eine Verknüpfung der vier historischen Varietäten auf zwei Ebenen: Einerseits zwischen den einzelnen Lemmata (z. B. ahd. *bigraban* – mhd. *begraben* – as. *bigravan* – mnd. *begrāven*, vgl. Tab. 4), andererseits über die Wortfamilien (z. B. ahd. *GRABAN* – mhd. *GRABEN* – as. *GRAVAN* – mnd. *GRĀVEN*),[7] wodurch auch miteinander verwandte Lemmata derselben Familie, die keine etymologische Entsprechung in einer der anderen Varietäten haben, einander zugeordnet sind (z. B. mnd. *grefte*, f. ‚Graben' und mnd. *grāvengenger*, m. ‚Wächter für Wälle und Gräben').

6 Die Untersuchung und Diskussion der Frage, ab welchem Zeitpunkt es sich um ein Fugenelement und nicht mehr um ein Flexionsmorphem handelt, kann im Rahmen von WoDia nicht geleistet werden. Auf Flexion basierende potenzielle Fugenelemente werden daher grundsätzlich als Fugenelement annotiert.

7 In der Forschungsumgebung kann bei der Suche zwischen Lemma, Wortfamilie und Strukturformel differenziert werden. Eine Suche nach z. B. mhd. *graben* auf der Ebene der Wortfamilie gibt auch die Wortfamilien der anderen Varietäten aus.

Sprachstufe	Lemma	Segmentierung	Wortfamilie
Althochdeutsch	bigraban	bi-grab-an	GRABAN
Mittelhochdeutsch	begraben	be-grab-en	GRABEN
Altsächsisch	bigravan	bi-grav-an	GRAVAN
Mittelniederdeutsch	begrāven	be-grāv-en	GRĀVEN

Tab. 4: Verknüpfung zwischen den Varietäten über das Einzellemma und die Wortfamilie

Die Tatsache, dass für das Althochdeutsche die Segmentierung, die Zuordnung zu Wortfamilien und die wortbildungsmorphologische Analyse bereits im AWB^{Spl} vorliegen, kann in WoDia auch für die Arbeiten an den anderen Sprachstufen genutzt werden. Die Verknüpfung mit den etymologisch entsprechenden Einträgen einer bereits analysierten Lemmaliste (Mittelhochdeutsch mit Althochdeutsch, Altsächsisch mit Althochdeutsch, Mittelniederdeutsch mit Mittelhochdeutsch) bietet den Vorteil, dass deren strukturelle Analyse (vgl. Abschn. 3.3) übernommen werden kann.

3.3 Erstellung von Strukturformeln

Die wortbildungsmorphologische Analyse der Lemmata in WoDia wird in einer Strukturformel wiedergegeben. Das Vorbild hierfür bieten die Wortfamilienwörterbücher zum Althochdeutschen (1993) und zur deutschen Gegenwartssprache (2009) von Jochen Splett. Diese bieten nicht nur eine Gliederung des Wortschatzes in Wortfamilien, sondern liefern für jedes Lemma eine hierarchische, geklammerte Strukturformel mit Informationen zur Konstituentenstruktur, zur Wortart der Stämme und zur Wortbildungsart (Komposition, Derivation, Konversion). Einige Beispiele aus dem Wortfamilienwörterbuch zur Gegenwartssprache (Splett 2009, Bd. 1, S. 416–417) verdeutlichen dieses Prinzip:

- *Bau-er*: (wV)sS – zu lesen als: substantivische Ableitung aus dem Verbstamm BAU mit dem Suffix *-er* (w = Wurzel/Stamm; V = Verb; s = Suffix; S = Substantiv).
- *Acker-bau-er*: (wS)((wV)sS) – Kompositum mit den Konstituenten ACKER und BAU-ER.
- *Boots-bau-er*: ((wS)((wV)S))sS – substantivische Ableitung aus dem zusammengesetzten Substantiv BOOTS-BAU mit dem Suffix *-er*; *-bau* wiederum ist eine substantivische Konversion des Verbstamms BAU.
- *Er-bau-er*: (p(wV))sS – Substantivableitung aus dem präfigierten Verbstamm ER-BAU mit dem Suffix *-er* (p = Präfix).

Aus Doppel- oder Mehrfachmotivationen ergeben sich Mehrdeutigkeiten, denen durch die Angabe alternativer Strukturformeln Rechnung getragen wird, z. B. im Fall des Adjektivs *baukünstlerisch* (Splett 2009, Bd. 1, S. 414) mit drei Formeln für die drei möglichen Analysen: Mit *baukünstler-isch* vs. *bau-künstlerisch* als unmittelbaren Konstituenten, und wiederum im ersten Fall mit *bau-künstler* vs. *baukunst-ler* als unmittelbaren Konstituenten des ersten Elements.

Die Strukturformeln bieten mit ihrer formalisierten Beschreibung der zugrunde liegenden Wortbildungsschritte und der beteiligten Morpheme eine einzigartige Möglichkeit zur Untersuchung historischer Wortbildungsregularitäten, die in den Printpublikationen der Splettschen Wortfamilienwörterbücher bislang kaum genutzt werden konnte. Mit der Online-Publikation des AWB^{Spl} wird in einem ersten Schritt die Suche über die Strukturformeln für das Althochdeutsche ermöglicht. Später wird die WoDia-Forschungsumgebung für alle erwähnten historischen Wortschätze die Strukturformeln sämtlicher Lemmata durchsuchbar und exportierbar machen.

3.4 Teil-Automatisierung der Arbeitsschritte

Die oben beschriebenen Arbeitsschritte können durch den Einsatz regelbasierter Programme unterstützt werden. Als methodische Leitlinie für alle Arbeitsschritte gilt ein iteratives teil-automatisiertes Verfahren. Dies bedeutet, dass mithilfe von Skripten automatische Vorschläge, z. B. zur Segmentierung der Lemmata, zum Wortfamilienstamm, zur Verknüpfung der Lemmata und zu den Strukturformeln, erzeugt werden, die in einem nächsten Schritt händisch zu prüfen sind. Dabei können Fehlerquellen entdeckt und die Skripte daraufhin angepasst und erneut eingesetzt werden, wodurch die automatischen Vorschläge sukzessive verbessert werden. Die Skripte zur Bearbeitung des mittelhochdeutschen Wortschatzes wurden bereits im Rahmen der erwähnten Vorarbeiten erstellt (zur technischen Umsetzung siehe auch Abschn. 3.5). Für das Mittelniederdeutsche müssen sie entsprechend angepasst werden. Am Beispiel der Arbeiten am mittelniederdeutschen Wortschatz soll im Folgenden die Teil-Automatisierung der Arbeitsschritte genauer beschrieben werden.

Für die Segmentierung der mittelniederdeutschen Lemmata wird ein Skript verwendet,[8] das Präfixe und Suffixe formal identifiziert und abtrennt (vgl. Tab. 3 in Abschn. 3.2).[9] In das Skript werden dabei zusätzlich für einzelne Suffixe zu vermeidende Fehltreffer eingegeben, die formal wie das Suffix enden, bei denen das jeweilige Suffix aber nicht vorliegt und deshalb keine Segmentierung durchgeführt werden soll (zum Beispiel ist *dēgedinge* keine Bildung mit *-inge* und sind Komposita mit *-stê^{i}n* keine Bildungen mit *-in* usw.). In einem iterativen Verfahren werden im Zuge der Durchsicht der Vorschläge weitere solcher Fehltreffer identifiziert und von der Segmentierung ausgeschlossen. Auf der Basis der formal identifizierten Affixe und der übrigbleibenden Stämme wird zudem automatisch ein Vorschlag für eine nicht-hierarchische Strukturformel erstellt,[10] d. h., die Reihenfolge, in der sich die jeweiligen Konstituenten zusammenfügen, wird als Formel dargestellt, z. B. für das oben erwähnte mnd. *be-grāv-en* die Formel *p|wV*. Für alle Lemmata, bei denen im automatischen Prozess keine Segmentierung erfolgt (d. h., die als reine Stämme aufgefasst werden), wird zunächst manuell geprüft, ob es sich tatsächlich um Simplizia handelt. Die daraus resultierende Simplizialiste wird wiederum verwendet, um – analog zum Vorgehen bei den Affixen – für diejenigen Lemmata, bei denen keine Segmentierung der Komposition aus den Vorarbeiten vorliegt, Segmentierungsvorschläge zu produzieren. Auch fließt die Wort-

8 Alle im Rahmen von WoDia entstandenen Skripte sollen zu einem späteren Zeitpunkt veröffentlicht werden.

9 Die Affixlisten basieren auf den Vorarbeiten von Dieter Möhn und Ingrid Schröder zur Wortbildung des Mittelniederdeutschen (vgl. z. B. zur Adjektivbildung Möhn/Schröder 2009), die auch in eine geplante korpusbasierte mittelniederdeutsche Grammatik einfließen sollen (vgl. Ihden/Schröder 2021). Für weitere Vorarbeiten zur mittelniederdeutschen Wortbildung, die Wörterbuch- und/oder Korpusdaten einbeziehen, vgl. Ihden/Schröder (2024) sowie Schröder (2022).

10 Für bestimmte Lemmata, die Kompositionsprodukte sind, kann zusätzlich auf Vorarbeiten aus dem ReN-Projekt zurückgegriffen werden, da diese dort bereits manuell segmentiert wurden.

artangabe der Simplizia direkt in den Strukturformel-Vorschlag ein. So werden schrittweise neue Teilmengen des Wortschatzes segmentiert und analysiert, wodurch die Skripte jeweils erweitert und verbessert werden können.

Ein weiterer Schritt, bei dem regelbasierte Skripte zum Einsatz kommen, ist die Verknüpfung des Wortschatzes; beispielhaft wird hier das Vorgehen für die Verknüpfung von Mittelhochdeutsch und Mittelniederdeutsch skizziert.[11] Ein Skript produziert auch hier Vorschläge, die anschließend zu überprüfen sind. Die Verknüpfungsrichtung verläuft dabei vom Hochdeutschen zum Niederdeutschen, da die lautlichen Korrespondenzen in diese Richtung überwiegend eindeutig sind.[12] Auf der Grundlage von knapp 70 Ersetzungsregeln, die für hochdeutsche Laute (bzw. Grapheme) die niederdeutschen Entsprechungen aufführen, erstellt das Skript für jedes mittelhochdeutsche Lemma eine ‚niederdeutsche' Form. Beispielsweise wird mhd. <zz> grundsätzlich zu mnd. <t> verändert, wodurch für mhd. *wazzer* mnd. *water* generiert wird. Diese Ersetzungsregeln sind je nach Laut (bzw. Graphem) unterschiedlich komplex und beziehen in einigen Fällen die Position im Wort sowie die folgenden Laute mit ein. Ausgenommen von diesen allgemeinen Lautkorrespondenzen werden Affixe, die systematisch durch ihre mittelniederdeutsche Entsprechung ersetzt werden, da die Lautkorrespondenzen hier nicht greifen. Schließlich werden einige Sonderregeln angewandt, wie z. B., dass mhd. *alt* mnd. *ōlt* entspricht oder dass die Korrespondenz von mhd. <ou> zu mnd. <ô> im Falle des Lexems *vrouwe* (mhd. und mnd.) nicht besteht.

Das Skript gleicht die künstlich hergestellten niederdeutschen Formen mit der mittelniederdeutschen Lemmaliste ab. Existiert eine Form in der mittelniederdeutschen Lemmaliste, wird diese als Verknüpfungsvorschlag zum mittelhochdeutschen Lemma ausgegeben.[13] Um weiter einzugrenzen, ob es sich um eine Entsprechung handelt, wird außerdem auf beiden Seiten die Wortart der Lemmata mit einbezogen. Werden für ein Lemma mehrere mögliche Verknüpfungen gefunden – beispielsweise bei Homonymen –, werden alle Möglichkeiten vorgeschlagen und bei der anschließenden Durchsicht wird eine Entscheidung getroffen.

Derzeit liegen bereits für knapp 15% der mittelniederdeutschen Lemmata Verknüpfungsvorschläge vor, wobei zu berücksichtigen ist, dass auf mittelniederdeutscher Seite eine noch unvollständige Lemmaliste zum Einsatz kommt, der zudem die Angaben zur Wortart fehlen. Daher ist mit einer deutlichen Verbesserung dieser Bilanz zu rechnen, sobald die mittelniederdeutsche Lemmaliste komplett vorliegt und auf beiden Seiten die Wortart mit einbezogen werden kann. Da es zudem nur eine Teilüberschneidung beider Wortschätze gibt, stellt dieser Anteil automatischer Verknüpfungsvorschläge bereits einen enormen Mehrwert für das Projekt dar. Neben ihrem übergeordneten Zweck, varietätenübergreifende Studien zu ermöglichen, erleichtert die Verknüpfung der verschiedenen Wortschätze die Arbeit im Projekt, da, wie erwähnt, bei der Bearbeitung des Mittelniederdeutschen für

11 Hilfreich waren hierfür insbesondere die Vorüberlegungen von Thomas Klein (2013, Folie 25–36) zur Verknüpfung der mittelhochdeutschen und mittelniederdeutschen Lemmalisten.

12 Vgl. die Übersicht über die lautlichen Korrespondenzen, zur Verfügung gestellt von Lourens Visser unter https://osf.io/yvfjh (Stand: 1.7.2025).

13 Weitere Angleichungen sind nötig, um unterschiedlich gesetzten Klammerungen in beiden Wörterbüchern zu begegnen. Enthält ein Lemma, für das so keine Entsprechung im Mittelniederdeutschen gefunden werden kann, Klammern, wird entweder eine Form ohne die Klammern generiert (so wird für mhd. *acker(e)n* > *ackeren* mnd. *ackeren* gefunden), oder es wird eine Form ohne Klammern und ohne Klammerinhalt erstellt (mhd. *ê(we)brëchunge* > *êbrëchunge* führt zu mnd. *êbrēkinge*). Ebenso werden auf mittelniederdeutscher Seite Formen ohne Klammern bzw. ohne Klammern und ohne Klammerinhalt generiert (so kann mnd. *ping(e)stâvent* zu mhd. *pfingestâbend* und mnd. *bû(w)knecht* zu mhd. *bûknëht* gefunden werden).

das Alt- oder Mittelhochdeutsche bereits angefertigte Segmentierungen und Strukturformeln übernommen werden können.

3.5 Technische Umsetzung der digitalen Forschungsumgebung

Parallel zu den skizzierten Arbeitspaketen läuft die Konzeption, Entwicklung und Erweiterung der digitalen Forschungsumgebung. Dazu zählen unter anderem die Modellierung der Datenbank sowie die dafür notwendige Datenintegration aus den unterschiedlichen vorliegenden Quellen. Für das Projektteam wird eine Bearbeitungsschnittstelle bereitgestellt. Daneben wird das Frontend der Forschungsumgebung gestaltet und es werden Anwendungskonzepte entwickelt, die verschiedene Ausgabe-, Export- und Visualisierungstypen auf der vereinheitlichten Datengrundlage ermöglichen. So sollen in der Forschungsumgebung komplexe Suchabfragen anhand intuitiv bedienbarer Steuerelemente zusammengestellt sowie dem individuellen Forschungsinteresse entsprechend die Ergebnisse ausgegeben werden können. Die Ergebnislisten lassen sich z. B. alphabetisch nach Lemma oder Wortfamilie sowie nach der Strukturformel sortieren.

4. Anwendungsbeispiele

Die beschriebenen Strukturformeln eröffnen neue Zugänge für die historische Wortbildungsforschung. Drei exemplarische Untersuchungsfelder sollen dies verdeutlichen und die Analysemöglichkeiten von WoDia aufzeigen. Die beiden Beispiele für den hochdeutschen Sprachraum befassen sich mit den doppelsuffigierten Adjektiven (vgl. Abschn. 4.1) und der Suffixablösung bei den Nomina agentis (vgl. Abschn. 4.2.1). Für eine Analyse im Niederdeutschen wird ein Blick auf funktionsgleiche Movierungssuffixe im Mittelniederdeutschen geworfen (vgl. Abschn. 4.2.2). Beide sind klassische Themenfelder der historischen Morphologie (vgl. Wilmanns 1899; Erben 2006) und erfahren durch die computergestützte Analyse in WoDia eine neue empirische Tiefenschärfe sowie zum ersten Mal eine vergleichbare systematische Aufbereitung.

4.1 Doppelsuffigierte Adjektive

Die Doppelsuffigierung von abgeleiteten Adjektiven wird in der historischen Wortbildungslehre des Deutschen vielfach kritisch betrachtet. Gemeint sind Bildungen auf (nhd.) *-haft-ig*, *-ig-lich* usw. (*wahrhaftig*, *gemeiniglich* usw.), die sich bereits im Althochdeutschen finden. Die älteren Wortbildungslehren sprechen hier von Bildungen, in denen das zweite Suffix zum Teil als „blosse Wucherung“ ohne Bedeutungsunterschied erscheine (Wilmanns 1899, S. 490), und ähnlich wird das aus doppelt suffigierten (ahd.) *-ig-lîh-*/(mhd.) *-ec-lich*-Bildungen als eigenes Suffix entwickelte mhd. *-eclich* (bei Bildungen ohne entsprechendes Adjektiv auf einfaches *-ec*) als „Wuchersuffix“ bezeichnet (Henzen 1965, S. 203, 274; KSW III, S. 313, § A 105).

Splett (1995) hat gezeigt, wie man die Behandlung dieser Erscheinung fürs Althochdeutsche auf eine breite und systematisch erhobene Materialgrundlage stellen kann, nämlich durch Nutzung der Strukturformeln seines Wörterbuchs. In diesem Fall können mit der Suche nach der Formel ((…)*sA*)*sA* (Stamm + Adjektivsuffix + Adjektivsuffix) 140 doppelsuffigierte althochdeutsche Adjektive erhoben werden, in denen von den 41 althochdeutschen Adjektivsuffixen 28 in erster oder zweiter Position oder in beiden vorkommen können.

Zu 89 von 140 dieser doppelsuffigierten Bildungen ist ein mit dem an Erstposition stehenden Suffix einfach suffigiertes Adjektiv bezeugt (z. B. *lîb-haft* neben *lîb-haft-îg*). Diese 89 Wortpaare liefern ein reichhaltiges Material für die genauere Untersuchung der Funktionen der Doppelpräfigierung am althochdeutschen Adjektiv. Entsprechende Materialzusammenstellungen werden in WoDia dann auch für Altsächsisch, Mittelhochdeutsch und Mittelniederdeutsch und den diachronen Vergleich möglich sein.

4.2 Persönliche Agentiva

4.2.1 Suffixablösung bei den Nomina agentis

Wie die Doppelaffigierung von Adjektiven hat auch die Suffixablösung bei den persönlichen Nomina agentis im Althochdeutschen und Mittelhochdeutschen von jeher große Aufmerksamkeit der historischen Wortbildungslehre gefunden (vgl. Erben 2006, S. 150–155). Bei diesem Wortbildungswandel wird der ältere germanische Typus der Ableitung von Nomina agentis durch das *-an-*/*-jan-*Suffix der schwachen Maskulina (z. B. ahd. *gebo* swM. ‚Geber, Spender, Wohltäter') durch den jüngeren mit dem (bereits voralthochdeutsch gebrauchten) Lehnsuffix aus lat. *-ārius*, ahd. *-âri*, mhd. *-ære*/*-er*, nhd. *-er* abgelöst (z. B. *gebâri*, mhd. *geber*, nhd. *Geber* stM.). Die entsprechenden femininen Nomina agentis wurden ursprünglich ebenfalls als schwache (*-ōn-*, *-jōn-*) Stämme gebildet (z. B. ahd. *becka* swF. ‚Bäckerin'). Die Nebensilbenabschwächung zum Mittelhochdeutschen hin hat dann der deutlicheren neuen Bildungsweise mit dem *-āri*/*-ære*-Suffix zum Durchbruch verholfen, wobei die Feminina (im Hochdeutschen) jetzt durch Anfügung des Movierungssuffixes ahd. *-in*, mhd. *-inne*, nhd. *-in* an das Lehnsuffix gebildet wurden (nhd. *Bäcker*, *Bäckerin*).

Charakteristisch für den alten Bildungstyp ist, dass er (schon germanisch) vor allem in Rektionskomposita (vgl. Krahe/Meid 1969, S. 26) bzw. Zusammenbildungen (vgl. Henzen 1965, S. 61) des Typs ahd. *brôt-becko*, *gast-geba* usw. vorkommt, wobei das Zweitglied häufig gar nicht selbstständig belegt ist (wie ahd. im Falle von **becko* swM. und **geba* swF.) und vielleicht primär gar nicht frei gebraucht wurde. Als Erstglied kommen vor allem Substantive vor, daneben auch Adjektive wie in *wîs-sago* swM. ‚Wahrsager', *blint-slîcho* swM. ‚Blindschleiche' (Henzen 1965, S. 69). Die Nomina agentis im Zweitglied können nicht nur von Verben abgeleitet sein, sondern auch von Substantiven wie z. B. in den *jan*-stämmigen Bildungen ahd. *scerio* swM. ‚Scharmeister' (nhd. *Scherge*) und *sidilo* swM. ‚Einwohner' zu ahd. *scara* ‚Schar', *sidil* ‚Siedel, Sitz' (Henzen 1965, S. 133).

Dieser Wortbildungswandel verdiente eine erneute Untersuchung auf der umfassenden Materialgrundlage des heutigen Stands der Lexikographie des älteren Deutsch, insbesondere des Althochdeutschen. Die Materialerhebung zu den Bildungen älteren Typs wird dabei durch Spletts Strukturformeln stark erleichtert. Die Suche in ihnen zu den deverbalen Bildungen ergibt

- für die Formel (*wV*)*San* der maskulinen einfachen Bildungen 87 Treffer,
- für die Formel (*wV*)*Sôn* der femininen einfachen Bildungen 45 Treffer,
- für die Formeln (*wS*)((*wV*)*San*) bzw. ((*wS*)+(*wV*))*San* der maskulinen Komposita bzw. Zusammenbildungen mit substantivischem Erstglied 39 bzw. 78 Treffer,
- für die Formeln (*wS*)((*wV*)*Sôn*) bzw. ((*wS*)+(*wV*))*Sôn* der femininen Komposita bzw. Zusammenbildungen mit substantivischem Erstglied 8 bzw. 17 Treffer.

Abbildung 1 zeigt exemplarisch einen Ausschnitt des Suchergebnisses auf AWB^{Spl} Online für die Formel *(wV)San*.

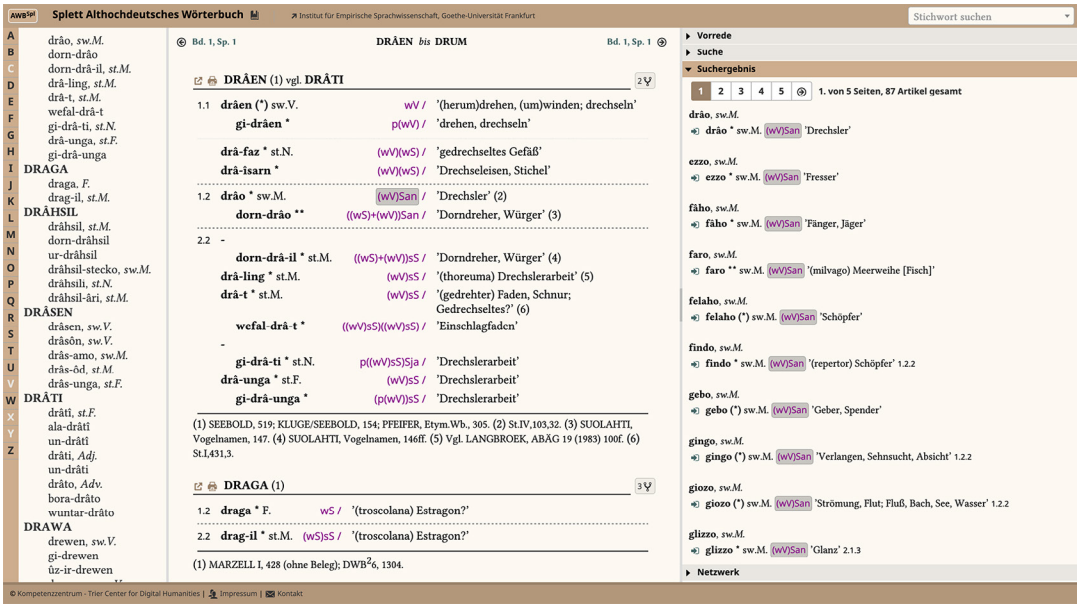

Abb. 1 : Screenshot AWB^{Spl} (2009) Online im Wörterbuchnetz (interner Prototyp)

Entsprechend kann nach desubstantivischen Nomina agentis und nach Bildungen mit deadjektivischen Erstgliedern gesucht werden. Die Trefferlisten sind natürlich einer genauen Kontrolle zu unterziehen und ggf. zu bereinigen, z. B. um die nicht-persönlichen Bildungen (Nomina instrumenti, Tier- und Pflanzennamen usw.). Außerdem sind unter Umständen Formeln für besondere weitere Bildungsweisen einzubeziehen. Für eine nähere Untersuchung könnten in einer Pilotstudie die Strukturformeln um die Unterscheidung der persönlichen und nicht-persönlichen Bildungen angereichert werden. Solche Zusatzinformationen können später in der Forschungsumgebung als optionales Angebot zur Verfügung gestellt werden.

Für den Vergleich mit dem Mittelhochdeutschen und Mittelniederdeutschen ergibt sich durch die Nebensilbenabschwächung die Schwierigkeit des Zusammenfalls verschiedener älterer Suffixe. Im Falle der hier interessierenden agentiven schwachen Maskulina und Feminina des besprochenen alten Bildungstyps werden die ahd. Ausgänge *-o* und *-a* zu mhd. *-e* abgeschwächt. Das mhd. Suffix *-e* vereinigt nach Ring (2008, S. 40–118) und KSW III (2009, S. 61–72) mindestens acht verschiedene Funktionsklassen, von denen die deverbalen Agentiva nur eine ausmachen (Ring *-e*^{3}, KSW III *-e*^{4}), die zudem wiederum auch nicht-persönliche Agentiva enthält. Hierzu könnte ähnlich wie bei den personalen und nicht-personalen Nomina agentis des Althochdeutschen in einer Detailstudie die genauere Unterscheidung der Funktionsklassen in die Formeln eingebracht und in der Forschungsumgebung zusätzlich zur Verfügung gestellt werden.

4.2.2 Funktionsgleiche Movierungssuffixe

Das Mittelniederdeutsche weist in mehreren Bereichen Affixkonvergenz auf, d. h., es gibt mehrere formal unterschiedliche Affixe, die (in etwa) die gleiche Funktion erfüllen (vgl. Barz 2021). Am Beispiel der frequentesten Movierungssuffixe soll gezeigt werden, wie WoDia für Untersuchungen solcher Konkurrenzkandidaten eingesetzt werden kann.

Im Neuhochdeutschen wird die Femininmovierung durch das Derivationssuffix *-in* dominiert, das z. B. aus dem *Koch* eine *Köchin* macht. Weitere Suffixe treten nur sehr eingeschränkt auf (vgl. Fleischer/Barz 2012, S. 239). In den älteren Sprachstufen des Deutschen finden sich hingegen mehrere Derivationssuffixe, die nebeneinander stehen.

Im Altsächsischen sind die Suffixe *-in* und *-inne* belegt, die an deverbale Nomina agentis gehängt werden, z. B. *mak-en* mit der Formel *wV* > *mak-āri* mit der Formel (*wV*)*sS* > *mak-ār-in* mit der Formel ((*wV*)*sS*)*sS* (vgl. Lasch 1914, § 213, 372; Hucko 1904, S. 72 f.; Wilmanns 1899, S. 310–312). Die Distinktion zwischen der kürzeren Form *-in* und der längeren Form *-inne* ist hierbei auf eine morphophonologische Differenzierung zwischen dem Nominativ und den obliquen Kasus zurückzuführen. Ein weiteres Movierungssuffix ist *-sche*, dessen Ursprung auf lat. *-issa* bzw. frz. *-esse* zurückgeht.

Im Mittelniederdeutschen sind die beiden frequentesten Movierungssuffixe *-inne* und *-sche* (vgl. Seelmann 1921).[14] Werth (2022, S. 3, 2015, S. 61) stellt in norddeutschen Hexenverhörprotokollen des 16. und 17. Jahrhunderts eine funktionale Differenzierung dieser beiden fest, wobei *-inne* das Lemma semantisch um das Merkmal [+weiblich] erweitert und *-sche* eine onymische Zugehörigkeitsrelation (‚Ehefrau von XName') etabliert.

Für eine umfassende Untersuchung lässt sich mit Hilfe von WoDia eine Trefferliste aller Movierungen anzeigen.[15] Das Suchergebnis in WoDia könnte wie in Abbildung 2 aussehen.

Abb. 2: Suche in der WoDia-Forschungsumgebung am Beispiel der Movierung[16]

Durch die Auszeichnung der Wortbildungsstruktur kann direkt nach den beiden Suffixen gesucht werden, formgleich endende Simplizia müssen nicht mehr aussortiert werden. Die dargestellte Abfrage liefert 126 Lemmata mit *-inne* und 165 Lemmata mit *-sche*. Bei jeweils dreien davon liegt die Kombination *-ster-inne* bzw. *-ster-sche* vor. Nomina agentis auf *-er* liegen bei 50 der *-inne*-Bildungen und bei 121 *-sche*-Bildungen zugrunde.[17] Insgesamt 30

14 Außerdem finden sich die weniger frequente, kurze Form *-in* sowie, ebenfalls weit weniger verbreitet, das aus dem Mittelniederländischen stammende *-(e)ster* (auch in Kombination mit *-inne* oder mit *-sche*) (vgl. Seelmann 1921). Nur bei Familiennamen dienen überdies *-s* und *-sen* zur weiblichen Sexusmarkierung (vgl. Werth 2015, S. 57–58).

15 So wird die Suche nach Lemmata, die auf *-sche* oder *-inne* enden und deren Strukturformel auf *sS* (= Substantivsuffix) ausgeht, möglich sein.

16 Es handelt sich um einen ersten Entwurf für das Design der Forschungsumgebung, das im weiteren Verlauf der Projektarbeiten noch bearbeitet wird.

17 Die Zahlen basieren auf einem präfinalen Stand der Lemmaliste, bei der neben der im MNWB noch nicht bearbeiteten Strecke einzelne weitere Lemmata fehlen können, da Nestartikel sowie bestimmte Verweise in der Auszeichnung aktuell noch nicht korrekt erfasst sind. In der finalen Fassung der Datenbank können die Trefferzahlen daher abweichen.

verschiedene Stämme liegen sowohl mit *-inne* als auch mit *-sche* vor (vgl. *kōkinne*/*kōksche* ‚Köchin' in Abb. 2). Die in Abbildung 2 ausgewählten Beispiele gehören (ihrer Bedeutungsangabe im MNWB nach) alle zum ersten Typ, bei dem die Basis um das Merkmal [+weiblich] ergänzt wird. Da im Wörterbuch Eigennamen nur in Ausnahmefällen aufgenommen sind, lässt sich die von Werth als onymischer Typ bezeichnete Movierungsform vor allem in Kombinationen mit Nomina agentis wie Berufsbezeichnungen, z. B. *vinstermēkersche* ‚Frau des Fenstermachers', finden. Eine eingehende Untersuchung der Movierung unter Nutzung der Wörterbuchdaten einerseits (Wortbildungsstruktur, Basen, Verbindung mit *-er*) und des über die Lemmaliste verknüpften Referenzkorpus Mittelniederdeutsch/Niederrheinisch (räumliche und zeitliche Verbreitung, kontextuelle Einflüsse) andererseits ist derzeit als Fallstudie in Vorbereitung.

5. Fazit

Im Projekt ‚Wortfamilien diachron' (WoDia) entsteht eine digitale Forschungsumgebung, die für Untersuchungen zur historischen Wortbildung des Deutschen eine umfassende Materialbasis zur Verfügung stellt. Sie bietet für alle Lemmata der Referenzwörterbücher des Althochdeutschen, Altsächsischen, Mittelhochdeutschen und Mittelniederdeutschen eine wortbildungsmorphologische Analyse und Wortfamilienzuordnung, die auf der Segmentierung der Lemmata und der Rückführung der dabei freigelegten Stämme auf ihren Wortfamilienstamm beruht. Das Ergebnis der Analyse wird in einer hierarchischen Strukturformel angegeben, in der die Konstituentenstruktur, die Wortart der Stämme und die Wortbildungsart notiert sind. Um varietätenübergreifende Analysen zu ermöglichen, sind die vier Wortschätze sowohl auf der Ebene der Einzellemmata als auch über die Wortfamilien miteinander verbunden. Alle Lemmata sind mit den entsprechenden Artikeln in ihren Referenzwörterbüchern verknüpft, darüber hinaus werden sie mit den Referenzkorpora zur deutschen Sprachgeschichte verlinkt.

Das Potenzial der in WoDia aufbereiteten Daten wurde am Beispiel dreier wortbildungsmorphologischer Phänomene in verschiedenen Sprachstufen verdeutlicht, der Doppelsuffigierung von Adjektiven, der Suffixablösung bei den Nomina agentis und der funktionsgleichen Movierungssuffixe. Über die Suche nach Strukturformeln eines bestimmten Typs in der Forschungsumgebung können die betreffenden Lemmata zusammengestellt werden. Je nach Phänomen und konkreter Forschungsfrage sind weitere Annotationen anzuschließen, z. B. im Fall der Nomina agentis zur Differenzierung zwischen persönlichen und nichtpersönlichen Bildungen oder bei den Movierungssuffixen zur Unterscheidung verschiedener Funktionstypen. WoDia ermöglicht Detailstudien nicht nur zu Wortbildungsstrukturen, Wortbildungsmitteln und deren Wandel im älteren Deutschen, sondern auch zu Veränderungen im Bereich der Wortfamilien, beispielsweise durch einen diachronen Ausbau bestimmter Wortfamilien. Die Verknüpfung zwischen den vier Wortschätzen erlaubt zudem varietätenübergreifende Untersuchungen, sodass im Ergebnis ein umfassenderes Bild der Entwicklungen im Bereich des Wortschatzes und der Wortbildung im historischen Deutschen gezeichnet werden kann.

Literatur

Wörterbücher

ASWB = Tiefenbach, Heinrich (2010): Altsächsisches Handwörterbuch. Berlin/New York: De Gruyter.

AWB = Sächsische Akademie der Wissenschaften zu Leipzig (2021): Althochdeutsches Wörterbuch. Auf Grund der von Elias von Steinmeyer hinterlassenen Sammlungen im Auftrag der Sächsischen Akademie der Wissenschaften zu Leipzig. Bd. 8,1: S – Sn-. Berlin/Boston: De Gruyter.

AWB^{Spl} = Splett, Jochen (1993): Althochdeutsches Wörterbuch. Analyse der Wortfamilienstrukturen des Althochdeutschen. Zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. Bd. 1–3. Berlin/New York: De Gruyter.

Lexer = Lexer, Matthias (1992): Mittelhochdeutsches Handwörterbuch. Zugleich als Supplement und alphabetischer Index zum Mittelhochdeutschen Wörterbuche von Benecke-Müller-Zarncke. Bd. 1–3. Stuttgart: Hirzel. [Nachdruck der Ausgabe Leipzig 1872–1878].

LW = Lübben, August/Walther, Christoph (1888): Mittelniederdeutsches Handwörterbuch. (= Wörterbücher 2). Leipzig: Soltau.

MNWB = Lasch, Agathe/Borchling, Conrad (1956–): Mittelniederdeutsches Handwörterbuch. Fortgef. von Gerhard Cordes und Dieter Möhn. Hg. von Ingrid Schröder. Bd. 1 ff. Zuletzt Bd. 3, Lfg. 44., 2023. Kiel/Hamburg: Wachholtz.

MWB = Akademie der Wissenschaften und der Literatur | Mainz und der Niedersächsischen Akademie der Wissenschaften zu Göttingen (2006–): Mittelhochdeutsches Wörterbuch. Zuletzt Bd. 3, Lfg. 3, 2024. Stuttgart: Hirzel.

Splett, Jochen (2009): Deutsches Wortfamilienwörterbuch. Analyse der Wortfamilienstrukturen der deutschen Gegenwartssprache. Zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. Bd. 1–18. Berlin/New York: De Gruyter.

Forschungsliteratur

Barz, Irmhild (2021): Affixkonvergenz. www.degruyterbrill.com/database/WSK/entry/wsk_id_wsk_artikel_artikel_17524/html (Stand: 1.7.2025).

Duden (2016): Der Duden in zwölf Bänden. Bd. 4: Die Grammatik. Unentbehrlich für richtiges Deutsch. 9., vollst. überarb. u. aktual. Aufl. Berlin: Dudenverlag.

Duden (2022): Der Duden in zwölf Bänden. Bd. 4: Die Grammatik. Unentbehrlich für richtiges Deutsch. 10., völlig neu verf. Aufl. Berlin: Dudenverlag.

Eisenberg, Peter (2020): Grundriss der deutschen Grammatik. Das Wort. 5., aktual. und überarb. Aufl. Stuttgart: Metzler.

Erben, Johannes (2006): Einführung in die deutsche Wortbildungslehre. (= Grundlagen der Germanistik 17). 5., durchges. u. erg. Aufl. Berlin: ESV.

Fleischer, Wolfgang/Barz, Irmhild (2012): Wortbildung der deutschen Gegenwartssprache. 4., völlig neu bearb. Aufl. Berlin/Boston: De Gruyter.

Henzen, Walter (1965): Deutsche Wortbildung. 3., durchges. u. erg. Aufl. (= Sammlung kurzer Grammatiken germanischer Dialekte B 5). Tübingen: Niemeyer.

Hucko, Matthias (1904): Bildung der Substantiva durch Ableitung und Zusammensetzung im Altsächsischen. Straßburg: DuMont-Schauberg.

Ihden, Sarah/Schröder, Ingrid (2021): Mittelniederdeutsche Grammatik: Konzeption und erste Analysen. In: Jahrbuch des Vereins für niederdeutsche Sprachforschung 144, S. 79–104.

Ihden, Sarah/Schröder, Ingrid (2024): Das Referenzkorpus Mittelniederdeutsch/Niederrheinisch als Quelle für wortbildungsmorphologische Studien. Zur Varianz des Suffixes *-schop*. In: Coniglio, Marco/Recker, Anabel/Sahm, Heike (Hg.): Mittelniederdeutsch zwischen Korpuslinguistik und Literaturwissenschaft. Göttingen: Universitätsverlag, S. 15–43.

Ihden, Sarah/Schnelle, Gohar/Schröder, Ingrid/Zeige, Lars E. (2023): Der Verbund ‚Deutsch Diachron Digital – Referenzkorpora zur deutschen Sprachgeschichte'. Strategien der Erschließung, Analyse und nachhaltigen Nutzung historischer Sprachdaten. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache/Corpus Linguistics and Interdisciplinary Perspectives on Language 11). Tübingen: Narr, S. 11–31.

Klein, Thomas (2013): Verknüpfung digitaler Lemmalisten historischer Sprachstufen des Deutschen – wie und wozu? www.uni-trier.de/fileadmin/forschung/maw/MWB/Arbeitsgespraech2013/Vortrag_Bullay_Klein.pdf (Stand: 1.7.2025). [Präsentationsfolien des Vortrags im Rahmen des Arbeitsgesprächs zur historischen Lexikographie vom 12. bis 14. April 2013 in Bullay, 13.4.2013].

Klein, Thomas (2018): Mittelhochdeutsche Wortfamilien: Ermittlung und Perspektiven. In: Zeitschrift für Wortbildung 2, 1 (Sonderheft: Wortbildung – historisch, mehrsprachig, kontrastiv), S. 11–31. https://journals.linguistik.de/zwjw/article/view/18/103 (Stand: 8.2.2026).

Krahe, Hans (1969): Germanische Sprachwissenschaft. Bd. 3: Wortbildungslehre. 7. Aufl. (= Sammlung Göschen 2234). Berlin: De Gruyter.

KSW III = Klein, Thomas/Solms, Hans-Joachim/Wegera, Klaus-Peter (2009): Mittelhochdeutsche Grammatik. Bd. 3: Wortbildung. Tübingen: Niemeyer.

Möhn, Dieter/Schröder, Ingrid (2009): Lexembildung im Aufriss einer Grammatik des Mittelniederdeutschen. Das Adjektiv als Exempel. In: Lenz, Alexandra N./Gooskens, Charlotte/Reker, Simon (Hg.): Low Saxon Dialects across Borders – Niedersächsische Dialekte über Grenzen hinweg. (= Zeitschrift für Dialektologie und Linguistik Beiheft 138). Stuttgart: Steiner, S. 38–59.

Plate, Ralf (2020): Computergestützte Etablierung epochenübergreifender Wortfamilienstrukturen. Abschlussbericht. https://zhistlex.de/ziele/wortfamilien/ (Stand: 1.7.2025).

Plate, Ralf (2022): Word families in diachrony. An epoch-spanning structure for the word families of older German. In: Klosa-Kückelhaus, Anette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (Hg.): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. 12–16 July 2022, Mannheim, Germany. Mannheim: Leibniz-Institut für Deutsche Sprache, S. 605–613. https://doi.org/10.14618/ids-pub-11253 (Stand: 1.7.2025).

Ring, Uli (2008): Substantivderivation in der Urkundensprache des 13. Jahrhunderts. Eine historisch-synchrone Untersuchung anhand der ältesten deutschsprachigen Originalurkunden. (= Studia Linguistica Germanica 96). Berlin: De Gruyter.

Schröder, Ingrid (2022): Stand und Perspektiven mittelniederdeutscher Grammatikographie am Beispiel der expliziten Derivation. In: Denkler, Markus/Mähl, Stefan (Hg.): Beiträge zur historischen Wortbildung des Niederdeutschen. (= Niederdeutsche Studien 61). Wien/Köln: Böhlau, S. 13–36.

Seelmann, Wilhelm (1921): Mittelniederländische Wörter in der Mark Brandenburg. In: Jahrbuch des Vereins für niederdeutsche Sprachforschung 47, S. 40–45.

Splett, Jochen (1995): Der Adjektivtyp ((…)sA)sA) im Althochdeutschen. In: Cajot, José/Kremer, Ludger/Niebaum, Hermann (Hg.): Lingua theodisca. Bd. 1: Beiträge zur Sprach- und Literaturwissenschaft. Jan Goossens zum 65. Geburtstag. (= Niederlande-Studien 16.1). Münster u. a.: Lit, S. 89–101.

Wilmanns, Wilhelm (1899): Deutsche Grammatik. Gotisch, Alt-, Mittel- und Neuhochdeutsch. 2. Aufl. Bd. 2: Wortbildung. Straßburg: Trübner.

Werth, Alexander (2015): *Gretie Dwengers*, genannt *die Dwengersche*. Formale und funktionale Aspekte morphologischer Sexusmarkierung (Movierung) in norddeutschen Hexenverhörprotokollen der Frühen Neuzeit. In: Jahrbuch des Vereins für niederdeutsche Sprachforschung 138, S. 53–75.

Werth, Alexander (2022): Der Abbau der onymischen Movierung im 18. Jahrhundert: Eine Auswertung des Deutschen Textarchivs. In: Havinga, Anna D./Lindner-Bornemann, Bettina (Hg.): Deutscher Sprachgebrauch im 18. Jahrhundert: Sprachmentalität, Sprachwirklichkeit, Sprachreichtum. (= Germanistische Bibliothek 71). Heidelberg: Winter, S. 93–113.

Online-Ressourcen

Deutsch Diachron Digital (o. D.): Referenzkorpora zur deutschen Sprachgeschichte. www.deutschdiachrondigital.de/ (Stand: 1.7.2025).

MWB Online (o. D.): Mittelhochdeutsches Wörterbuch. www.mhdwb-online.de/ (Stand: 1.7.2025).

Sächsische Akademie der Wissenschaften zu Leipzig (o. D.): Althochdeutsches Wörterbuch. https://awb.saw-leipzig.de/AWB (Stand: 1.7.2025).

Trier Center for Digital Humanities (o. D.): Wörterbuchnetz. Kompetenzzentrum. www.woerterbuchnetz.de (Stand: 1.7.2025).

Visser, Lourens (o. D.): Lautkorrespondenzen MHD-MND. https://osf.io/yvfjh (Stand: 1.7.2025).

Kontaktinformation

Thomas Burch
Kompetenzzentrum – Trier Center for Digital Humanities
Universität Trier
54286 Trier
E-Mail: burch@uni-trier.de

Jost Gippert
Centre for the Study of Manuscript Cultures
University of Hamburg
Warburgstr. 26
20354 Hamburg
E-Mail: jost.gippert@uni-hamburg.de

Sarah Ihden
Institut für Germanistik
Universität Rostock
Universitätsplatz 3
18055 Rostock
E-Mail: sarah.ihden@uni-rostock.de

Sarah Kwekkeboom
DFG-Projekt ‚Wortfamilien diachron‘
Institut für Empirische Sprachwissenschaft
Goethe-Universität
Rostocker Str. 2
60323 Frankfurt a. M.
E-Mail: sarah.kwekkeboom@adwmainz.de

Ralf Plate
Mittelhochdeutsches Wörterbuch,
Arbeitstelle der Akademie der Wissenschaften und der Literatur | Mainz
Universität Trier
Universitätsring 15
54286 Trier
E-Mail: plate@uni-trier.de

Lena Schnee
DFG-Projekt ‚Wortfamilien diachron'
Institut für Germanistik
Universität Hamburg
Von-Melle-Park 6
20146 Hamburg
E-Mail: lena.schnee@uni-hamburg.de

Ingrid Schröder
Institut für Germanistik
Universität Hamburg
Von-Melle-Park 6
20146 Hamburg
E-Mail: ingrid.schroeder@uni-hamburg.de

Roland Schuhmann
DFG-Projekt ‚Wortfamilien diachron'
Institut für Empirische Sprachwissenschaft
Goethe-Universität
Rostocker Str. 2
60323 Frankfurt a. M.
E-Mail: schuhmann@em.uni-frankfurt.de

Johanna Wittmack
DFG-Projekt ‚Wortfamilien diachron'
Institut für Germanistik
Universität Rostock
Universitätsplatz 3
18055 Rostock
E-Mail: johanna.wittmack@uni-rostock.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.79

Jenia Yudytska / Jannis Androutsopoulos

Reddit Corpus Keyword Search (ReCKS)

Ein Tool für die Gewinnung und Auswertung von Sprachdaten aus der Social Media-Plattform Reddit

Abstract ReCKS ("Reddit Corpus Keyword Search") is a web application for the linguistic research of Reddit comments. The current underlying dataset comes from the largest German-language subreddit, r/de, and includes user comments from 2006 to 2023 with a total of ca. 41 million tokens. As input, ReCKS allows both simple fixed keyword searches and complex search queries using regular expressions (RegEx). The output is given in the form of an exportable online table and a diagram that visualises the normalised frequency of the search term per year. This paper first explains the technical architecture of the application. It then briefly describes various usage scenarios and discusses in detail how the tool can be used for microdiachronic analyses. This is illustrated with an analysis of *Genderzeichen* ('gender signs', i. e., spelling variants that index gender-inclusivity, such as *Student:in* or *Student*in*) by r/de users over the last 15 years.

Keywords ReCKS, Reddit, user comments, RegEx, natively digital language corpora, microdiachronic analysis of digitally written language

1. Einleitung

Die Plattform Reddit zählt zu den meistbesuchten Websites weltweit (Semrush 2025) und ist in der Computerlinguistik eine etablierte Datenquelle (vgl. Blombach et al. 2020). In der deutschsprachigen Medien- und Diskurslinguistik hingegen hat Reddit bislang nur wenig Beachtung gefunden (vgl. Pfurtscheller 2023; Androutsopoulos 2023). Ein möglicher Grund für diese Forschungslücke liegt darin, dass Reddit-Daten zwar öffentlich als Datensatz zugänglich sind (Baumgartner et al. 2020), ihre schiere Größe jedoch einen niederschwelligen Zugang erschwert, insbesondere für Forschende ohne Programmierkenntnisse. So umfasst etwa die Textdatei mit sämtlichen Kommentaren des größten deutschsprachigen Subforums (r/de) von seiner Gründung 2006 bis 2023 ca. 20 GB, ein Volumen, das von gängigen Texteditoren nicht mehr verarbeitet werden kann.

Mit ReCKS („**R**eddit **C**orpus **K**eyword **S**earch") wird derzeit eine neue Web-Applikation für die niedrigschwellige Gewinnung und Auswertung von Daten aus Reddit-Kommentaren entwickelt. Ziel ist es, Sprachdaten aus nahezu zwei Jahrzehnten (2006–2023) für die linguistische Forschung und Lehre zugänglich zu machen und auf dieser Grundlage (mikro-)diachrone Untersuchungen des öffentlichen Online-Sprachgebrauchs zu ermöglichen. Die einfache Handhabung des Tools und seine freie Online-Zugänglichkeit machen es auch für Studierende, die etwa im Rahmen von Seminar- oder Abschlussarbeiten mit digitalsprachlichen Korpora arbeiten möchten, besonders attraktiv.

Reddit ist in Diskussionsforen (genannt „Communities", „Subreddits" bzw. „Subs") zu allen erdenklichen Themen und in zahlreichen Einzelsprachen unterteilt. Die Kommunikation ist anonym und asynchron und wird lokal moderiert. Die Reddit-Nutzerschaft ist mehr-

heitlich männlich und vergleichsweise jung, allerdings sind aufgrund der Anonymität der Plattform keine belastbaren soziodemografischen Generalisierungen möglich (Blombach et al. 2020, S. 6310; Proferes et al. 2021). Der in ReCKS verwendete Datensatz stammt derzeit aus dem Subreddit r/de, dem größten deutschsprachigen Sub mit derzeit über 3 Millionen registrierten Nutzer:innen und etwa 1.400 gleichzeitig aktiven Nutzer:innen während der üblichen Arbeitszeiten in Mitteleuropa (Stand: Mai 2025). Im Gegensatz zu thematisch oder regional spezialisierten Subs (etwa r/automobil oder r/Wien) ist r/de nicht auf ein spezifisches Thema oder Land ausgerichtet, sondern beschreibt sich selbst als „Sammelbecken für alle Deutschsprechenden". Abbildung 1 zeigt exemplarisch den Auftakt eines Threads auf r/de.

Abb. 1: Ein Thread auf r/de

Ein Initialbeitrag umfasst einen Titel (obligatorisch) und wahlweise Fließtext (wie in Abb. 1), ein Foto oder einen (bebilderten) Link zu einem Medienbeitrag. Zudem wird er mit einem thematischen Tag (sog. „Flair"), das aus einer vorgegebenen Liste ausgewählt wird, versehen. Flairs ermöglichen eine grobe inhaltliche Kategorisierung der Threads im Korpus und sind auch in den Metadaten des ReCKS-Korpus enthalten. Der hier abgebildete Thread trägt das Flair „Diskussion/Frage", weitere frequente Flairs auf r/de sind u. a. „Nachrichten", „Politik" oder „Humor". An den Initialbeitrag schließen Nutzerkommentare an, die sich entweder direkt auf den Initialbeitrag oder auf andere Nutzerkommentare beziehen können. Die in Abbildung 1 sichtbare Baumstruktur von Reddit erleichtert es dabei, Gesprächsverläufe innerhalb eines Threads nachzuvollziehen. Das ReCKS-Korpus besteht nur aus Nutzerkommentaren. Initialbeiträge wurden ausgeschlossen, weil sie strukturell viel zu heterogen sind. Beispielsweise haben in einer von uns untersuchten Stichprobe von ca. 20.000 Initialbeiträgen 85% davon keinen Fließtext und entsprechen somit nicht den Anforderungen einer korpuslinguistischen Auswertung. Allerdings sind die Titel der Initialbeiträge in den Metadaten des ReCKS-Korpus enthalten.

2. Architektur von ReCKS

Die Applikation ReCKS ist in drei Hauptbereiche gegliedert: Korpusdaten, Input/Abfrage und Output/Ergebnisse. Die technische Umsetzung erfolgt in der Programmiersprache R (R Core Team 2024) unter Verwendung des R-Pakets {shiny} (Chang et al. 2024), das zur Erstellung interaktiver Web-Applikationen dient.

2.1 Das ReCKS-Korpus

Das in ReCKS verwendete Korpus besteht aus Nutzerkommentaren aus r/de und ihnen zugeordneten Metadaten. Die Daten stammen aus dem Pushshift-Datensatz (Baumgartner et al. 2020), der alle Reddit-Initialbeiträge und -Kommentare enthält und monatlich aktualisiert wird. Verwendet wurde eine nach Subreddit reorganisierte Version dieses Datensatzes, die den gezielten Download einzelner Subreddits ermöglicht (Watchful1 o. D.). Aufgrund des riesigen Gesamtvolumens wird in der aktuellen Testphase nur eine Teilmenge des Datenbestandes von r/de herangezogen. Sie umfasst (a) alle Kommentare von 2006 bis 2014 sowie (b) die ersten 100.000 Kommentare pro Jahr von 2015 bis zum 17. Januar 2023. Diese Beschränkung ergibt sich dadurch, dass Reddit seine API-Richtlinien im Sommer 2023 drastisch eingeschränkt hat (Shakir 2023). Tabelle 1 stellt die Datenauswahl pro Jahr dar.

Jahr	Anzahl Kommentare	Anzahl Tokens
2006	118	2.356
2007	193	4.906
2008	241	6.646
2009	1.853	76.224
2010	2.555	95.865
2011	7.507	296.400
2012	19.684	730.581
2013	40.207	1.724.390
2014	86.444	4.062.803
2015	100.000	4.486.850
2016	100.000	3.905.236
2017	100.000	3.451.260
2018	100.000	3.421.205
2019	100.000	3.375.192
2020	100.000	3.460.093
2021	100.000	4.077.560
2022	100.000	4.016.031
2023	100.000	4.166.932
Summe	**1.058.802**	**41.360.530**

Tab. 1: Zusammenfassung der Korpusdaten, ReCKS v1.03

Wie Tabelle 1 zeigt, ist das Korpus nicht ausgewogen, da bis Anfang der 2010er Jahre weit weniger Kommentare pro Jahr verfügbar sind. Dies ist größtenteils auf das exponenzielle Wachstum von Reddit im deutschsprachigen Raum zurückzuführen, zum Teil jedoch wohl auch darauf, dass ein Teil der früheren Kommentare von den User:innen selbst oder von den Subreddit-Moderator:innen gelöscht wurde, bevor sie gescrapet werden konnten. Das Korpus wurde maschinell um Kommentare bereinigt, die ausschließlich aus den Hinweisen „[deleted]“ bzw. „[removed]“ bestanden, d. h. von User:in selbst bzw. von Moderator:in gelöscht wurden. Insgesamt beläuft sich die Datenbasis somit auf ca. 1 Million Kommentare bzw. ca. 41 Million Tokens.

Die Sprachdaten werden roh, d. h. ohne Lemmatisierung und POS-Tagging, eingegeben. Die Baumstruktur der Kommentare geht im für das Korpus verwendeten Pushshift-Datensatz verloren. Folglich wird jeder Kommentar nicht in seinem sequenziellen Kontext dargestellt, sondern als eigenständiger Text behandelt. In den Kommentaren enthaltene URLs wurden gelöscht und durch <URL> entsetzt, um technische Probleme mit der Suche mittels regulärer Ausdrücke zu vermeiden. Die erhobenen Metadaten sind:

- das Jahr
- der Zeitstempel
- der Titel des Initialbeitrags
- der Flair (Threads ohne Flair werden mit „[no flair]“ markiert)
- die (maschinell ausgelesene) Anzahl der Token pro Nachricht (als Token gelten ausschließlich Wörter, keine Emojis oder Interpunktion)

Auf weitere Metadaten wird aus forschungsethischen Gründen im Hinblick auf eine mögliche Rückverfolgung der Nutzerbeiträge bewusst verzichtet. Dies betrifft besonders direkte Permalinks zu jedem einzelnen Kommentar sowie Nutzer-Pseudonyme, die etwa eine Analyse der individuellen Posting-Frequenz und Themenwahl ermöglichen würden. In den Kommentaren vorkommende Nutzer-Pseudonyme wurden jedoch nicht gelöscht.

2.2 Input/Abfrage

Die Datengewinnung mit ReCKS erfolgt über eine Stichwortsuche, bei der zwei Strategien zum Einsatz kommen: die einfache exakte Suche sowie die Suche mittels regulärer Ausdrücke (Regular Expressions, RegEx). Abbildung 2 zeigt die Suchmaske mit beiden Optionen.

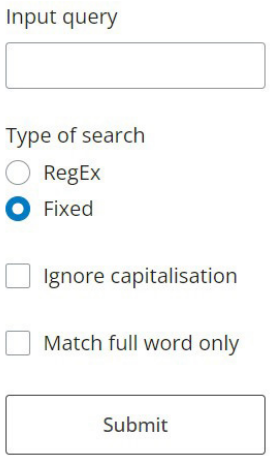

Abb. 2: Die Input-Optionen von ReCKS

Bei der **einfachen exakten Suche** wird exakt nach der im Feld „Input query“ eingegebenen Zeichenkette gesucht. Diese kann aus einem vollständigen Wort (z. B. <Lehrer>), einem Wortbestandteil (z. B. <lehr>), Interpunktionszeichen (z. B. <?!>) oder Emoji (z. B. <😂>) bestehen. Die Zeichenkette darf keine Leerzeichen enthalten. Soll etwa ein Phraseologismus oder sonstige Mehrworteinheit durchsucht werden, muss die Suche auf ein einzelnes, prägnantes Wort reduziert werden, eine gezielte Auswahl der relevanten Kommentare kann erst im Anschluss an den Export erfolgen. Beispielsweise muss die Phrase *i bims* („Jugendwort des Jahres“ 2017) über das Wort <bims> gesucht werden.

In der aktuellen ReCKS-Version ist eine Kombination von Buchstaben und Interpunktionszeichen innerhalb einer Zeichenkette nur eingeschränkt möglich. Erlaubt ist die Suche nach der Kombination von Buchstaben und den folgenden Interpunktionszeichen: Apostroph, Bindestrich, Sternchen, Doppelpunkt, Unterstrich. Ziel dieser Einschränkung ist eine optimierte Trefferausgabe: Bei der Suche nach einem bestimmten Wort wird die umgebende Zeichensetzung nicht in der Trefferspalte (query_hit, vgl. Abschn. 2.3) mitextrahiert, gleichzeitig wird die Suche nach potenziell interessanten graphischen Varianten ermöglicht. Beispielsweise ist eine gezielte Suche nach dem Ausdruck <Moin!> nicht umsetzbar, stattdessen muss nach <Moin> gesucht und anschließend in den exportierten Daten gezielt nach der Kette <Moin!> gefiltert werden (vgl. Abschn. 2.3). Dafür können beispielsweise durch Apostroph markierte Elisionen (<geh’>), Durchkopplungen (<Ost-West-Konflikt>) oder grafische Varianten des Genderns (<Lehrer*in>, <Lehrer:in>, <Lehrer_in>) gezielt durchsucht werden.

Bei der einfachen exakten Suche stehen zwei zusätzliche Optionen zur Verfügung: Großschreibung ignorieren („Ignore capitalisation“) sowie Wortgrenzen berücksichtigen („Match full word only“). Die erste Option ermöglicht es, unterschiedliche Varianten der Groß- und Kleinschreibung gleichzeitig zu erfassen: So liefert eine Suche nach der Form <moin> sowohl <moin> als auch <Moin> und <MOIN>. Bei der zweiten Option werden die Treffer auf die Tokens eingeschränkt, die exakt mit der eingegebenen Zeichenkette übereinstimmen. Dies kann bei der Suche nach bestimmten Flexionsformen oder lexikalischen Einheiten nützlich sein, da ansonsten die Suche nach <gehe> auch den Infinitiv und die Suche nach <moin> Treffer wie <Guantana**moin**haftierte> einschließt.

Die Suche mit **RegEx** ermöglicht die Formulierung präziserer und komplexerer Suchanfragen. Die verwendete RegEx-Syntax des R-Pakets {stringr} (Wickham 2023) basiert auf der standardisierten PCRE-Syntax, die auch in anderen Programmiersprachen wie Python oder JavaScript benutzt wird. Eine vollständige Liste zulässiger Ausdrücke findet sich in der Dokumentation des entsprechenden Pakets, die ReCKS-Anleitung (Yudytska 2025) fasst die wichtigsten davon zusammen. Auch in dieser Suchform darf die eingegebene Zeichenkette keine Leerzeichen enthalten, kann jedoch beliebige Kombinationen aus Buchstaben und Interpunktionszeichen umfassen, was dazu führt, dass gegebenenfalls auch angrenzende Interpunktionszeichen mitextrahiert werden (vgl. Abb. 2).

Die RegEx-Suche erlaubt u. a. den Einsatz von Wildcards, Quantoren, Wortgrenzen und dem ODER-Operator. Die vielfältigen Möglichkeiten lassen sich an der Suche nach Komposita mit dem Erstglied *Klima* exemplarisch veranschaulichen. Umgesetzt wird sie mit der folgenden Zeichenkette:

- \b(k|K)lima[:alpha:]{3,}

Hier steht <\b> für eine Wortgrenze und schließt somit Treffer mit *-klima* als Zweitglied (z. B. *Arbeitsklima*) aus. Die Schreibweise <(k|K)> erlaubt die Erfassung von groß- und kleingeschriebenen Komposita, d. h. sowohl *Klimaaktivist* als auch *klimaschädlich* werden

erfasst. Schließlich gibt die Kette <[:alpha:]{3,}> an, dass auf <Klima> mindestens drei beliebige Buchstaben folgen müssen, wodurch sichergestellt wird, dass ein Zweitglied im Kompositum vorhanden ist. Diese Zeichenkette lässt sich je nach Untersuchungsziel weiter verfeinern. Beispielsweise kann bei Interesse am Klimawandeldiskurs das mögliche Zweitglied *-anlage* durch eine weitere RegEx, nämlich <(?!anlage)>, ausgeschlossen werden. Damit sind im ReCKS-Korpus recht präzise Suchanfragen möglich.

Zusammenfassend eignet sich die einfache exakte Suche besonders für einen schnellen, explorativen Zugriff auf das Datenmaterial, während die Regex-Suche besser für gezielte Anfragen und die Untersuchung graphemischer Variation geeignet ist.

2.3 Output/Ergebnisse

In der Web-Applikation werden die Treffer in Tabellen- und Diagrammform angezeigt und können zwecks Weiterbearbeitung bequem exportiert werden.

year	flair	kwic_line	query_hit
2017	Frage/Diskussion	ist der nicht gerade. > Klimawandel, ansteigen der ozeane -> du	Klimawandel,
2017	Frage/Diskussion	Menschen in den ohnehin bedrohten Klimaregionen leiden und es werden dann	Klimaregionen
2017	Frage/Diskussion	anbahnen. Falls Trump's Kabinett den Klimawandel als Hoax abtut und sich	Klimawandel
2017	Politik	das Wasserstoffauto als "...eines der klimafeindlichsten Autos überhaupt..." zu qualifizieren. [Link]([URL].	klimafeindlichsten

Abb. 3: Auszug aus der Ergebnistabelle für die RegEx-Suche <\b(k|K)lima[:alpha:]{3,}>

Die Ergebnistabelle (Abb. 3) enthält von links nach rechts das Jahr, das Flair des zugrundeliegenden Threads, eine KWIC-Zeile und den genauen Treffer. Die KWIC-Zeile zeigt fünf Tokens vor und nach dem Treffer, bzw. bis zur Textgrenze, an, was einen unmittelbaren Einblick in den Gebrauchskontext ermöglicht. Als Treffer wird je nach Suche eine oder mehrere mögliche Wortformen angezeigt, bei einer RegEx-Suche schließt dies auch die umgebende Zeichensetzung mit ein, wie in der ersten Zeile in Abbildung 3. Die Ergebnisse werden in der Voreinstellung in zeitlicher Reihenfolge dargestellt und lassen sich über alle vier Spalten alphabetisch bzw. numerisch umsortieren.

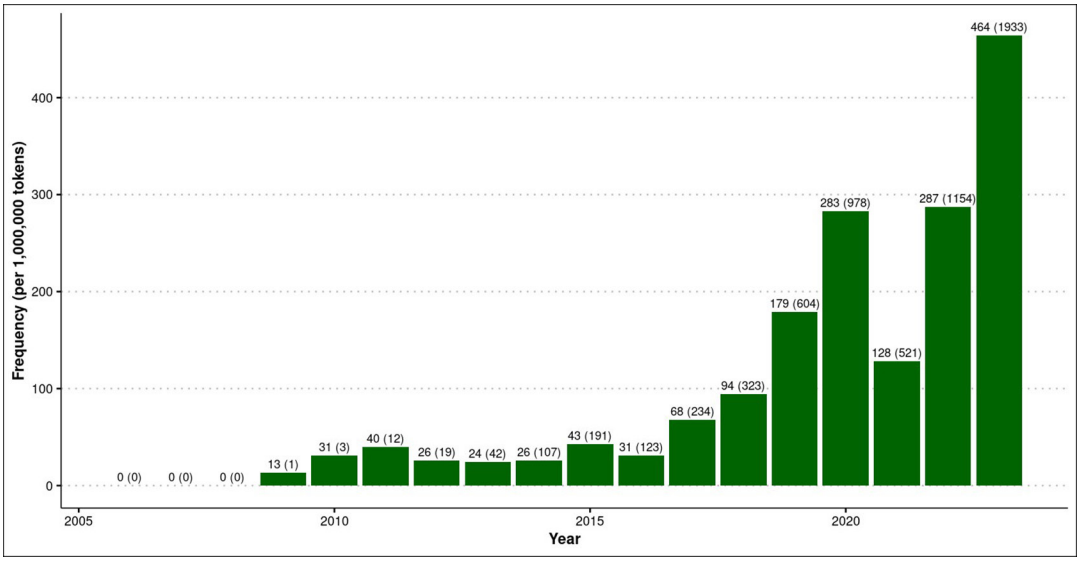

Abb. 4: Histogramm-Visualisierung der Ergebnisse für die RegEx-Suche <\b(k|K)lima[:alpha:]{3,}>

Weiterhin werden die Ergebnisse in Form eines Balkendiagrammes visualisiert, das die normalisierte Frequenz der Treffer pro Jahr darstellt (Abb. 4). Die erste Zahl über jedem Balken gibt die Anzahl der Treffer pro Million Tokens an, die zweite Zahl (in Klammern) die absolute Tokenanzahl. Auf diese Weise lässt sich die diachrone Entwicklung des Suchbegriffs im Zeitraum von 2006 bis 2023 auf einen Blick nachvollziehen.

year	timestamp	query_hit	kwic_line	submission_title	flair	body	token_count	ID
2017	17/01/2017 16:30	Klimawandel,	ist der nicht gerad	Gibt es auf /r/de auc	Frage/Diskus	> Die hast du auf Abruf	96	502
2017	17/01/2017 17:58	Klimaregionen	Menschen in den	Wovon seid ihr felse	Frage/Diskus	Ich wüsste wirklich gen	256	503
2017	17/01/2017 17:58	Klimawandel	anbahnen. Falls Tr	Wovon seid ihr felse	Frage/Diskus	Ich wüsste wirklich gen	256	503
2017	18/01/2017 09:04	klimafeindlichsten	das Wasserstoffau	DIE ZEIT: Weltkonzer	Politik	Liebe Zeit, das werden	260	504

Abb. 5: Exportierte Tabelle mit den Ergebnissen für die Beispielsuche <\b(k|K)lima[:alpha:]{3,}>

Schlussendlich kann die Tabelle als CSV-Datei (mit Semikolon als Trennzeichen) exportiert und gespeichert werden (Abb. 5). Die exportierte Datei enthält neben den in der Ergebnistabelle enthaltenen Metadaten zusätzlich auch den genauen Zeitstempel, den Titel des Initialbeitrags, den vollständigen Kommentartext („body“), die Anzahl der Tokens im Kommentar sowie eine eindeutige Kommentar-ID. So zeigt Abbildung 5, dass die beiden mittleren Zeilen dieselbe ID aufweisen, was bedeutet, dass beide Treffer (*Klimaregionen* und *Klimawandel*) aus demselben Kommentar stammen. Die CSV-Datei bildet eine Grundlage für verschiedene weiterführende Analysen, etwa eine korpuslinguistisch unterstützte Kollokationsanalyse mit AntConc (Anthony 2016) oder eine qualitative Auswertung der Kommentartexte.

3. Anwendungsbereiche

Mit ReCKS können Sprachdaten unterschiedlicher Art gewonnen und insbesondere (aber nicht nur) im Hinblick auf mikrodiachrone Entwicklungen und Dynamiken im öffentlichen Online-Sprachgebrauch ausgewertet werden. Drei Anwendungsbereiche stehen dabei im Vordergrund:

Lexikalische Innovationen in der Mikrodiachronie: Neologismen im Online-Sprachgebrauch können durch die Auswertung der Zeitstempel und Visualisierung der Ergebnisse präzise dokumentiert werden. Damit lassen sich u. a. Anglizismen, das „Jugendwort des Jahres“ und andere Arten lexikalischer Innovation im öffentlichen Online-Sprachgebrauch untersuchen. Speziell mit Bezug auf die verschiedenen „Jugendwörter“ lassen die Balkendiagramm-Visualisierungen der Verlaufsstrukturen schnell erkennen, ob es sich um Modewörter handelt, die schlagartig aufkommen und ebenso schnell wieder verschwinden. Beispielsweise ist der Anglizismus *cringe*, „Jugendwort des Jahres“ 2021, im ReCKS-Testkorpus bereits seit 2013 mit stabiler Verwendungshäufigkeit belegt. Dagegen findet sich für *i bims*, „Jugendwort“ 2017, eine einzelne Nutzungsspitze im Jahr 2018 mit danach steil abfallender Frequenz.

Graphemische Variation: Mittels RegEx-Suche können Variationsmuster im Bereich der Phono- und Graphostilistik (Androutsopoulos/Busch (Hg.) 2020) recherchiert und miteinander bzw. mit der jeweiligen Normalform verglichen werden. Exemplarisch werden in Abschnitt 4 Varianten des Genderzeichens mit Hilfe von ReCKS untersucht.

Diskursrelevante Begriffe: Mit ReCKS können diskursrelevante Begriffe recherchiert und in ihrer diachronen Entwicklung sowie im Kontext ihrer Kommentare untersucht werden. Beispielsweise lassen sich mit dem einfachen Suchwort <Klima> das gleichlautende Simplex und zahlreiche Komposita ermitteln. Dabei zeigt es sich, dass *Klima-Wörter* ab 2009 als Types wie auch Tokens deutlich zunehmen und in mehrere thematische Kon-

texte (Umwelt, Nachrichten, Politik, Wissenschaft/Technik, Verkehr/Reisen, Humor) eingebunden sind. Nach dem Datenexport lassen sich die einschlägigen, oft längeren Kommentare zudem auch qualitativ untersuchen.

4. Anwendungsbeispiel: Genderzeichen

Der Gebrauch von und das Interesse an geschlechtergerechter Sprache sind im frühen 21. Jahrhundert deutlich gewachsen, insbesondere nach dem Urteil des Bundesverfassungsgerichts von 2017 zur dritten Geschlechtsoption (Schneider 2021; Bast et al. 2024). Im Deutschen lassen sich Personenbezeichnungen auf unterschiedliche Weise als geschlechterinklusiv markieren (gendern). Müller-Spitzer/Ochs (2023) unterscheiden folgende Strategien: Doppelformen (*Studenten und Studentinnen*), Neutralisierungen (*Studierende*) und Verwendung verschiedener Genderzeichen, darunter das Binnen-I (*StudentIn*), der Schrägstrich (*Student/in*), die Klammerschreibung (*Student(in)*), das Gendersternchen (*Student*in*), der Unterstrich (*Student_in*) und der Doppelpunkt (*Student:in*). Die letzten drei Varianten (Sternchen, Unterstrich, Doppelpunkt) werden gezielt verwendet, um nichtbinäre Geschlechtsidentitäten sprachlich sichtbar zu machen (Kotthoff 2020; Waldendorf 2024). Im öffentlichen Diskurs sind Genderzeichen und insbesondere das Gendersternchen umstritten, da wortinterne typografische Zeichen im Widerspruch zu etablierten orthografischen Konventionen des Deutschen stehen (vgl. Sökefeld 2021; Krome 2022; Link 2024).

Bisherige korpuslinguistische Untersuchungen zum Gebrauch geschlechtergerechter Sprachformen konzentrieren sich auf institutionelle Kontexte, etwa journalistische Texte (Krome 2022; Waldendorf 2024; Link 2024) oder behördliche Kommunikation (Elmiger/Schaeffer-Lacroix/Tunger 2017; Müller-Spitzer/Ochs 2023). ReCKS ermöglicht nun Einblicke in den Gebrauch von Genderzeichen in der außerinstitutionellen Online-Öffentlichkeit. Reddit-Kommentare unterliegen keiner institutionellen Sprachregelung, was besonders für die Erforschung orthografisch nichtstandardisierter Formen wertvoll ist.

Genderzeichen lassen sich in Reddit-Kommentaren mittels RegEx präzise identifizieren. Tabelle 2 zeigt für sechs verschiedene Genderzeichen jeweils den eingesetzten RegEx-Ausdruck sowie ein Beispiel aus dem Korpus.

Genderzeichen	**RegEx**	**Beispiel**
Binnen-I	\b[:upper:][:alpha:]+In(\b|nen\b)	*PolitologInnen*
Schrägstrich	\b[:upper:][:alpha:]+/-*(i|I)n(\b|nen\b)	*Ärzt/in, EU-Bürger/-innen*
Klammerschreibung	\b[:upper:][:alpha:]+\(-*(i|I)n(\)|nen\))	*Kanzler(in), Autor(-in)*
Sternchen	\b[:upper:][:alpha:]+*(i|I)n(\b|nen\b)	*Partner*in*
Unterstrich	\b[:upper:][:alpha:]+_(i|I)n(\b|nen\b)	*Kund_innen*
Doppelpunkt	\b[:upper:][:alpha:]+:(i|I)n(\b|nen\b)	*Verteidigungsminister:in*

Tab. 2: Formen von Genderzeichen

Allen RegEx-Ausdrücken gemeinsam ist, dass sie ein Wort erfassen, das mit einem Großbuchstaben beginnt (<[:upper:]>), also ein Nomen darstellt, und mit dem femininen Suffix

-in im Singular oder Plural endet. Freilich sind sowohl falschnegative (d. h. zielkonforme, aber durch die RegEx nicht erfasste) als auch falschpositive (d. h. nicht zielkonforme, aber zu einer RegEx passende) Treffer möglich. Die verwendete RegEx schließt kleingeschriebene Nomen aus; da jedoch nicht alle Nutzer:innen der Substantivgroßschreibung folgen, werden einzelne Fälle wie <fotograf:in> nicht erfasst (falschnegativ). Umgekehrt werden insbesondere bei der Schrägstrich-Variante vereinzelt auch Konstruktionen wie <im Land/in Europa> erfasst (falschpositiv). Beim Erstellen dieser RegEx ist daher ein Ausgleich zwischen falschnegativen und falschpositiven Treffern notwendig, da Letztere eine nachgelagerte manuelle Prüfung und ggf. Bereinigung der exportierten Daten erfordern. Eine solche Nachbearbeitung ist grundsätzlich möglich, wird im Folgenden jedoch nicht weiter ausgeführt.

Alle untersuchten Formen mit Genderzeichen sind im gegenwärtigen ReCKS-Korpus vertreten, wie in Tabelle 3 veranschaulicht. Ihre Verwendung ist insgesamt sehr gering: Zusammengenommen erreichen alle Varianten lediglich eine normalisierte Häufigkeit von 76 Treffern pro Million Tokens. Auch der Anteil der gegenderten Formen eines Substantivs im Vergleich zu allen Vorkommen ist insgesamt sehr gering. Beispielsweise ist *Schüler* ein häufig gegendertes Wort, es erscheint im ReCKS-Korpus 108-mal mit Genderzeichen und insgesamt erscheint *Schüler* 2.994-mal im Korpus, der Anteil der gegenderten Formen beträgt somit nur ca. 3,6%.

Form	**Anzahl Treffer**
Binnen-I	1.005
Schrägstrich	444
Klammerschreibung	63
Sternchen	696
Unterstrich	275
Doppelpunkt	654
Gesamt	**3.137**

Tab. 3: Vorkommen der Genderzeichen im ReCKS-Korpus (absolute Zahlen)

Wie aus Tabelle 3 hervorgeht, variiert die Beliebtheit der einzelnen Gender-Varianten deutlich. Das Binnen-I ist mit Abstand die am häufigsten verwendete Form, die Klammerschreibung die mit Abstand seltenste, die übrigen Varianten bewegen sich zwischen diesen beiden Extremen. Bemerkenswert ist, dass die beiden inklusiven Varianten (Gendersternchen und Doppelpunkt) zu den frequentesten gehören. Wenn Reddit-Nutzer:innen sich also für eine gendergerechte Schreibweise entscheiden, greifen sie oft zu denjenigen Formen, die nichtbinäre Geschlechtsidentitäten sprachlich sichtbar machen.

Die drei häufigsten Varianten (Binnen-I, Gendersternchen, Doppelpunkt) weisen genügend Treffer auf, um ihre Verwendung im mikrodiachronen Vergleich zu vertiefen. Die drei untenstehenden Diagramme (Abb. 6–8) zeigen jeweils ihre diachrone Verteilung von 2006–2023. Insgesamt lässt sich feststellen, dass im Verlauf der 2010er und ganz besonders zu Beginn der 2020er Jahre zunehmend mit diesen Formvarianten gegendert wird.

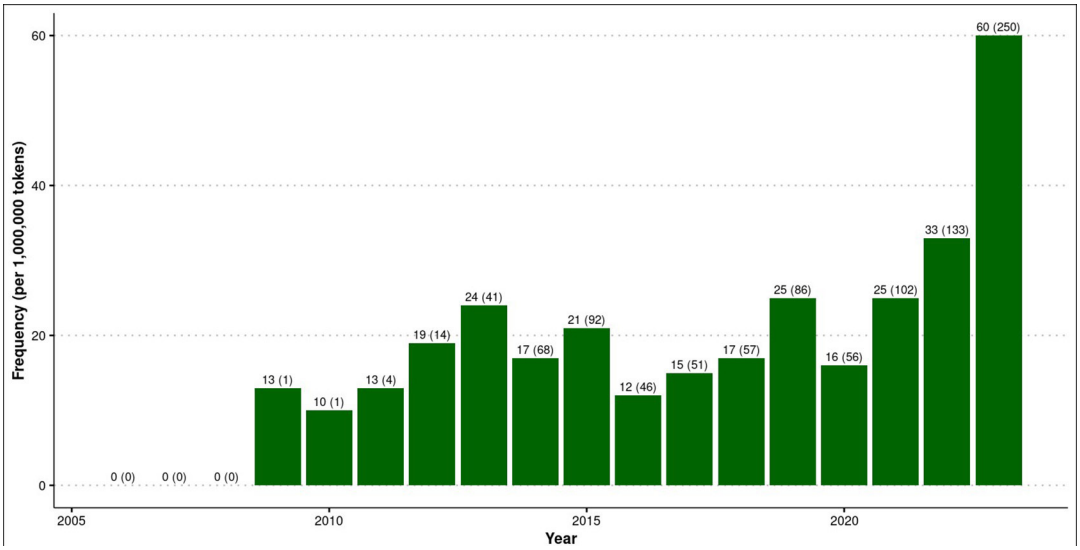

Abb. 6: Diachrone Verteilung der Binnen-I-Variante (normalisierte Häufigkeit pro Jahr)

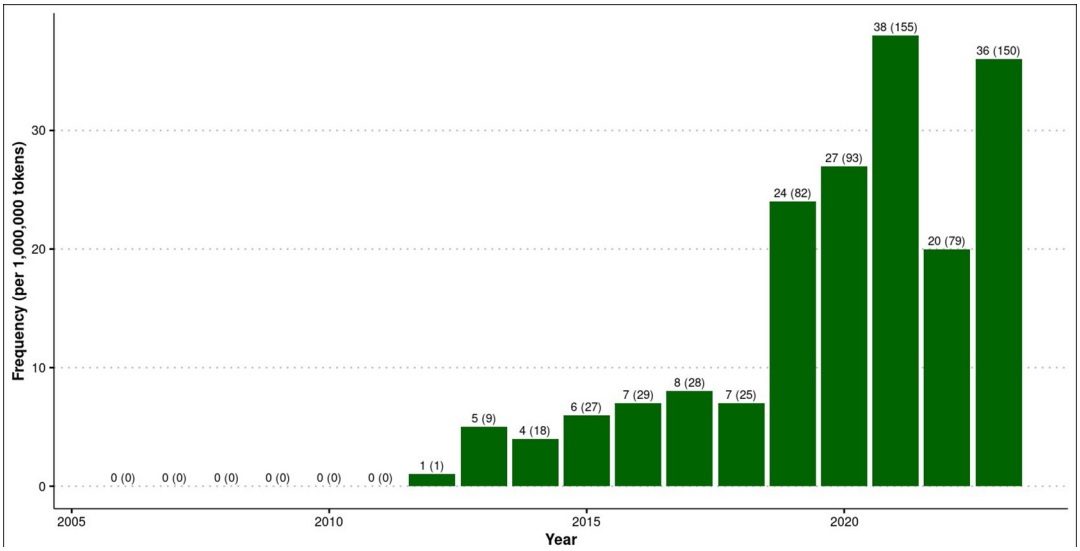

Abb. 7: Diachrone Verteilung der Sternchen-Variante (normalisierte Häufigkeit pro Jahr)

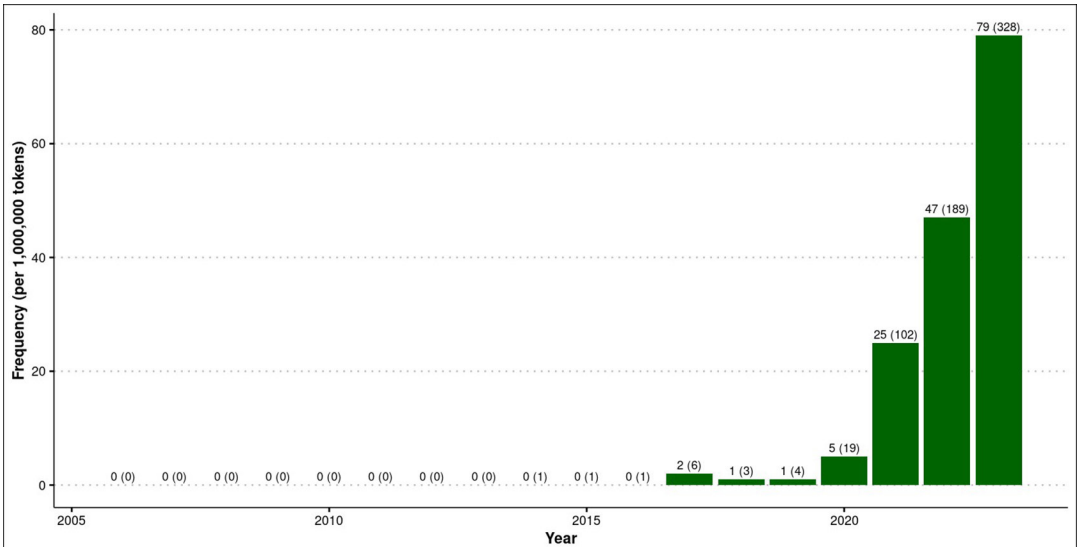

Abb. 8: Diachrone Verteilung der Doppelpunkt-Variante (normalisierte Häufigkeit pro Jahr)

Die älteste der drei Varianten, das Binnen-I, findet sich in den Reddit-Daten bereits ab 2009 (Abb. 6). Das Gendersternchen tritt erstmals 2012 auf (Abb. 7), der Doppelpunkt erscheint ab 2017 (Abb. 8) unter Ausklammerung einiger früheren falschpositiven Treffer (z. B. <Edit:in den USA.>). Zum Vergleich sind der Schrägstrich ab 2009, die Klammerschreibung ab 2011 und der Unterstrich ab 2012 belegt. Bemerkenswert ist der rasche Aufstieg des Doppelpunktes. Obwohl er die jüngste der untersuchten Varianten ist, weist er die dritthöchste Trefferzahl auf. Im Jahr 2022 übertrifft er erstmals das Sternchen, Januar 2023 sogar das Binnen-I.

Insgesamt stehen die Befunde im Einklang mit der bisherigen Forschung. Frühere Studien zu geschlechtergerechten Sprachformen erwähnen den Doppelpunkt entweder nur am Rande oder gar nicht (vgl. Sökefeld 2021; Krome 2022). Demgegenüber zeigen neuere diachrone Untersuchungen auf Basis von Zeitungskorpora ebenfalls einen deutlichen Anstieg im Gebrauch des Doppelpunkts ab den späten 2010er Jahren (Link 2024; Waldendorf 2024). In den von Waldendorf (2024) untersuchten Zeitungen wird der Doppelpunkt 2023 ca. viermal so oft verwendet wie das nächsthäufigste Genderzeichen, das Sternchen. Die rasche Verbreitung der Doppelpunkt-Variante in den Medien könnte darauf zurückzuführen sein, dass der Doppelpunkt als besonders blinden- und sehbehindertenfreundliche Form des Genderns gilt, weil er in der Sprachsynthese angeblich erkannt und ausgelassen werde (Deutscher Blinden- und Sehbehindertenverband e. V. 2024), auch wenn dies nicht immer zuverlässig der Fall sei (ebd.). Unsere Ergebnisse zeigen, dass der öffentliche Online-Diskurs auf Reddit diese zunehmende gesamtgesellschaftliche Beliebtheit widerspiegelt.

Interessant ist zudem die Beobachtung, dass Genderzeichen auch metasprachlich bzw. metapragmatisch gebraucht werden. In der Tat trifft dies bereits für den ersten Beleg des Gendersternchens aus dem Jahr 2012 zu. In einer Diskussion über das Gendern bringt ein Kommentar eine ambivalente Haltung zum Ausdruck: *Interessant finde ich auch ausserdem, dass ich irgendwie nie „Mörder/in“, „Gewalttäter*in“, „Steuerhinterzieher_in“ lese.* Die implizite Kritik lautet, dass geschlechtergerechte Formen tendenziell bei positiv oder neutral konnotierten Personenbezeichnungen verwendet werden, während bei negativ konnotierten Nomina weiterhin das generische Maskulinum dominiert. Angesichts der öffentlichen metasprachlichen Debatten um das Gendersternchen steht zu erwarten, dass sich ein solcher metasprachlicher Gebrauch quer durch das Korpus wiederfindet.

Daran anknüpfend werden nun die drei nichtbinär-inklusiven Genderzeichen (Sternchen, Unterstrich und Doppelpunkt) auf ihre jeweils zehn häufigsten Stammwörter untersucht. Die Auswertung erfolgt auf Grundlage des Datenexportes, indem jedes Token mithilfe der Suchen/Ersetzen-Funktion vom femininen Singular-/Plural-Suffix sowie von Interpunktionszeichen bereinigt wird. Die Häufigkeitszählung pro Type erfolgt anschließend über eine Pivot-Tabelle. Die Ergebnisse sind Tabelle 4 zu entnehmen.

Sternchen	Treffer	Unterstrich	Treffer	Doppelpunkt	Treffer
Schüler	26	Feminist	13	Polizist	44
Kolleg	24	Polizist	10	Schüler	23
Politiker	20	Mitarbeiter	6	Kolleg	21
Bürger	16	Student	5	Politiker	19
Lehrer	16	Patient	5	Bürger	19
Student	12	Ärzt	5	Mitarbeiter	17
Aktivist	12	Beamt	5	Lehrer	17
Feminist	12	Teilnehmer	4	Wähler	13
Mitarbeiter	11	Veganer	4	Ärzt	13
Freund	11	Rassist	4	User	12

Tab. 4: Die zehn häufigsten Stämme für die drei nichtbinär-inklusiven Genderzeichen (absolute Zahlen)

Bei allen drei Genderzeichen lassen sich Überschneidungen sowohl auf der Ebene konkreter Personenbezeichnungen als auch hinsichtlich ihrer semantischen Felder feststellen. Diese umfassen die Bereiche Bildung (*Schüler*, *Lehrer*, *Student*), Arbeitswelt (*Kollege*, *Mitarbeiter*) und öffentliches Leben (*Politiker*, *Bürger*, *Polizist*). Dennoch scheinen die beiden älteren Varianten, Sternchen und Unterstrich, stärker mit feministischen oder anderweitig politisierten Diskursen verbunden, was Bezeichnungen wie *Aktivist*, *Feminist* sowie potenziell auch *Veganer* und *Rassist* nahelegen. Der jüngere Doppelpunkt weist auf den ersten Blick keine derartigen Bezüge auf. Im Zusammenspiel mit den diachronen Verläufen legt dies nahe, dass der Erfolg des Doppelpunkts teilweise darauf zurückzuführen sein könnte, dass er als politisch neutralere Alternative die älteren Formen ersetzt.

Insgesamt handelt es sich hierbei um erste, vorläufige Befunde zur diachronen Entwicklung der Genderzeichen, die sich mithilfe von ReCKS rasch erheben und auswerten lassen. Für präzisere und vertiefte Analysen ist eine manuelle Bereinigung der Daten unerlässlich. Eine ausführlichere Studie könnte zudem diachrone Entwicklungen in Zusammenhang mit den Flairs untersuchen, um Rückschlüsse auf die diskursive Einbettung der verschiedenen Genderzeichen zu ziehen. Eine qualitative Analyse wäre auch notwendig, um zwischen objekt- und metasprachlichen Verwendungen zu unterscheiden. Ironische oder absichtlich „falsche“ Verwendungen (z. B. *Reicht mir mal jemand die **Salzstreuer_in**?*) wurden in dieser Analyse kaum berücksichtigt. Dennoch zeigt sich anhand dieses Beispiels das Potenzial von ReCKS, Reddit-Kommentare gezielt über Suchwörter zu extrahieren und für mikrodiachrone Analysen zugänglich zu machen.

5. Aktueller Entwicklungsstand und Ausblick

ReCKS (Version 1.03) ist als Web-Applikation über den ShinyApps-Server online verfügbar. Die im Hamburger Datenrepositorium veröffentlichte Anleitung zu ReCKS enthält einen Link zur aktuellen Version der Web-Applikation sowie eine detaillierte Beschreibung ihrer Verwendung (Yudytska 2025). Zukünftige Updates werden dort fortlaufend erfasst. Die Applikation ReCKS wird unter der Lizenz CC BY-NC-SA 4.0 International veröffent-

licht.[1] Diese erlaubt die Weitergabe und Anpassung des Codes, sofern die Urheber:innen genannt werden. Die Nutzung ist jedoch ausschließlich für nicht-kommerzielle Zwecke gestattet. Bei der Verwendung von ReCKS in wissenschaftlichen Arbeiten sollte auf den aktuellen Artikel verwiesen werden.

Die aktuelle Online-Version ist stabil genug, um Suchanfragen mit bis zu 100.000 Treffern zu verarbeiten; bei größeren Datenmengen kann es vorkommen, dass die Applikation hängen bleibt und keine Rückgabe liefert. Daher steht auf Anfrage auch eine Offline-Version, die das Korpus selbst im SQLite-Format und den Code beinhaltet, zur Verfügung. Diese setzt zwar grundlegende Kenntnisse in R voraus, jedoch enthält die beiliegende Anleitung sämtliche notwendige Informationen zur Nutzung.

Der nächste geplante Entwicklungsschritt (v2.0) umfasst die Migration der Anwendung auf einen leistungsstärkeren und stabileren Server. Weiterführende Pläne betreffen eine Vergrößerung des zugrundeliegenden Reddit-Korpus sowie die Integration zusätzlicher deutschsprachiger Subreddits, um diachrone Vergleiche innerhalb des deutschsprachigen Raums auszuweiten und durch sozio-thematisch gesteuerte Vergleiche zu ergänzen. Bezüglich der Funktionalität stehen Verbesserungen beim Input- und Output-Handling im Vordergrund, insbesondere die Möglichkeit, mehrere Suchausdrücke zu einer kombinierten Anfrage zusammenzufassen und direkt in einem gemeinsamen Diagramm zu vergleichen. Bei Interesse an der Anwendung oder für Zugang zur Offline-Version können die Autor:innen über recks.slm@uni-hamburg.de kontaktiert werden.

Literatur

Androutsopoulos, Jannis (2023): Punctuating the other: Graphic cues, voice, and positioning in digital discourse. In: Language & Communication 88, S. 141–152. https://doi.org/10.1016/j.langcom.2022.11.004.

Androutsopoulos, Jannis/Busch, Florian (Hg.) (2020): Register des Graphischen: Variation, Interaktion und Reflexion in der digitalen Schriftlichkeit. (= Linguistik – Impulse & Tendenzen 87). Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110673241.

Anthony, Lawrence (2016): AntConc. Tokyo: Waseda University. www.laurenceanthony.net/software (Stand: 19.12.2025).

Bast, Jennifer/Maier, Jürgen/Albert, Georg/Schneider, Jan G. (2024): Gendered debates? The use of gender-sensitive language in German televised debates, 1997–2022. In: European Journal of Politics and Gender 8, 3, S. 645–669. https://doi.org/10.1332/25151088Y2024D000000023.

Baumgartner, Jason/Zannettou, Savvas/Keegan, Brian/Squire, Megan/Blackburn, Jeremy (2020): The Pushshift Reddit Dataset. In: Proceedings of the International AAAI Conference on Web and Social Media 14, 1, S. 830–839. https://doi.org/10.1609/icwsm.v14i1.7347.

Blombach, Andreas/Dykes, Natalie/Heinrich, Philipp/Kabashi, Besim/Proisl, Thomas (2020): A corpus of German Reddit exchanges (GeRedE). In: Proceedings of the Twelfth Language Resources and Evaluation Conference, S, 6310–6316. https://aclanthology.org/2020.lrec-1.774.pdf.

Chang, Winston/Cheng, Joe/Allaire, J.J./Sievert, Carson/Schloerke, Barrett/Xie, Yihui/Allen, Jeff/McPherson, Jonathan/Dipert, Alan/Borges, Barabara (2024): Shiny: web application framework for R. R package version 1.9.1. https://CRAN.R-project.org/package=shiny (Stand: 19.12.2025).

1 Der Pushshift-Datensatz wurde ursprünglich unter der Lizenz CC-BY-4.0 International veröffentlicht (vgl. das in Baumgartner et al. 2020 verlinkte Zenodo-Repository). Beide Lizenzen erlauben die Nutzung, Bearbeitung und Weitergabe des Materials unter Nennung der Urheber:innen.

Deutscher Blinden- und Sehbehindertenverband e. V. (2024): Gendern. www.dbsv.org/gendern.html (Stand: 19.12.2025).

Elmiger, Daniel/Schaeffer-Lacroix, Eva/Tunger, Verena (2017): Geschlechtergerechte Sprache in Schweizer Behördentexten: Möglichkeiten und Grenzen einer mehrsprachigen Umsetzung. In: OBST – Osnabrücker Beiträge zur Sprachtheorie, Sprache und Geschlecht 90, 1 (Sprachpolitiken und Grammatik), S. 61–90. https://hal.science/hal-02335049/ (Stand: 19.12.2025).

Kotthoff, Helga (2020): Gender-Sternchen, Binnen-I oder generisches Maskulinum, … (Akademische) Textstile der Personenreferenz als Registrierungen? In: Linguistik Online 103, 3, S. 105–127. https://doi.org/10.13092/lo.103.7181.

Krome, Sabine (2022): Gendern in der Schule: Zwischen Sprachwandel und orthografischer Norm. In: Mitteilungen des Deutschen Germanistenverbandes 69, 1, S. 86–110. https://doi.org/10.14220/mdge.2022.69.1.86.

Link, Sabrina (2024): The use of gender-fair language in Austria, Germany, and Switzerland: A contrastive, corpus-based study. In: Lingua 308, 103787. https://doi.org/10.1016/j.lingua.2024.103787.

Müller-Spitzer, Carolin/Ochs, Samira (2023): Geschlechtergerechte Sprache auf den Webseiten deutscher, österreichischer, schweizerischer und Südtiroler Städte. In: SPRACHREPORT 2/2023, S. 1–5. https://doi.org/10.14618/sr-2-2023_mue.

Pfurtscheller, Daniel (2023): Vom Fundus Zum Korpus: Reddit als Medium und digitale Sprachressource. In: Korpora Deutsch als Fremdsprache 3, 2, Art. 2. https://doi.org/10.48694/kordaf.3864.

Proferes, Nicholas/Jones, Naiyan/Gilbert, Sarah/Fiesler, Casey/Zimmer, Michael (2021): Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. In: Social Media + Society, April-June 2021, S. 1–14. https://doi.org/10.1177/20563051211019004.

R Core Team (2024): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (Stand: 19.12.2025).

Semrush (2025): Most visited websites in the world. Updated November 2025. www.semrush.com/website/top/ (Stand: 19.12.2025).

Schneider, Jan G. (2021): Zum prekären Status sprachlicher Verbindlichkeit: Gendern im Deutschen. In: Raab, Jürgen/Heck, Justus (Hg.): Prekäre Verbindlichkeiten: Studien an den Problemschwellen normativer Ordnungen. (= Wissen, Kommunikation und Gesellschaft). Wiesbaden: Springer VS, S. 17–43. https://doi.org/10.1007/978-3-658-34227-2_2.

Shakir, Umar (2023): Reddit's upcoming API changes will make AI companies pony up. In: The Verge. www.theverge.com/2023/4/18/23688463/reddit-developer-api-terms-change-monetization-ai (Stand: 19.12.2025).

Sökefeld, Carla (2021): Gender (un)gerechte Personenbezeichnungen: derzeitiger Sprachgebrauch, Einflussfaktoren auf die Sprachwahl und diachrone Entwicklung. In: Sprachwissenschaft 46, 1, S. 111–141. https://sprw.winter-verlag.de/article/SPRW/2021/1/7 (Stand: 19.12.2025).

Waldendorf, Anica (2024): Words of change: The increase of gender-inclusive language in German media. In: European Sociological Review 40, 2, S. 357–374. https://doi.org/10.1093/esr/jcad044.

Watchful1 (o. D.): Subreddit comments/submissions 2005–06 to 2024–12. www.reddit.com/r/pushshift/comments/1itme1k/separate_dump_files_for_the_top_40k_subreddits/? (Stand: 19.12.2025).

Wickham, Hadley (2023): Stringr: simple, consistent wrappers for common string operations. R package version 1.5.1. https://CRAN.R-project.org/package=stringr (Stand: 19.12.2025).

Yudytska, Jenia (2025): Anleitung zu ReCKS: Eine Webapplikation für die linguistische Erforschung von Reddit-Kommentaren. Version 3 (Mai 2025). http://doi.org/10.25592/uhhfdm.17586.

Kontaktinformation

Dr. Jenia Yudytska
Universität Hamburg
Fakultät für Erziehungswissenschaften
Forschungszentrum *Literacy in Diversity Settings*
Von-Melle-Park 8
20146 Hamburg
E-Mail: yevgeniya.yudytska@uni-hamburg.de

Prof. Dr. Jannis Androutsopoulos
Universität Hamburg
Fakultät für Geisteswissenschaften
Fachbereich SLM I, Institut für Germanistik, Institut für Medien und Kommunikation
Von-Melle-Park 6
20146 Hamburg
E-Mail: jannis.androutsopoulos@uni-hamburg.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.82

Peter Meyer | Doris Stolberg

Eine Plattform für deutschsprachige Varietäten im Kontakt

Korpusabfrage, Sprachanalyse und Wörterbuch zur Untersuchung lexikalischer Kontaktphänomene

Abstract Within the project *Lexical Dynamics of German Varieties in Contact* at the Leibniz Institute for the German Language (IDS), lexical structures of German varieties spoken outside Central Europe's main German-speaking area are analyzed through corpus-based linguistic methods. A central question is whether lexical changes arise predominantly from particular language constellations or more broadly from general contact conditions. In addition, borrowability across word classes is investigated. This paper introduces a web-based research and analysis platform currently under development. It allows linguistic researchers to annotate corpus data flexibly, produce structured lexicographic documentation, and interactively explore linguistic phenomena. Content is presented through systematically interconnected units covering individual lexical items, word combinations, and abstract lexical patterns (e. g., discourse markers, modal particles, auxiliary selection). Navigation is facilitated by full-text searches and a thematic tagging system. The individual units of information provide dynamic access to corpus excerpts, query results, and interactive statistical analyses.

Keywords spoken language corpora, variational linguistics, language contact, lexicographical databases

1. Grundsätzliches: Das Projekt *Lexikalische Dynamik deutschsprachiger Varietäten im Kontakt*

Sprachkontaktprozesse sind durch eine besondere Dynamik gekennzeichnet, speziell im mündlichen und informellen Sprachgebrauch. Im Rahmen unseres Projekts zu deutschsprachigen Kontaktvarietäten untersuchen wir komparativ die Lexik ausgewählter Varietäten des Deutschen, insbesondere solcher, die außerhalb des geschlossenen deutschen Sprachgebiets in Mitteleuropa gesprochen werden.

Beim Vergleich der Lexik unterschiedlicher extraterritorialen Varietäten des Deutschen ergibt sich aus methodischer Sicht die Herausforderung, Kontaktphänomene auf der Grundlage von Daten aus verwandten, aber unterschiedlichen deutschen Ausgangsvarietäten zu vergleichen. Extraterritoriale Varietäten des Deutschen basieren auf verschiedenen Ausgangsvarietäten, z. B. im Fall von neuseeländischem Deutsch (Bairisch) oder pennsylvanischem Deutsch (Pfälzisch) im Vergleich zu australischem Deutsch und namibischem Deutsch, die dem heutigen Standarddeutsch näherstehen. Es fehlt also ein zuverlässiges Tertium Comparationis. Dieser Herausforderung kann zumindest partiell dadurch begegnet werden, dass bei einer anscheinenden Parallelität der Kontaktfolgen die tatsächliche Ausgangsvarietät bestmöglich berücksichtigt wird und die unterschiedlichen dialektalen

Ausgangslagen in Betracht gezogen werden. Des Weiteren können die außersprachlichen Bedingungen, z. B. die Sprachkontaktsituationen, sehr unterschiedlich sein. So ist z. B. in Namibia das Standarddeutsche medial und in der schulischen Bildung im Umfeld der deutschsprachigen Gemeinschaft sehr präsent (Wiese et al. 2017). Demgegenüber findet die Weitergabe des Deutschen in Australien, wenn überhaupt, innerfamiliär und primär mündlich statt (Clyne 1981; Riehl 2015). Diese Faktoren wirken sich auf Entlehnungen und andere lexikalische Kontakteinflüsse in unterschiedlicher Weise aus.

Zu den (mikro-)diachronen Sprachkontaktprozessen im Lexikon, die das Projekt untersucht, gehören u. a. Phänomene wie lexikalische Entlehnungen in verschiedenen Stadien der Integration, Verschiebungen der Bedeutung und der Verwendungsmuster bestehender Lexeme oder neue Formen und Funktionen lexikalisch-pragmatischer Elemente. Neben der Untersuchung von einzelwortbezogenen Entlehnungen stehen dabei insbesondere Phänomene im Bereich der lexikalischen Syntagmatik und Pragmatik im Fokus.

Gerade bei mündlichen Daten ist es oft schwierig, zwischen spontanen Phänomenen des mehrsprachigen Sprachgebrauchs (z. B. Code-Switching oder Ad-hoc-Entlehnungen) und etablierten lexikalischen Kontaktprodukten zu unterscheiden (Poplack 2018). Hier können Unterschiede in der Verwendungshäufigkeit wichtige Anhaltspunkte liefern, in dem Sinne, dass spontane Verwendungen nur vereinzelt auftreten, etablierte Entlehnungen dagegen in höherer Zahl und personenübergreifend verwendet werden. Große, digital durchsuchbare Korpora stellen somit ein wichtiges Instrument in Hinblick auf den Entlehnungsstatus von Lexemen oder ihren Verwendungsweisen dar, da solche Unterschiede erst bei einer ausreichend großen Datenmenge zuverlässig identifizierbar sind.

Die zunehmende Standardisierung (vgl. dazu Hedeland/Schmidt 2022) und Verfügbarkeit größerer mündlicher Korpora ermöglicht es daher, den Untersuchungsschwerpunkt auf gesprochensprachliche Daten zu legen und darauf aufbauend internetlexikographische Dokumentationsformate zu entwickeln. Eine wesentliche Datengrundlage des Projekts sind gesprochensprachliche Korpora zu Diaspora-Varietäten in der Datenbank für Gesprochenes Deutsch (DGD) des Leibniz-Instituts für Deutsche Sprache (IDS), zunächst zum Australiendeutschen, Unserdeutschen (Papua-Neuguinea, Australien), Namibiadeutschen und zukünftig zum Kaukasiendeutschen. In der projektbezogenen Forschung geht es primär um die Untersuchung und Erfassung unterschiedlicher Arten von lexikalischen Strukturmustern, die in ihrer Sprachkontaktabhängigkeit, ihrer Entwicklung und ihren (De-)Stabilisierungstendenzen in verschiedenen Typen deutscher Varietäten untersucht und verglichen werden (vgl. für einschlägige Untersuchungen zu Australiendeutsch u. a. Clyne 1981; Clyne 2003; Riehl 2015; Riehl/Plewnia 2018; zu Namibiadeutsch u. a. Shah 2007; Shah/Zimmer 2023; Zimmer et al. 2020; Wiese et al. 2014; Wiese et al. 2017; zu Unserdeutsch u. a. Lindenfelser 2021; Maitz 2016, 2019; Maitz/Volker 2017). Eine der zentralen Forschungsfragen ist, inwieweit Sprachwandel von spezifischen Kontaktsprachkonstellationen geprägt oder aber allgemeiner durch die Kontaktsituation an sich bedingt ist.

Ein weiteres Forschungsinteresse ist darin begründet, dass unterschiedliche lexikalische Elemente verschieden häufig entlehnt werden. Es wird seit langem angenommen, dass lexikalische Entlehnbarkeit in engem Zusammenhang mit der Zugehörigkeit zu einer bestimmten Wortart steht. Diese Annahme beruht auf der Beobachtung, dass in Sprachkontaktumgebungen, die lexikalische Entlehnungen begünstigen, Substantive in vergleichsweise höherer Zahl entlehnt werden, gefolgt von Verben und Adjektiven, während Funktionswörter deutlich seltener entlehnt zu werden scheinen (vgl. bereits Haugen 1950; Weinreich 1953). Entlehnbarkeitsskalen, die der Abstraktion dieser Beobachtungen dienen, können lexikalische Entlehnungsmuster beschreiben und bis zu einem gewissen Grad vorhersagen.

Zusätzlich werden soziolinguistische Faktoren zur Feinabstimmung der Skalen in entsprechende Modelle integriert und zur Erklärung von beobachteten Abweichungen argumentativ herangezogen, wie bereits z. B. Thomason/Kaufman (1988) darlegen.

Als mögliche Ursache für die beobachteten Unterschiede in der lexikalischen Entlehnbarkeit wird, neben sozialen Faktoren, häufig eine unterschiedliche grammatische Komplexität der Wortarten angenommen, die z. B. verhindert, dass Verben mit der gleichen Häufigkeit entlehnt werden wie Substantive, da die Verwendung von Verben morphosyntaktisch komplexer sei als die von Substantiven (vgl. Haspelmath 2009 für eine kurze Darstellung dieser Annahme). Matras/Adamou (2023, S. 5) hingegen argumentieren, dass Entlehnbarkeit eher als ein direktes Produkt kommunikativer, sozialer und pragmatischer Motivationen zum Entlehnen betrachtet werden sollte und weniger als eine Frage formaler Zwänge,[1] da die Sprachkontaktforschung gezeigt hat, dass grundsätzlich alle lexikalischen Elemente entlehnbar sind. Aus ihrer Sicht stellt eine Wortart zudem keine einheitliche Kategorie dar, die eine hohe oder niedrige Entlehnbarkeit vorhersagt. Vielmehr kann es innerhalb einer Wortart ‚gespaltene Kategoriewerte' („split category values", Matras/Adamou 2023, S. 6) geben. So zeigen verschiedene Gruppen von Funktionswörtern eine unterschiedliche Entlehnungshäufigkeit; Diskursmarker beispielsweise werden wesentlich häufiger entlehnt als Pronomina. Jedoch auch innerhalb von Wortklassen im engeren Sinne gibt es Abstufungen, so dass z. B. Präpositionen, die Distanz und komplexe Bezugspunkte oder kontrastive Bedeutung kodieren, häufiger sprachübergreifend entlehnt werden als andere Arten von Präpositionen.

Eines der Ziele der zu entwickelnden Online-Ressource, die im Folgenden genauer beschrieben wird, ist es, einen Beitrag zur Beantwortung der oben umrissenen Forschungsfragen zu leisten. Die vergleichende Untersuchung extraterritorialer Varietäten des Deutschen ermöglicht es, nicht nur allgemeine Entlehnungstendenzen von Partikularentwicklungen zu unterscheiden, sondern erlaubt auch einen Einblick in Entlehnungspräferenzen in vergleichbaren Kontexten, nämlich der informellen mündlichen Interaktion in verwandten Varietäten, in jeweils unterschiedlichen Kontaktkonstellationen.

2. Eine Forschungsplattform für Diaspora-Varietäten des Deutschen

Der vorliegende Beitrag stellt eine derzeit in Entwicklung befindliche webbasierte Forschungs- und Analyseplattform vor. Ihr Ziel ist es, einem sprachwissenschaftlich informierten Publikum innovative Möglichkeiten zur Exploration von Untersuchungsergebnissen des Projekts zu bieten, indem Beschreibungstexte mit interaktiven Zugängen zu den zugrundeliegenden Korpusdaten verknüpft werden. Die Plattform wird voraussichtlich im Laufe des Jahres 2026 als eigenständiges Onlineangebot des IDS zugänglich werden und nach einer kostenfreien Registrierung kostenfrei nutzbar sein.

Die verschiedenen Informationsangebote und Funktionalitäten der Plattform werden im Folgenden jeweils aus drei Perspektiven vorgestellt:

1 „‚Borrowability' should […] be considered a direct product of communicative, social and pragmatic motivations to borrow rather than a matter of formal constraint." (Matras/Adamou 2023, S. 5).

a. Zunächst wird die Verwendung der Online-Forschungsplattform, die über einen Browser aufgerufen wird, aus der Sicht der Nutzer:innen betrachtet.

b. Aus der Perspektive der am Aufbau der Plattform beteiligten Linguist:innen werden anschließend die redaktionellen Abläufe, die für die Erstellung der Inhalte erforderlich sind, sowie die dabei verwendeten Softwarewerkzeuge skizziert.

c. In einem letzten Schritt wird dann aus Entwickler:innensicht jeweils die zugrundeliegende Datenmodellierung, die Umsetzung durch Datenbanken und Schnittstellen sowie die technische Umsetzung in der Webapplikation diskutiert. Der Fokus der Betrachtungen liegt dabei auf der Konzeption und grundsätzlichen Architektur des Systems; Hinweise auf konkrete Implementierungsstrategien sind vorläufiger Natur.

Als konkretes Fallbeispiel sollen im Folgenden Untersuchungen zu Neubedeutungen (Sinnerweiterungen und -verschiebungen) deutscher lexikalischer Einheiten[2], insbesondere Verben, dienen, die zumindest potenziell aus Kontaktsprachen übernommen wurden.

2.1 Informationseinheiten

Das Informationsangebot der Plattform ist ganz allgemein in Einzeltexte gegliedert, die hier als *Informationseinheiten* bezeichnet werden sollen. Informationseinheiten können zum einen lexikologische Phänomene einer Einzelsprache beschreiben, von konkreten Lexemen bis hin zu bestimmten Lexemklassen, syntagmatischen Mustern und Lexikalisierungs- und Univerbierungsprozessen. Ein konkretes Beispiel wäre eine Informationseinheit zu Verwendungsweisen des Verbs *nehmen* im Australiendeutschen (DGD-Korpus: AD), die mutmaßlich durch Sprachkontakt mit dem Englischen entstanden sind, etwa die Lesart ‚jemanden/etwas an einen Ort bringen' wie in

(1) Ich mußte erst deutsch lernen, dann hat meine Tante mich zum Adelaide genommen […]

(IDS, DGD, AD--_E_00057_SE_01_T_01, [00:05:04.08]) [http://dgd.ids-mannheim.de, Stand: 3.6.2025]

Entsprechend sind auch Informationseinheiten denkbar, die das Auftreten solcher Lehnbedeutungen für mehrere australiendeutsche Lexeme gemeinsam beschreiben.

Informationseinheiten können zum anderen aber auch kontrastiv bzw. mit den Mitteln der Variationslinguistik sprachübergreifende Untersuchungsergebnisse präsentieren, in Bezug auf (1) beispielsweise allgemein Lehnbedeutungen, die in Varietäten wie dem Australien- und Namibiadeutschen auf englischen Sprachkontakt zurückgehen können. Auf einer noch allgemeineren Ebene sind Übersichtseinheiten zu bestimmten Themen angesiedelt; so könnte es etwa eine allgemeine Informationseinheit zu Lehnbeziehungen geben. Durch Querverweise auf Einheiten, in denen konkrete Aspekte des Themas behandelt werden, können solche Überblicksartikel auch als Navigationshilfen im Gesamtangebot der Plattform dienen.

Aus redaktioneller Sicht sind Informationseinheiten XML-Dokumente, die in einem auf der Plattform bereitgestellten Redaktionssystem von autorisierten Linguist:innen erstellt

2 Lexikalische Einheiten können Einzelwörter, aber auch Mehrworteinheiten, d. h. mehr oder weniger feste Kombinationen aus mehreren Wörtern, sein.

und kollaborativ bearbeitet werden. Dies wirft die Frage nach der Modellierung und internen Struktur dieser Dokumente auf. Anders als bei einem Wörterbuch werden, wie oben beschrieben, auf der Plattform sehr heterogene Klassen von Inhalten verwaltet. Je nach linguistischer Thematik stehen aber auch schon bei Informationseinheiten, die beispielsweise „nur" Einzellexeme einer Sprache behandeln, recht unterschiedliche lexikalische Aspekte im Vordergrund, so dass nicht einmal in dieser Untergruppe von Informationseinheiten eine einheitliche, granulare mikrostrukturelle Modellierung praktikabel erscheint. Informationseinheiten werden daher überwiegend als hierarchisch nach Abschnitten und Unterabschnitten gegliederter Freitext gestaltet. Dennoch sind für bestimmte Klassen von Einheiten jeweils unterschiedliche formalisierte Abschnitte sinnvoll. Im Falle von Einheiten zu Einzelwörtern entsprechen diese beispielsweise typischen lexikographischen Angaben zur Wortart, Morphosyntax, Etymologie usw. Das Redaktionssystem muss also verschiedene Typen von Informationseinheiten anbieten. Jedem Typ ist – zur Validierung der Eingaben – ein eigenes XML-Schema sowie ggf. zusätzlich eine eigene Dokumentation mit redaktionellen Vorgaben zugeordnet.

Auf der Ebene der Softwarearchitektur ist eine dokumentenorientierte Datenbank die zentrale Komponente des Plattform-Backends. Jede Informationseinheit wird zusammen mit Verwaltungsinformationen, u. a. einem volltextindizierbaren Auszug aus dem XML-Quelltext, als separates Dokument in der Datenbank abgelegt. Neue Typen von Informationseinheiten können jederzeit modular dadurch hinzugefügt werden, dass ein zugehöriges XSLT-Stylesheet und ein Schemadokument in entsprechende Verzeichnisse abgelegt werden.

2.2 Zugriffsstrukturen

Die inhaltlichen Beziehungen zwischen Informationseinheiten der Plattform sind im Allgemeinen nicht systematischer oder hierarchischer Natur, auch wenn es natürlich „vertikale" thematische Verhältnisse von über- zu untergeordneten Einheiten geben kann. Offensichtlich lassen sich Informationseinheiten auch nicht auf naheliegende Weise als sortierte Liste anbieten, abgesehen vielleicht von Einheiten zu Einzellexemen. Es gibt also keine für die Plattform geeignete elementare Makrostruktur, die einer Lemmaliste im Wörterbuch vergleichbar wäre. Das System muss aus diesem Grund geeignete Zugriffsstrukturen und Suchoptionen für Endnutzer:innen anbieten.

- Grundlegend sind Verweisverfahren: Die Informationseinheiten sind über Querverweise systematisch miteinander vernetzt. Im laufenden Text einer Einheit funktionieren diese Verweise als Verlinkungen, mit denen jeweils eine andere Informationseinheit bzw. darin enthaltene Abschnitte aufgerufen werden können. Zusätzlich kann jede Einheit separate Abschnitte anbieten, die, soweit möglich und sinnvoll, thematisch über- und untergeordnete sowie anderweitig inhaltlich zugehörige Einheiten auflistet, ähnlich wie dies von Wiki-Datenbanken vertraut ist. Spezielle Informationseinheiten können als Einstieg in eine bestimmte Thematik dienen und liefern in strukturierter Weise kommentierte Verweise auf zugehörige Einheiten der Plattform. Auch die Startseite des Gesamtsystems kann letztlich eine solche – im Weiteren als *navigationell* bezeichnete – Informationseinheit sein, die ihrerseits auf wichtige thematische Übersichtseinheiten, aber auch auf einzelne aktuell neu hinzugekommene oder anderweitig relevante Einheiten verweist.

- Wesentliche Zugriffsstruktur der Plattform ist ein dynamisch erweiterbares System von ‚semantischen' Etiketten (Tags), die Informationseinheiten beispielsweise hinsichtlich der in ihnen behandelten deutschen Varietäten und Kontaktsprachen sowie linguistischen Themen einordnen. Durch Auswählen eines oder mehrerer Tags können die Nutzer:innen der Plattform die Gesamtliste der Einheiten effektiv filtern, wobei navigationelle Einheiten in der Resultatliste besonders hervorgehoben werden. Die einer angezeigten Einheit zugeordneten Tags erscheinen visuell prominent in ihrer Online-Präsentation. In unserem Beispiel (1) einer Einheit zu kontaktinduzierten Lesarten des australiendeutschen *nehmen* sind naheliegende Tags „Varietät: Australiendeutsch", „Kontaktsprache: Englisch", „Thema: Lehnbeziehungen" und ggf. auch „Thema: Argumentstrukturen".
- Gefiltert werden können Informationseinheiten aber auch durch eine Volltextsuche. Dies kann die Möglichkeit einschließen, nach „Metawörtern" zu suchen, die im eigentlichen Text der betreffenden Einheit womöglich gar nicht auftauchen, aber für die Einheit inhaltlich einschlägig sind, ähnlich den Keywords, die für Webseiten zur Suchmaschinenoptimierung hinterlegt werden können. Das können beispielsweise standarddeutsche Metalemmata sein, wenn das in der Einheit beschriebene Wort auf eine dialektale Form der deutschen Ausgangsvarietät zurückgeht (z. B. das diskursstrukturierende Element „weißt du" als Metalemma zur dialektalen Ausprägung „weischt").

Die beschriebenen Zugriffsstrukturen lassen sich in der Benutzeroberfläche als responsive Seitenleiste gestalten, die Texteingaben bzw. Dropdown-Listen für die auswählbaren Filter zur Verfügung stellt und eine entsprechend gefilterte und sortierbare Liste von Informationseinheiten anbietet. Das Anklicken eines Elements der gefilterten Liste öffnet dann im zentralen Bereich die ausgewählte Einheit. Abbildung 1 zeigt einen Entwurf der für alle Nutzer:innen der Plattform zugänglichen Such- und Ansichtsseite für Informationseinheiten.

Im Redaktionsprozess werden Tags und Metawörter als Metadaten, d. h. letztlich durch XML-Elemente in einem ‚Header'-Bereich, realisiert. Dazu gehört auch der Status einer Einheit als navigationell. Verweise auf andere Einheiten werden über eineindeutige IDs ermöglicht und können im Redaktionssystem durch direkte Auswahl der Zieleinheit erzeugt werden. Implementierungsseitig wird die Suche durch Indizierung der Metadatenfelder der Datenbankdokumente sowie durch einen Volltextindex umgesetzt.

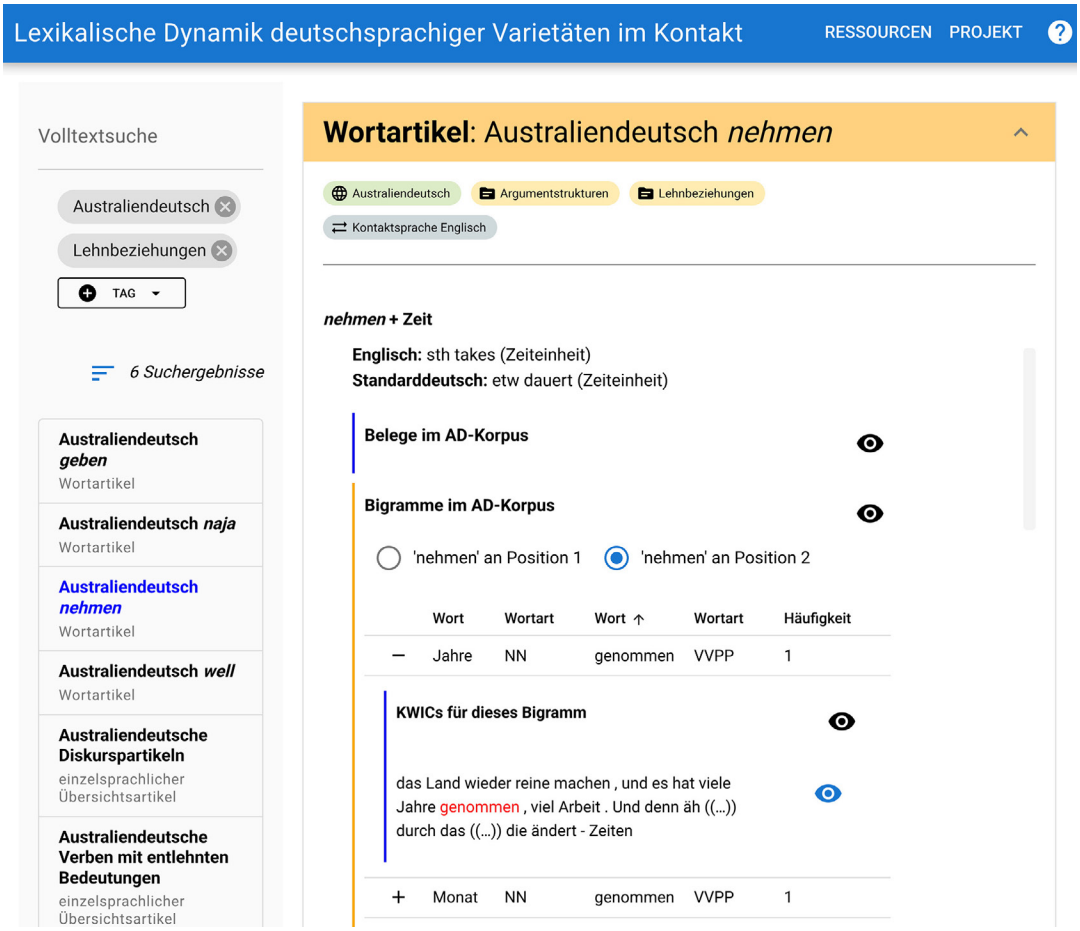

Abb. 1: Entwurf der Nutzeroberfläche des Onlinesystems. Gezeigt wird eine Informationseinheit zu kontaktinduzierten Lesarten des australiendeutschen Verbs *nehmen*. Im linken Seitenbereich erscheint eine Liste von Informationseinheiten, die hier bereits nutzerseitig nach den Tags „Australiendeutsch“ sowie „Lehnbeziehungen“ gefiltert wurde.

2.3 Interaktive Komponenten und Korpuszugriff

Die Informationseinheiten der Plattform sind nicht als vorgefertigte, statische lexikografische Beschreibungen konzipiert, sondern sollen einen konfigurierbaren, auf den jeweils relevanten Forschungskontext zugeschnittenen Zugriff auf zugehörige Daten aus den verwendeten Korpora gewähren. Eine besondere Rolle spielen für das Gesamtprojekt dabei die mündlichen Korpora zu deutschen Diasporavarietäten in der *Datenbank für Gesprochenes Deutsch* (DGD) des IDS, zu denen man nach kostenfreier Registrierung Zugang erhält. Die Ein- oder Anbindung weiterer, auch schriftsprachlicher, Korpora ist jedoch möglich und vorgesehen.

Ähnlich wie in LeGeDe („Lexik des gesprochenen Deutsch“, www.owid.de/legede/, Stand: 17.12.2025), einem am IDS entwickelten Prototyp eines Wörterbuchs zur Lexik des gesprochenen Deutsch, können insbesondere, nach Anmeldung mit den DGD-Zugangsdaten, zum jeweiligen Kontext in der Informationseinheit passende Transkriptausschnitte aus den DGD-Korpusdaten präsentiert und dabei dergestalt mit einer Audiowiedergabe kombiniert werden, dass die aktuell zu hörende Position im Transkript unterlegt ist. Dabei wird der unter dem Namen ‚ZuViel‘ firmierende Visualisierungszugang der für audiovisuelle Korpora entwickelten und u. a. für die DGD genutzten ZuMult-Softwarearchitektur (zumult.org) genutzt; vgl. Abbildung 2. Im Gegensatz zu LeGeDe soll es die hier beschriebene Plattform

aber ermöglichen, nicht nur einzelne redaktionell ausgewählte Transkriptausschnitte anzeigen zu lassen, sondern sämtliche der im Kontext passenden Korpusbelege in tabellarischer, durchsuchbarer, filterbarer und sortierbarer Form als Keywords in Context (KWICs), zu denen dann per Mausklick interaktive Transkripte erscheinen.

Abb. 2: Audiovisuelle Präsentation eines Transkriptausschnittes mit dem DGD-Visualisierungszugang ZuViel. Der Ausschnitt enthält den in Abbildung 1 gezeigten KWIC zum Verb *nehmen* in der entlehnten Bedeutung ‚dauern'.

In unserem Beispiel einer Informationseinheit zu Lehnbedeutungen der australiendeutschen Verwendung von *nehmen* heißt dies, dass zu jeder einzelnen Lehnbedeutung alle relevanten KWICs aus dem AD-Korpus der DGD angezeigt werden können. Neben der oben genannten Bedeutung ‚jemanden/etwas an einen Ort bringen' ist dies etwa die Bedeutung ‚eine bestimmte Zeitspanne in Anspruch nehmen', wie in

(2) […] das hat sie sechs Monat genommen, bis sie sind hergekommen hier.

(IDS, DGD, AD--_E_00039_SE_01_T_01, [00:08:48.17]) [http://dgd.ids-mannheim.de, Stand: 3.6.2025]

Um eine solche Funktionalität zu ermöglichen, muss das betreffende Korpus mithilfe eines für autorisierte Linguist:innen zugänglichen Redaktionswerkzeugs annotiert werden. Dazu wird für die Lehnbedeutungen von *nehmen* eine eigene Annotationsschicht angelegt, die nicht in der DGD, sondern in der Datenbank der Plattform verwaltet und gespeichert wird. Mithilfe einer DGD-Abfrage in der Abfragesprache CQP (vgl. Frick/Helmer/Wallner 2023, S. 50 mit Verweisen auf weitere Informationen) kann die bearbeitende Person im Annotationstool z. B. alle KWICs im Korpus AD anzeigen lassen, die eine Form des Lexems *nehmen* enthalten. Jedem *nehmen*-Token kann dann, im elementarsten Fall mit einem Dropdown-Auswahlmenü, von dem/der Annotator:in eine der im Tool vordefinierten Lesarten zugeordnet werden. Um sodann in der Informationseinheit eine KWIC-Liste für *nehmen*-Token in einer bestimmten Lesart zu generieren, muss von der Autorin bzw. dem Autor der Einheit im Editor lediglich ein einzelnes, für diesen Zweck vordefiniertes, XML-Element eingefügt werden, das über Attribute die Annotationsschicht und den annotierten Wert (Lesart) spezifiziert.

In der technischen Realisierung des Annotationswerkzeuges entspricht jedes annotierte Token einem eigenen Dokument in der Datenbank der Plattform. Dieses Dokument nimmt

zum einen mit bestimmten IDs auf das betreffende annotierte Token in der DGD Bezug, z. B. mit einer ID für das jeweilige DGD-Transkript sowie einer ID für das konkrete Token darin. Zum anderen speichert das Dokument die jeweils zugehörige Annotationsinformation, im konkreten Beispielfall also die Zuordnung zu einer Annotationsschicht für Lehnbedeutungen eines bestimmten Verbs sowie die dem Token zugeordnete Lehnbedeutung. Durch dieses Vorgehen entfällt die Notwendigkeit, Daten der DGD-Datenbank zu duplizieren. Das Annotationswerkzeug kommuniziert über leistungsfähige Schnittstellen der bereits erwähnten ZuMult-Architektur mit der DGD (vgl. Frick/Helmer/Wallner 2023). Lässt ein:e Endnutzer:in der Plattform die KWICs einer Lesart anzeigen, werden diese ebenfalls in Echtzeit über die ZuMult-Schnittstelle abgerufen.

Aus dem XML-Element generiert ein XSLT-Stylesheet die browserseitige dynamische KWIC-Listendarstellung für die Nutzer:innen. Dazu wird eine HTML-Komponente sowie eine Serverschnittstelle für die erforderlichen komponentenspezifischen Datenbankabfragen implementiert. Im Projekt kann für verschiedene interaktive Elemente eine ganze Bibliothek solcher HTML-Elemente erstellt werden, die dann von Autor:innen in jeder Informationseinheit nach Belieben genutzt werden kann.

Auf einer abstrakteren Ebene können aber auch aggregierende, quantitative Auswertungen über dem Korpusdatenbestand in eine Informationseinheit integriert werden. Für derzeit zwei DGD-Korpora ist eine externe Webanwendung, der Lexical Explorer (Lemmenmeier-Batinić 2020), entwickelt worden, die vorausberechnete Häufigkeitslisten und tabellarische Auswertungen zu den in den betreffenden Korpora verfügbaren Annotationsmerkmalen und Metadaten bietet. Vergleichbare Funktionalität soll, direkt eingebettet in die Informationseinheiten der neuen Plattform und jeweils kontextuell begrenzt auf relevante Wörter, insbesondere auch für die plattformspezifischen Annotationen zur Verfügung stehen.

Im Beispiel des australiendeutschen *nehmen* können beispielsweise bei der Behandlung einer bestimmten Lehnbedeutung interaktive Tabellen von Bi- und Trigrammen sowie Kookkurrenzen von *nehmen* in der betreffenden Lesart angezeigt werden, einschließlich statistischer Angaben zu Frequenz und Signifikanz der jeweiligen Wortverbindung und der Möglichkeit, sich alle KWICs mit der jeweiligen Wortverbindung anzeigen zu lassen. Die Lesarten können auch mit Metadaten wie dem Sprechergeschlecht oder dem Typ der mündlichen Interaktion kreuztabelliert werden. Visualisierungen, z. B. Balken- oder Tortendiagramme zur Häufigkeitsverteilung der einzelnen Lehnbedeutungen von *nehmen* im Korpus, lassen sich ebenfalls realisieren.

Obwohl derlei quantitative Auswertungen im Prinzip automatisch über Abfragen im Korpus sowie in der Plattform-Datenbank gewonnen werden können, ist in vielen Fällen eine Echtzeitberechnung nicht möglich. So müssen für Signifikanzmaße bei Bigrammen, wie etwa der Log Likelihood Ratio, für jedes in einer Informationseinheit zu behandelnde Bigramm sämtliche anderen Bigramme des Korpus ausgezählt werden, die dasselbe Erst- bzw. Zweitelement aufweisen. Solche vorzuberechnenden Auswertungen müssen nach jeder Veränderung am Korpus- oder Annotationsbestand erneut programmgesteuert durchgeführt und die Resultate in der Datenbank der Plattform hinterlegt werden. Im Falle einer Bigrammtabelle würde jedes einzelne Bigramm mitsamt zugehörigen statistischen Maßen durch ein eigenes Datenbankdokument repräsentiert.

Abschließend sei auf die Möglichkeit hingewiesen, dass eine Informationseinheit über ein speziell zu diesem Zweck implementiertes benutzerdefiniertes HTML-Element Zugriff auf eine gesamte, in die Plattform eingebundene, Ressource gewährt und dabei auf diese Ressource zugeschnittene Suchfunktionen bereitstellt. Ein einfaches Beispiel wäre ein Wör-

terbuch zu einer bestimmten Diasporavarietät mitsamt filterbarer Lemmaliste. Auch kleinere schriftsprachliche Korpora könnten auf diese Weise eingebunden werden. Im Falle einer integrierten lexikographischen Ressource kann es von der Anzahl der Lemmata abhängen, ob es sinnvoller ist, jedes einzelne Lemma als separate Informationseinheit anzubieten, statt das gesamte Wörterbuch in einer einzigen Einheit zu präsentieren.

3. Fazit

Die hier vorgestellte, in der Entwicklung befindliche Plattform trägt dazu bei, mündliche Korpora deutschbasierter Kontaktvarietäten systematisch in Bezug auf lexikalische Kontaktphänomene zu untersuchen. Sie bietet vielfältige und neuartige technische Möglichkeiten, um kontaktinduzierte Entwicklungen in verschiedenen extraterritorialen Varietäten des Deutschen vergleichend zu betrachten und zu beschreiben. Dabei kann der Fokus je nach Forschungsinteresse auf bestimmte Wortarten, auf Einzellexeme oder auf andere Ordnungskategorien gelegt werden, um unterschiedliche Kontaktkonstellationen und deren lexikalische Auswirkungen strukturiert zu untersuchen. Dadurch ermöglicht sie es, sich der Frage nach unterschiedlichen Entlehnbarkeiten zu nähern und zu deren theoretischer Modellierung beizutragen. Ebenso erlaubt sie es, Unterschiede zwischen allgemeinen Sprachwandeltendenzen in deutschen Kontaktvarietäten einerseits und konstellationsspezifischen Entwicklungen andererseits schärfer herauszuarbeiten. Indem die traditionellen Grenzen zwischen Korpusabfrage, Sprachanalyse und Wörterbuch zunehmend aufgelöst werden, wird so aus einem digitalen Wörterbuchportal ein ebenso komplexes wie flexibles lexikologisches Forschungsinstrument.

Literatur

Clyne, Michael (1981): Deutsch als Muttersprache in Australien. (= Deutsche Sprache in Europa und Übersee 8). Wiesbaden: Steiner.

Clyne, Michael (2003): Dynamics of language contact. (= Cambridge Approaches to Language Contact). Cambridge: Cambridge University Press.

DGD = Leibniz-Institut für Deutsche Sprache (o. D.): Datenbank für Gesprochenes Deutsch (DGD). https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome (Stand: 15.12.2025).

Frick, Elena/Helmer, Henrike/Wallner, Franziska (2023): ZuRecht: Neue Recherchemöglichkeiten in Korpora gesprochener Sprache für Gesprächsanalyse und Deutsch als Fremd- und Zweitsprache. In: Korpora Deutsch als Fremdsprache 3, 1, S. 44–71.

Haugen, Einar (1950): The analysis of linguistic borrowing. In: Language 26, S. 210–231.

Hedeland, Hanna/Schmidt, Thomas (2022): The TEI-based ISO Standard ‚Transcription of spoken language' as an exchange format within CLARIN and beyond. In: Monachini, Monica/Eskevich, Maria (Hg.): Proceedings of the CLARIN Annual Conference 2021. Virtual Event, 2021, 27–29 September. (= Linköping Electronic Conference 189). Linköping: Linköping University, S. 34–45.

Lindenfelser, Siegwalt (2021): Kreolsprache Unserdeutsch. Genese und Geschichte einer kolonialen Kontaktvarietät. (= Koloniale und Postkoloniale Linguistik/Colonial and Postcolonial Linguistics [KPL/CPL] 17). Berlin/Boston: De Gruyter.

Maitz, Péter (2016): Unserdeutsch. Eine vergessene koloniale Varietät des Deutschen im melanesischen Pazifik. In: Lenz, Alexandra N. (Hg.): German abroad – Perspektiven der Variationslinguistik, Sprachkontakt- und Mehrsprachigkeitsforschung. (= Wiener Arbeiten zur Linguistik 4). Göttingen: V&R unipress, S. 211–240.

Maitz, Péter (2019): Deutsch als Minderheitensprache in Australien und Ozeanien. In: Herrgen, Joachim/Schmidt, Jürgen E. (Hg.): Sprache und Raum. Ein internationales Handbuch der Sprachvariation. Bd. 4: Deutsch. (= Handbooks of Linguistics and Comunication Science 30.4). Berlin/Boston: De Gruyter, S. 1191–1209.

Maitz, Péter/Volker, Craig A. (2017): Documenting Unserdeutsch: Reversing colonial amnesia. In: Journal of Pidgin and Creole Languages 32, 2, S. 365–397.

Matras, Yaron/Adamou, Evangelia (2023): Word classes in language contact. In: van Lier, Eva (Hg.): The Oxford handbook of word classes. (= Oxford Handbooks in Linguistics). Oxford: Oxford University Press, S. 887–898.

Poplack, Shana (2018): Borrowing. Loanwords in the speech community and in the grammar. Oxford u. a.: Oxford University Press.

Riehl, Claudia M. (2015): Language attrition, language contact, and the concept of relic variety: The case of Barossa German. In: International Journal of the Sociology of Language 236, S. 1–33.

Riehl, Claudia M./Plewnia, Albrecht (2018): Australien. In: Plewnia, Albrecht/Riehl, Claudia M. (Hg.): Handbuch der deutschen Sprachminderheiten in Übersee. Tübingen: Narr, S. 9–32.

Shah, Sheena (2007): German in a contact situation: The case of Namibian German. In: eDUSA 2, 2, S. 20–45.

Shah, Sheena/Zimmer, Christian (2023): Grammatical innovations in German in multilingual Namibia: The expanded use of linking elements and *gehen* ‚go' as a future auxiliary. In: Journal of Germanic Linguistics 35, 3, S. 205–265.

Thomason, Sarah G./Kaufman, Terrence (1988): Language contact, creolization and genetic linguistics. Berkeley, CA: University of California Press.

Weinreich, Uriel (1953): Languages in contact. (= Publications of the Linguistic Circle of New York 1). Den Haag: De Gruyter.

Wiese, Heike/Simon, Horst J./Zappen-Thomson, Marianne/Schumann, Kathleen (2014): Deutsch im mehrsprachigen Kontext: Beobachtungen zu lexikalisch-grammatischen Entwicklungen im Namdeutschen und im Kiezdeutschen. In: Zeitschrift für Dialektologie und Linguistik 81, 3, S. 274–307.

Wiese, Heike/Simon, Horst J./Zimmer, Christian/Schumann, Kathleen (2017): German in Namibia: A vital speech community and its multilingual dynamics. In: Maitz, Péter/Volker, Craig A. (Hg.): Language & Linguistics in Melanesia (Sonderheft: Language Contact in the German Colonies: Papua New Guinea and Beyond), S. 221–245.

Zimmer, Christian/Wiese, Heike/Simon, Horst J./Zappen-Thomson, Marianne/Bracke, Yannic/Stuhl, Britta/Schmidt, Thomas (2020): Das Korpus Deutsch in Namibia (DNam): Eine Ressource für die Kontakt-, Variations- und Soziolinguistik. In: Deutsche Sprache 48, 3, S. 210–232.

Kontaktinformation

Dr. Peter Meyer
Leibniz-Institut für Deutsche Sprache (IDS)
R 5, 6-13
68161 Mannheim
E-Mail: meyer@ids-mannheim.de

Dr. Doris Stolberg
Leibniz-Institut für Deutsche Sprache (IDS)
E-Mail: stolberg@ids-mannheim.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.84

Thomas Eckart / Ulrike Ertel / Christine Rains

Deutsche Wortfeldetymologie

Bedeutungswandel im deutschen Wortschatz

Abstract The research project „Semantic Field Etymology in German and in European Context – Man in Nature and Culture“ (DWEE), funded by the Saxon Academy of Sciences and Humanities in Leipzig and affiliated with the Department of Indo-Germanic Studies at the Friedrich-Schiller University Jena, aims to make semantic change within the German lexicon accessible to systematic investigation within a broader European framework. Drawing on a curated, interactively accessible corpus and regular publication of research results, the project reveals synchronic and diachronic patterns of lexical and semantic change within the German lexicon by tracing the evolution of semantic fields from present-day German back to Old High German.

Keywords semantic field, semantic change, etymology, word history

1. Das Projekt

1.1 Forschungsziele

Für die formale Rekonstruktion von Wörtern verfügt die Indogermanistik über ein ausgefeiltes methodisches Instrumentarium, das eine zuverlässige Basis für die Forschung zum Sprachwandel bietet. Schwieriger ist die Situation bei der Untersuchung des semantischen Wandels, der ja nicht den algebraischen Regeln des Lautwandels folgt, sondern von den verschiedensten Faktoren gesteuert wird: Wortbedeutungen ändern sich durch Veränderungen der Realien, durch Sprachtabu, Sprachkontaktphänomene, durch Registerverschiebungen, geändertes Sprachprestige und alle möglichen anderen zum Teil schwer vorhersagbare Faktoren. Daher bleibt die Beschreibung von Bedeutungswandel in der historischen Sprachforschung oft etwas intuitiv, während die synchronen Modelle zum semantischen Wandel noch zu wenig an großen, auch historischen Korpora erprobt sind. Das Projekt DWEE möchte diese Lücke füllen. Ausgangspunkt ist dabei die schon von Jost Trier (1931) begründete, seither vielfach modifizierte Wortfeldtheorie, die davon ausgeht, dass die Wörter im mentalen Lexikon nach Sachgruppen organisiert sind, und dass sich Veränderungen an einer Stelle dieser Gruppe jederzeit auf andere, verwandte Wörter auswirken können. Wir fragen also nicht nur danach, wie der Bedeutungswandel jeweils abläuft, sondern auch danach, warum er überhaupt stattfindet. Neu ist dabei, dass wir vom Korpus und nicht von der Theorie ausgehen, dass also nicht ein vordefiniertes Wortfeld untersucht wird, sondern die Konfigurationen von Wortfeldern innerhalb eines größeren und weit definierten „semantischen Bereichs“.

Da die Zusammensetzung von Wortfeldern jedoch diachron instabil ist, muss die Untersuchung für die verschiedenen Sprachstufen jeweils gesondert durchgeführt werden. Wir erfassen daher auch das Aufkommen und Verschwinden einzelner Wortfeldbestandteile (Wörter und Teilwortfelder) im Laufe der Sprachentwicklung. Ziel ist die Etablierung einer

korpusbasierten, empirisch geprüften Wortfeldtheorie anhand einer von DWEE gesammelten Datenbasis[1]. Daher ist die Dokumentation der sprachhistorischen Primärdaten von besonderer Wichtigkeit.

Entwickelt wurde deshalb eine Lemma-Struktur nach Wortfeldern sowie deren interne diachrone Schichtung, ausgehend vom Gegenwartsdeutschen bis hin zu den Wortentsprechungen im Althochdeutschen. Durch die Rekonstruktion der historischen Bedeutung durch die Etymologie kann man in vielen Fällen den Beobachtungszeitraum noch erweitern. So ist ein sprachhistorisch aufbereitetes Korpus entstanden, das die folgenden Forschungsfragen zu beantworten hilft:

- Woher kommen die Wörter des Deutschen?
- Warum sterben bestimmte Wörter aus?
- Warum dringen neue Wörter ein?
- Wie breiten sich Wörter aus?
- In welchen semantischen Feldern ist lexikalische Konstanz besonders häufig, welche Felder unterliegen dagegen eher Sprachwandelphänomenen?
- Stehen Sprachwandelmotive mit onomasiologischen Prozessen oder mit bestimmten Sprachstufen in Zusammenhang?
- Existieren Muster für die Versprachlichung bestimmter Konzepte, und wenn ja, inwieweit lassen sich daraus Prognosen für zukünftigen Sprachwandel ableiten?

1.2 Beschreibung unserer Wortfelder

Statt eines klassischen semasiologischen Herangehens bei der Lemma-Bearbeitung hat DWEE sich, wie beschrieben, für eine Anordnung der behandelten Artikelwörter innerhalb von Wortfeldern entschieden. Ein Wortfeld ist eine semantisch zusammengehörende, in sich taxonomisch strukturierte Gruppe von Lexemen einer bestimmten Wortart, in unserem Falle ausgewählte Substantive des deutschen Wortschatzes. Durch die Analyse von Ober- und Unterbegriffen, die zum Teil ererbt sind, zum Teil aber auch erst im Laufe der Sprachgeschichte entstehen, ergibt sich eine innere Wortfeldstruktur, die wiederum untergeordnete Wortfelder enthalten kann.

Mithilfe dieser onomasiologischen Vorgehensweise kann DWEE die oben genannten Forschungsthemen untersuchen, nämlich wie bestimmte Sachverhalte und Konzepte versprachlicht werden, welche parallelen Repräsentationen existieren und wie sie sich gegebenenfalls in stilistischer, quantitativer oder anderer Hinsicht zueinander verhalten. Durch die zusätzliche Einbindung von Europäismen kann außerdem analysiert werden, inwiefern deutsche Sprachkonzepte sowohl auf diejenigen verwandter und nichtverwandter Sprachen abbildbar sind, als auch sich als universell oder doch kulturspezifisch herausstellen lassen. An einem auf unseren Daten basierenden Forschungsbeispiel soll eine solche Analyse später näher ausgeführt werden.

1 Zur Erstellung unserer Datensammlung nutzen wir verschiedenste literarische Quellen, historische Sprach- und Fachwörterbücher, Lexika und Online-Bibliotheken (bspw. Zeno.org oder dwds.de, beide Stand: 20.5.2025). In der online nutzbaren DWEE-Datenbank (siehe Kap.2) ist unseren aufgenommenen Wörtern jeweils ihr Quellnachweis angefügt.

Unsere Wortfelder erschließen den „Menschen in Natur und Kultur“: Ausgehend von der natürlichen Perspektive „Was ist der Mensch?“ und der philosophischen Frage „Wer ist der Mensch?“ wurden zunächst die biologischen und schließlich die vielzähligen kulturellen Eigenschaften des Menschen in semantische Felder eingeteilt, unsere „Hauptwortfelder“.

- Der Mensch im Alltag
- Mensch und Mitmensch
- Religion und Ethik
- Wirtschaft
- Bildung, Wissenschaft und Kunst
- Technologie

Wortfelder werden als Container semantischer Unterfelder begriffen, in die die Artikelwörter eingeordnet sind. So ergibt sich eine hierarchische Wortfeldstruktur bedeutungsverwandter Begriffe. Das Lemma *Pupille* beispielsweise ist (wie *Iris*) dem Feld *Auge*, dieses wiederum (wie *Mund*, *Stirn* oder *Nase*) dem Feld *Gesicht* zugeordnet usw.:

Mensch > *Körper* > *Kopf* > *Gesicht* > *Auge* > *Pupille*

Es ist aber keineswegs so, dass sich alle Wortfelder in dieser einfachen Weise hierarchisch strukturieren lassen, was eine Modifikation des Theoriemodells erfordert. Ob man also überhaupt von „Wortfeldern“ sprechen kann, wenn Wörter nur lose zusammenhängen und keine weitere Binnenstruktur des Feldes zu erkennen ist, bleibt zunächst eine offene Forschungsfrage, die nach Abschluss der empirischen Arbeit geklärt werden soll.

Den von uns aufgenommenen Lemmata[2] ist jeweils ein Wortfeld-Artikel zugeordnet, der für jede der fünf deutschen Sprachstufen Listeneinträge zu sprachwissenschaftlichen Kategorien aufführt. Neben Synonymen, Hyperonymen, Meronymen, Ableitungen und weiteren Kategorien liegt ein Hauptaugenmerk des Projekts auf der Kategorie der Komposita, die ein Lemma als Vorder- oder Hinterglied bilden kann. Die Produktivität der Komposition zeigt in einer objektiv überprüfbaren Weise an, wann und wie schnell sich ein Lemma im Wortschatz ausbreitet.

2. Technische Infrastruktur

Als ein auf Langfristigkeit ausgelegtes Projekt spielt in DWEE die kontinuierliche Anpassung an technologische Entwicklungen sowie die Reaktion auf Änderungen des wissenschaftlichen und organisatorischen Umfeldes eine herausragende Rolle. Dieser Bedarf lässt sich beispielhaft anhand der ursprünglich geplanten Publikation von Forschungsergebnissen über CD-ROMs – neben einem ebenfalls von Beginn an vorgesehenen Onlineportal – illustrieren, die mittlerweile offensichtlich keine größere Relevanz für Disseminations- und Publikationsmaßnahmen spielen.

Die über die bisherige Projektlaufzeit vorgenommenen Überarbeitungen und Anpassungen betreffen entsprechend in der Breite sowohl die verwendete Software im Bereich Datenerfassung, Ergebnispräsentation und -publikation als auch das Thema Datenhaltung und Datenpublikation in einem sich stetig wandelnden Forschungsumfeld.

2 Unser Wort-Korpus wird sukzessive befüllt. Die Zahlen zum bereits öffentlich zugänglichen Datenbestand sind unter https://dwee.saw-leipzig.de/statistics/de (Stand: 14.7.2025) angegeben.

2.1 Portale und Editoren

Für die Gestaltung und Implementierung der zentralen Anwendungen des Projektes stand von Beginn an die benutzerfreundliche und leicht zugängliche UI-Gestaltung und – für die Editorkomponenten – die Unterstützung von Editorinnen und Editoren mit verschiedenen fachlichen Hintergründen und technischen Vorkenntnissen im Vordergrund.

Als Ergebnis präsentiert DWEE seine öffentlichen Datenbestände in einem eigenen Webportal[3] (siehe Abb. 1), das mehrfach grundlegend überarbeitet wurde. Die aktuelle Version steht seit 2021 der Öffentlichkeit zur Verfügung und wurde insbesondere mit Hinblick auf verbesserte Suchfunktionalitäten, die transparente Darstellung von Artikelstrukturen und interaktive Zugriffsformen erarbeitet. Ziel ist es, verschiedene Sichten auf den vorhandenen Datenbestand gleichrangig zu unterstützen. Dies umfasst klassische, alphabetisch sortierte Lemmalisten mit einfachen Filterfunktionen für einen schnellen Überblick, den Zugriff über den interaktiven Artikelbaum entsprechend der im Fokus des Projekts stehenden hierarchischen Wortfeldstruktur parallel zu den gedruckten Ausgaben, sowie eine Volltextsuche, die umfangreichere Filterfunktionen bzgl. der für die Nutzerinteressen jeweils relevanten Typen gefundener Listeneinträge unterstützt. Die gesamte Entwicklung fand auf Basis einer gesonderten Anforderungsanalyse statt, die in Zusammenarbeit mit interessierten Fachkolleginnen und -kollegen Entwicklungsziele bzgl. relevanter Nutzungsszenarien identifizierte und die Grundlage für die erste Priorisierung der Entwicklungsarbeit war. Seither werden neue Anforderungen in einem inkrementellen Prozess konzeptioniert und umgesetzt.

Abb. 1: Darstellung des Artikels *Basilikum* im DWEE-Portal

Der vorliegende Entwicklungsstand unterliegt weiterhin kontinuierlichen Anpassungen und Verbesserungen. Aktuell vorgesehene Erweiterungen betreffen insbesondere die verbesserte Erschließung der erfassten Komposita über eine gesonderte Suchfunktionalität.

Ebenfalls aufgrund der dargestellten Projektrahmenbedingungen und allgemeinen Entwicklungskriterien liegt der DWEE-Editor „dweedit“ mittlerweile in der dritten Fassung vor (siehe Abb. 2). Auch in diesem Fall stand die nutzerfreundliche Gestaltung und Unterstützung des gemeinsamen Editionsprozesses im Vordergrund. Entsprechend wurden bewusst Datenerfassung und -korrektur nicht über die Integration von XML-Editoren in die Arbeitsprozesse umgesetzt, sondern über eine eigene Webanwendung, die projektspezifische

3 https://dwee.saw-leipzig.de (Stand: 14.7.2025).

Besonderheiten (wie z. B. die Pflege der hierarchischen Wortfeldstruktur oder Publikation/Depublikation im Webportal) gesondert unterstützt.

Abb. 2: Darstellung von Einträgen des Artikels *Basilikum* im DWEE-Editor

2.2 Daten und ihre FAIRe Nutzung

Die frühe Entscheidung für die Nutzung von XML und darauf aufbauender Technologien für die Erstellung und Ablage von Artikeln und Etymologien war ein zentraler Baustein, der die Anpassung der genutzten Anwendungen über die letzten Jahre massiv unterstützt bzw. in Teilen erst ermöglicht hat. Anforderungen an die verwendeten Datenmodelle – getrennt für die hochstrukturierten Artikel und die eher Volltextnahen Etymologien – wurden in projektspezifischen Schemata abgebildet. Wichtige Bausteine der Datenmodell-Konzeption waren u. a. die Abbildung diachroner Angaben für die verschiedenen Sprachstufen, die semantische „Typisierung" erfasster Informationen, die enge Verzahnung zwischen Artikeln und Etymologien sowie die Einbindung im dynamisch erarbeiteten Taxonomie-Baum. Den sich teils aus dem jeweiligen Forschungsstand ergebenden Änderungsbedarfen wurde durch die Anpassung dieser Schemata begegnet.

Trotzdem war – unabhängig von der bzgl. der Projektziele angemessenen Datenhaltung – die gewählte technische Umsetzung für eine Bereitstellung der erarbeiteten Ergebnisse entsprechend der 2016 formulierten FAIR-Prinzipien (Wilkinson et al. 2016) nicht vollumfänglich geeignet. Als Reaktion auf die hohe Relevanz dieser Leitlinien wurde die Projektstrategie bzgl. der Dissemination von Datensätzen systematisch weiterentwickelt. Unter anderen liegen die DWEE-Daten mittlerweile auch als (Linguistic) Linked Open Data, unter Nutzung der Ontolex/Lemon (McCrae et al. 2011) bzw. der LexInfo-Ontologie (Cimiano et al. 2011), vor.[4] Die öffentliche Bereitstellung weiterer lexikalischer Datenformate ist in Vorbereitung.

Auch auf den 2020 begonnenen Aufbau der Nationalen Forschungsdateninfrastruktur (NFDI) wurde im Sinne einer FAIRen Bereitstellung der Projektergebnisse reagiert. Früh wurde der Kontakt zum NFDI-Konsortium Text+[5] hergestellt, das sich den verschiedenen Fachbereichen text- und sprachbasierter Forschung widmet. In einer ersten prototypischen Umsetzung stehen bereits Teile der DWEE-Daten über die dort entwickelte föderierte Wörterbuchplattform LexFCS (Eckart et al. 2023) zur Verfügung (siehe Abb. 3). Im

4 https://dwee.saw-leipzig.de/api (Stand: 20.5.2025).

5 https://text-plus.org/ (Stand: 20.5.2025).

gleichen Kontext ist die langfristige Sicherung der Projektergebnisse über die Langzeitarchivierung in einem etablierten Daten-Repositorium in der Planungsphase.

Abb. 3: Projektergebnisse in der föderierten Wörterbuchplattform des NFDI-Konsortiums Text+ (Prototyp)

3. Untersuchungsbeispiel zum Bedeutungswandel

DWEE sammelt und gliedert Belege seiner aufgenommenen Lemmata aus schriftlichen Quellen aller deutschen Sprachstufen innerhalb der besprochenen sprachwissenschaftlichen Artikel-Kategorien, sodass ein diachroner Vergleich der Wortbedeutungen möglich wird. Wie mithilfe unserer Daten Sprachwandel analysiert und mögliche Ursachen ermittelt werden können, sollen die nachfolgenden Erläuterungen skizzieren. Mit Einbeziehung des kulturellen Kontextes versuchen wir herauszustellen, welche Konnotationen eines Lemmas seit dem Althochdeutschen existieren, welche dagegen erst in späteren Sprachstufen hinzukommen oder aber verblassen bzw. gänzlich aus dem deutschen Lexikon verschwunden sind.

Zu unserer aktuellen Forschungsarbeit gehört das semantische Feld **„Bildung, Wissenschaft und Kunst“**.[6] Die folgenden Ausführungen stellen erste Forschungsergebnisse dar und sollen unsere Herangehensweise zur Analyse von Bedeutungswandel beschreiben.

3.1 „Bildung“

Das Wort *Bildung* geht auf ahd. *bilidunga* zurück, eine Ableitung des etymologisch umstrittenen ahd. *bilidi*, das seit dem 8. Jahrhundert neben ahd. *piladi*, *pilodi*, *biladi* bezeugt ist und sich im 10. Jahrhundert weitestgehend durchgesetzt hat.[7] Seine zahlreichen Bedeutungen gehen auf die unterschiedlichen lateinischen Übersetzungen zurück (wie von *exemplum*, *simulatio*, *forma*, *imaginatio*), so dass das semantische Spektrum im Althochdeutschen von *Abbild*, *Darstellung*, *Nachbildung* über *Muster*, *Vorbild*, *Gleichnis* bis hin zu *äußerer Form*, *Gebilde* und *Gestalt* reicht.

6 Der dieses Wortfeld behandelnde Projektband VI befindet sich in Vorbereitung, weshalb nachfolgende Verweise nur mit Kapitel-Titeln zitiert werden.

7 Vgl. Sächsische Akademie der Wissenschaften zu Leipzig (o. D.): Etymologisches Wörterbuch des Althochdeutschen. https://ewa.saw-leipzig.de/articles/14910 (Stand: 17.1.2026).

3.1.1 Diachroner Vergleich der sprachwissenschaftlichen Kategorien

Die Auswertung unserer bisherigen Daten zum Begriff *Bildung* zeigt, dass dem heutigen Wortverständnis im Sinne einer Fähigkeit, Zusammenhänge zu erfassen oder selbständig abzuleiten, Notkers des Deutschen (um 950–1022) Auffassung von *bilidunga* innerhalb seiner Erkenntnislehre (bzw. seiner Übersetzung von Boethius' „De consolatione Philosophiae") schon sehr nahekommt. Dort bezeichnet der Begriff „nicht nur die Gestalt, die äußere Erscheinung der Dinge, sondern auch das Bild, das der Mensch mit Hilfe seiner Sinne von der Außenwelt empfängt […], ferner den Begriff, den der Verstand daraus ableitet, und zuletzt das allen Dingen zugrunde liegende Urbild, die göttliche Idee".[8] Mit dieser Betrachtungsweise Notkers lässt sich eine semantische Brücke zum modernen Konzept von „Bildung" sogar noch früher schlagen als durch die späteren theologischen und philosophischen Abhandlungen Meister Eckharts (um 1260–1328) im Spätmittelalter.[9] Dieser Schluss lässt sich durch den im DWEE-Korpus abgebildeten Wortgebrauch im Alt- bis zum Neuhochdeutschen, das heißt durch ein Gegenüberstellen der Bedeutungen innerhalb der mit *Bildung* verwendeten Wortbildungen, verfestigen.

Heutige Hauptkonnotationen – *Wissenserwerb*, *Erziehung*, *Anstand*, *Formung* usw. – sind also teils aus den älteren Sprachstufen übernommen, teils durch Sprachwandelphänomene verschoben worden oder neu entstanden. Und obwohl wir bei Notker den Begriff *Bildung* schon hinsichtlich einer geistigen Entwicklung vorfinden können, etabliert sich diese und die ihr verwandten Wortbedeutungen (wie *erlernte Kenntnisse* oder *Erziehung*) erst spät im deutschen Lexikon. Um einzugrenzen, wann und über welche Schritte sich der Bedeutungswandel vollzogen hat, können unsere bisherigen Funde, in Tabelle 1 Auszüge der nominalen Derivationen mit Präfix[10], zur Untersuchung herangezogen werden.

Es lässt sich deutlich erkennen, dass sich eine Bedeutungsverschiebung des kognitiven Bildungsbegriffs vom theologischen und mystischen Verständnis (als *Urbild*, *Eingebung*, *religiöse Anschauung*), das noch bis ins Frühneuhochdeutsche vorherrscht, hin zu einer in gesellschaftspolitischem Kontext konnotierten Verwendung (im Sinne von *Lehrmaßnahme*, *Kenntnisvermittlung*) vollzogen hat. Die Verdrängung der mystisch-theologischen Semantik geschieht parallel zur historischen Epoche der naturwissenschaftlichen Revolution (16. und 17. Jh.) und der Aufklärung (18. Jh.). Die im allgemeindeutschen Wortschatz neu aufgekommenen Ableitungen zu *Bildung* als *Wissens-* bzw. *Fähigkeitsvermittlung* (wie *Fortbildung*, *Weiterbildung*) korrelieren wiederum mit Reformen in Sozialwesen und Pädagogik. Durch Einflüsse in diese Fachgebiete durch neue didaktische Modelle, in der frühen Neuzeit maßgeblich jene von Johann Heinrich Pestalozzi (1746–1827) und Maria Montessori (1870–1952)[11], sowie die sich ändernde Bildungspolitik Deutschlands nimmt außerdem die Anzahl der Komposita mit *Bildung* (mit der Bedeutung *Wissensvermittlung*, *erworbenes Wissen*) seit dem Älteren Neuhochdeutschen zu (vgl. Tab. 2).

8 Sächsische Akademie der Wissenschaften zu Leipzig (o. D.): Althochdeutsches Wörterbuch. https://awb.saw-leipzig.de/?sigle=AWB&lemid=B01213 (Stand: 14.6.2025).

9 Zumeist gilt Meister Eckhart als Begründer des modernen Bildungsbegriffes, vgl. hierzu ausführlicher Zeilfelder/Irslinger 2025, Kap. „*bildunge* bei Meister Eckhart".

10 Nachfolgende Aufzählungen in Tabelle 1 zeigen Auszüge aus den jeweils behandelten Kategorien. Vollständige Einträge samt ihren Quellenangaben sind demnächst auf unserem Webportal (https://dwee.saw-leipzig.de, Stand: 14.7.2025) abrufbar.

11 Die Einwirkung von historischen Entwicklungen in Pädagogik, Gesellschafts- und Sozialpolitik auf den Bedeutungswandel besprechen eingehender Zeilfelder/Irslinger (2025, Kap. „Abgrenzung von Bildung und Erziehung").

Nominale Präfixableitungen mit *bildunge* im Mittelhochdeutschen	
widerbildunge	Schaffung eines Ebenbildes; Einbildungskraft
īnbildunge	Einbildung; Einprägung, Prägung; Eingebung; Äußerlichkeit; Illusion, Vorstellung; Vorstellungskraft; Urbild
nāchbildunge	Nachbildung, Imitat
verbildunge	Verunstaltung; Umänderung, Umformung
vorbildunge	Vorbedeutung, Vorzeichen
Nominale Präfixableitungen mit *bildung* im Frühneuhochdeutschen	
ausbildung	Vorbildlichkeit, Musterhaftigkeit
einbildung	Einbildung; Illusion; Einsicht, Anschauung (Theologie); Selbstüberschätzung
nāchbildung	Nachbildung, Imitat
Nominale Präfixableitungen mit *Bildung* im Neuhochdeutschen	
Ausbildung	Vermittlung von Wissen und Fähigkeiten, die für spezifische Tätigkeiten qualifizieren; Entstehung, Entfaltung (Biologie, Botanik, Psychologie); Gestaltung, Formung
Einbildung	Hochmut, Selbstüberschätzung; falsche Vorstellung, Illusion
Fortbildung	Lehrmaßnahme zur Erweiterung beruflicher Fähigkeiten
Nachbildung	Imitat, Attrappe, Fälschung
Rückbildung	Rückentwicklung (Medizin, Biologie); Art der Wortbildung (Linguistik)
Verbildung	Deformation, Fehlentwicklung
Weiterbildung	Erweiterung von Kenntnissen und Kompetenzen

Tab. 1: Untersuchungen zum Bedeutungswandel anhand von Ableitungen

Hyponyme mit *Bildung* im Älteren Neuhochdeutschen	
Elementarbildung	Vermittlung von Wissen und Fähigkeiten im Kindesalter
Kunstbildung	Vermittlung von Kunstverständnis
Volksbildung	Vermittlung von Wissen und Fähigkeiten an Erwachsene
Hyponyme mit *Bildung* im Neuhochdeutschen	
Lehrerbildung	Schulung für den Lehrberuf
Volksbildung	Vermittlung von Wissen und Fähigkeiten an Erwachsene, organisierte Wissensvermittlung eines Staates
Determinativkomposita mit *Bildung* als Vorderglied im Neuhochdeutschen	
Bildungsabschluss	erworbene Qualifikation
Bildungsarmut	sozial oder wirtschaftlich bedingter Mangel an Kenntnissen
Bildungsauftrag	Verpflichtung zur Wissensvermittlung
Bildungspolitik	Bestimmungen für das Lehrwesen; staatliches Ressort für Schule und Ausbildung
Bildungsschranke	sozial oder wirtschaftlich bedingte Hinderung der Teilhabe an Wissensvermittlung

Tab. 2: Untersuchungen zum Bedeutungswandel anhand von Komposita

Für eine vertiefende Untersuchung des Bedeutungswandels mithilfe der Komposita strukturiert DWEE die gesammelten Wortzusammensetzungen zusätzlich nach ihren grammatischen und semantischen Kategorien. So lässt sich z. B. gezielt nach adjektivischen Determinations- oder nach Klammerkomposita suchen, aber auch nach spezifischer Semantik im Vorder- und/oder Hinterglied, wie nach Pflanzen-, Material- oder präpositionaler Bezeichnung.

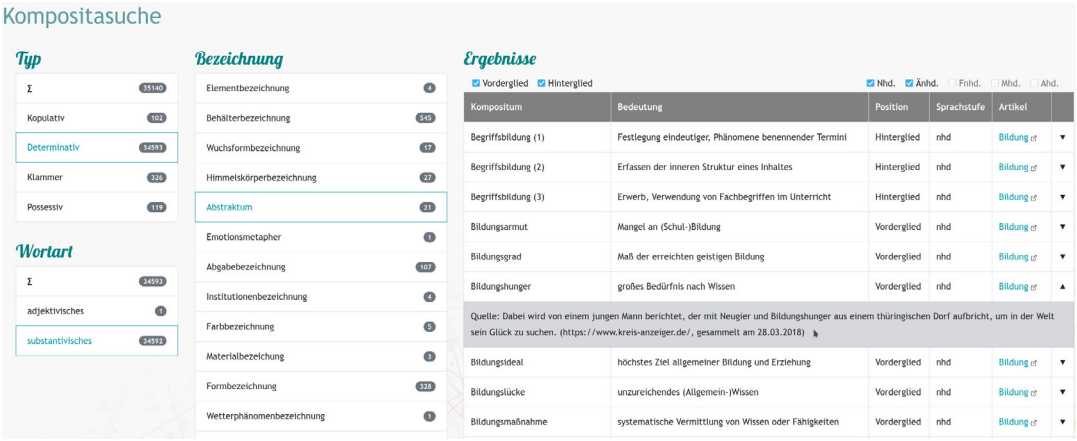

Abb. 4: Komposita-Suche im DWEE-Portal (Prototyp)

Die Zunahme der Komposita korrespondiert mit den angerissenen gesellschaftlichen Phänomenen einerseits und der zunehmenden Notwendigkeit der Differenzierung in sozialen, wirtschaftlichen und fachlichen Bereichen auf der anderen Seite. Im Laufe der erhöhten Publikation von Sach- und Lehrbüchern seit der Aufklärungszeit und fortwährendem wissenschaftlichem Fortschritt zeigt sich der Bedarf nach genaueren Abgrenzungen oder Wortneubildungen. Wir erkennen einen Anstieg der Zusammensetzungen hinsichtlich der semantischen Verwendung als *etwas Entstandenes*, *Gewachsenes* seit dem 17. Jahrhundert in vielen Bereichen, z. B. der Botanik (*Knospenbildung*, *Blattbildung*), Medizin und Zoologie (*Blutbildung*, *Knochenbildung*), der Philosophie und Psychologie (*Urteilsbildung*, *Willensbildung*), der Physik und Chemie (*Kristallbildung*, *Säurebildung*) oder auch der Geologie (*Felsbildung*, *Talbildung*).

3.1.2 Diachroner Vergleich der Wortfelder

Innerhalb des DWEE-Hauptwortfeldes „Bildung, Wissenschaft und Kunst“ steht *Bildung* in seiner Verwendung als gesellschaftskultureller Begriff. Wir unterscheiden weitere semantische Teilfelder wie *Erziehung*, *Wissenschaft* oder *Schule* und schlüsseln auch diese wiederum in eigene Untergruppen auf (zu *Schule* bspw. *Unterricht*, *Übung*, *Gymnasium*, *Sportschule*, *Schüler* usw.). Dabei ist die Abgrenzung oft schwierig oder fluide, wie man bei den bedeutungsnahen Begriffen *Lehre*, *Bildung* und *Erziehung* erahnen kann. Um uns einer hierarchischen Gruppierung bestmöglich anzunähern, werden die Wortfelder in ihrer jeweiligen Sprachstufe untersucht. Grundlage bildet dazu zunächst die semantische Taxonomie im Neuhochdeutschen. Anhand eines Vergleichs der Wortfunde und ihrer Bedeutungen in älteren deutschen Sprachstufen lässt sich erkennen, wann sich welche Abgrenzung oder auch ein Zusammenfall von Konnotationen vollzogen hat. Daraus kann wiederum abgeleitet werden, ob sich Gemeinsamkeiten innerhalb eines Wortfeldes evaluieren lassen oder sich dessen bedeutungsverwandte Wörter heterogen entwickeln.

Als Beispiele für einen Sprachwandel des deutschen Wortschatzes im hier skizzierten Wortfeld seien die Wörter *List* und *Edukation* genannt: Während im Mittelhochdeutschen *List* neben *Täuschung* noch als *Wissen*, *Erfahrung*, *Können*, *Kunst* u. w. mit den Bedeutungsmerkmalen <Intelligenz> und <erlerntes Wissen> auftritt, ist sie mittlerweile gänzlich aus dem Wortfeld *Bildung* verdrängt und ihre Konnotation *Erlerntes* verblasst.[12] Dagegen gibt es Entlehnungen, die erst spät in dieser semantischen Gruppe auftauchen, wie *Edukation* (fachsprachlich: *Erziehung*, *Belehrung*, *Ausbildung*), das im 16. Jahrhundert aus lat. *ēducātio* entlehnt und im späteren Früh- und Älteren Neuhochdeutschen von franz. *éducation* beeinflusst wurde.[13]

3.2 „Bildung“ im europäischen Kontext

Unsere Datenerhebungen und deren erste Untersuchungen zusammenfassend lässt sich erkennen, dass der Wortgebrauch von *Bildung* im Sinne von *Form* und *Gestalt* seit dem Mittelhochdeutschen abnimmt, während sich ein Anstieg der Verwendung als *etwas Entstandenes*, *Geschaffenes* oder *Gegründetes* abzeichnet. Inwiefern diese Tendenz auch in anderen europäischen Sprachen vorzufinden ist, wird zurzeit untersucht. Es lassen sich

12 Hierzu und zu weiterem Sprachwandel des Wortfelds vgl. die weiteren Ausführungen von Zeilfelder/Irslinger (2025, Kap. „Lehre“).

13 Vgl. Berlin-Brandenburgischen Akademie der Wissenschaften (o. D.): Digitales Wörterbuch der deutschen Sprache. www.dwds.de/wb/Edukation (Stand: 15.6.2025).

jedoch bereits allgemeine Konzeptmetaphern mit teils langer Tradition herausstellen, die als Europäismen gelten können:

> *Bildung* als Wanderung ‚nach vorn'
>
> *Bildungsweg*: engl. *course* (*Kurs*) *of study*, span. *vía* (*Straße*) *de formación*, russ. *учебная тропа* (*Pfad*) [*učebnaja tropa*]
>
> *Bildung* als Wanderung ‚nach oben'
>
> *Bildungsleiter*, *Bildungstreppe*: engl. *educational ladder*, alb. *shkallë arsimore/edukatore*
>
> *Bildungsniveau*, *Bildungsstufe*: engl. *level of education*, ital. *livello culturale*, port. *nivel de formação*
>
> *Bildung* als angehäuftes Gut
>
> *Bildungsgut*: ital. *bagaglio* (*Gepäck*) *culturale*, port. *patrimonio* (*Besitz*) *cultural*, rum. *bun* (*Besitz*) *educațional*

Eines der sprachhistorisch prägnantesten Konzepte ist die metaphorische Übertragung von Aufzucht und Pflege im Pflanzen- und Ackerbau auf die Begriffe des Wortfelds *Bildung* (wie z. B. *Schulzweig*, *Schulreife*, *Zögling*). Bereits in der Klassischen Antike nutzt Cicero (106–43 v.Chr.) in seinen „Gesprächen in Tusculum" diese Verbildlichung: „wie ein Acker, auch wenn er fruchtbar ist, ohne Pflege keine Frucht tragen kann, so auch die Seele nicht ohne Unterweisung".[14] Diese Wachstums- bzw. Ackerbaumetaphorik setzt sich diachron fort, wie bspw. De Haan (1991, S. 364) belegt: „Hatten ‚gute Gärtner oder Bauern' Fröbels Anstalten noch bevölkert, so ist es der durch Wissenschaft und Technik geschulte ‚moderne Bauer', der den Kindergarten Montessoris bestellt."

Auch noch im heutigen europaweiten Sprachgebrauch lassen sich viele Beispiele für dieses Konzept finden. So kann *Bildung* Früchte tragen (span. *hacer fructificar esta experiencia – diese Erfahrung Früchte tragen lassen*) oder verkümmern (engl. *isolation can cause ideas to wither – Isolation kann Ideen verkümmern/welken lassen*), mit ihr kann eine Saat gelegt werden (franz. l*e livre plantera les semences de tant de ses pensées profondes – das Buch wird die Saat für viele seiner großen Ideen legen/pflanzen*) oder sie kann sich verzweigen (russ. *ветвь в древе знаний* [*vetv' vdreve znanij*] – *Zweig vom Baum des Wissens*). Man kann hier eine universelle Konzeptmetapher vermuten, zumindest erweist sie sich für Acker- und Gartenbaukulturen als weit verbreitet.

Danksagung

Wir bedanken uns bei Projektleiterin Rosemarie Lühr sowie bei allen weiteren aktuellen und ehemaligen DWEE-Mitarbeitern: Bettina Bock, Britta Irslinger, Stefan Lotze, Sabine Ziegler und Stefan Wegerhoff. Ein spezieller Dank gilt unserer Arbeitsstellenleiterin Susanne Zeilfelder für ihre Unterstützung bei der Formulierung oben angeführter Forschungsfragen.

14 Übersetzt nach Gigon (1984, S. 123 f.); vgl. auch Zeilfelder/Irslinger (2025, Kap. „Abgrenzung von Bildung und Erziehung").

Literatur

Berlin-Brandenburgischen Akademie der Wissenschaften (o. D.): Digitales Wörterbuch der deutschen Sprache. www.dwds.de/wb/Edukation (Stand: 15.6.2025).

Cimiano, Philipp/Buitelaar, Paul/McCrae, John/Sintek, Michael (2011): LexInfo: A declarative model for the lexiconontology interface. In: Journal of Web Semantics 9, 1, S. 29–51. https://pub.uni-bielefeld.de/record/2002936 (Stand: 17.1.2026).

De Haan, Gerhard (1991): Über Metaphern im pädagogischen Denken. In: Zeitschrift für Pädagogik 27 (Beiheft), S. 361–375.

Eckart, Thomas/Herold, Axel/Körner, Erik/Wiegand, Frank (2023): A federated search and retrieval platform for lexical resources in Text+ and CLARIN. In: Electronic Lexicography in the 21st Century. Proceedings of the 8th eLex 2023 Conference, Brno, Czech Republic, 27–29 June 2023, S. 280–292. https://elex.link/ojs/index.php/elex/article/view/31/17 (Stand: 20.11.2025).

Gigon, Olof (Hg.) (1984): Marcus Tullius Cicero. Gespräche in Tusculum. Lateinisch-deutsch. 5., durchges. Aufl. München: Artemis.

McCrae, John/Spohr, Dennis/Cimiano, Philipp (2011): Linking lexical resources and ontologies on the Semantic Web with Lemon. In: Antoniou, Grigoris/Grobelnik, Marko/Simperl, Elena/Parsia, Bijan/Plexousakis, Dimitris/Leenheer, Pieter/Pan, Jeff (Hg.): The Semantic Web: Research and Applications. Proceedings of the 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29–June 2, 2011. Part I. Berlin/Heidelberg: Springer, S. 245–259.

Sächsische Akademie der Wissenschaften zu Leipzig (o. D.): Althochdeutsches Wörterbuch. https://awb.saw-leipzig.de/?sigle=AWB&lemid=B01213 (Stand: 14.6.2025).

Sächsische Akademie der Wissenschaften zu Leipzig (o. D.): Etymologisches Wörterbuch des Althochdeutschen. https://ewa.saw-leipzig.de/articles/14910 (Stand: 17.1.2025).

Trier, Jost (1931): Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes. Heidelberg: Winter.

Wilkinson, Mark D./Dumontier, Michel/Aalbersberg, IJsbrand Jan/Appleton, Gabrielle/Axton, Myles/Baak, Arie/Blomberg, Niklas/Boiten, Jan-Willem/da Silva Santos, Luiz Bonino/Bourne, Philip E./Bouwman, Jildau/Brookes, Anthony J./Clark, Tim/Crosas, Mercè/Dillo, Ingrid/Dumon, Olivier/Edmunds, Scott/Evelo, Chris T./Finkers, Richard/Gonzalez-Beltran, Alejandra/Gray, Alasdair J. G./Groth, Paul/Goble, Carole/S. Grethe, Jeffrey/Heringa, Jaap/'t Hoen, Peter A. C/Hooft, Rob/Kuhn, Tobias/Kok, Ruben/Kok, Joost/Lusher, Scott J./Martone, Maryann E./Mons, Albert/Packer, Abel L./Persson, Bengt/Rocca-Serra, Philippe/Roos, Marco/van Schaik, Rene/Sansone, Susanna-Assunta/Schultes, Erik/Sengstag, Thierry/Slater, Ted/Strawn, George/Swertz, Morris A./Thompson, Mark/van der Lei, Johan/van Mulligen, Erik/Velterop, Jan/Waagmeester, Andra/Wittenburg, Peter/Wolstencroft, Katherine/Zhao, Jun/Mons, Barend (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific Data 3, 1, S. 1–9. https://doi.org/10.1038/sdata.2016.18 (Stand: 15.6.2025).

Zeilfelder, Susanne/Irslinger, Britta (2025): Bildung, Wissenschaft und Kunst. (= Deutsche Wortfeldetymologie in europäischen Kontext 6). Wiesbaden: Reichert. [Unveröffentlichtes Manuskript].

Kontaktinformation

Ulrike Ertel
Sächsische Akademie der Wissenschaften zu Leipzig
Karl-Tauchnitz-Str. 1
04107 Leipzig
E-Mail: ulrike.ertel@uni-jena.de

Thomas Eckart
Sächsische Akademie der Wissenschaften zu Leipzig
Karl-Tauchnitz-Str. 1
04107 Leipzig
E-Mail: eckart@saw-leipzig.de

Christine Rains
Sächsische Akademie der Wissenschaften zu Leipzig
Karl-Tauchnitz-Str. 1
04107 Leipzig
E-Mail: drains@smail.uni-koeln.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.85

Susanne Kabatnik/Anne Klee

Historischer Pandemiewortschatz und Linked Open Data (LOD)

Lexikalische Veränderungen und kulturelle Reaktionen im Spiegel der Cholera

Abstract The study examines pandemic-related language change using historical lexicons on cholera. It is based on a cholera corpus generated from the historical archives of the DWDS and DeReKo, which was analysed using Sketch Engine and examined for frequencies, compound words and semantic patterns. Historical dictionaries and secondary literature were also consulted to explore concepts such as the feminisation of the disease. The analysis shows that pandemic language change is particularly visible in thematically focused fields such as infrastructure and medical measures. Frequency data show wave-like patterns, with pandemic compound words following productive patterns. Semantic fields are cyclically reactivated across pandemics. Methodologically, the study combines corpus linguistic methods with cultural studies contextualisation and links the results in Linked Open Data structures, enabling comparative analyses between pandemics. The study demonstrates how historical language data can be integrated into modern digital infrastructures, thus opening new perspectives for diachronic analyses of pandemic vocabulary.

Keywords digital lexicography, linked open data, corpus linguistics, language change, historical pandemic vocabulary

1. Untersuchung von Pandemiewortschatz als Sprachwandelphänomen

Sprachwandel kann sich in zahlreichen Bereichen und auf verschiedenen linguistischen Ebenen vollziehen. Besonders augenfällig sind lexikalische Veränderungen wie die Entstehung neuer Lexeme, Neologismen, Wortneubildungen oder Entlehnungen (vgl. Innerwinkler 2015, S. 7 f.). Auch auf semantischer Ebene lassen sich Prozesse wie Bedeutungsveränderungen, -erweiterungen oder -spezifikationen beobachten (vgl. Traugott 2000 S. 385–387). Diesen Prozessen liegt ein gemeinsamer Mechanismus zugrunde: Veränderungen der Lebenswirklichkeit, etwa durch technischen, sozialen oder kulturellen Fortschritt, die sich auf der sprachlichen Ebene niederschlagen. Sprachverwendung ist somit nicht nur Mittel der Kommunikation, sondern auch Ausdruck zeittypischer Einstellungen, kultureller Praktiken und Denkweisen (vgl. Linke 2003, S. 44). Entsprechend erlaubt der Blick auf die Sprachverwendung früherer Zeitstufen Rückschlüsse auf gesellschaftliche Gegebenheiten, kulturelle Dynamiken und kollektive Haltungen einer bestimmten Epoche (vgl. Epps 2015, S. 579 f.).

Pandemien – also Erkrankungen mit weltweiter Ausbreitung – beeinflussen sowohl zahlreiche Lebensbereiche als auch den Wortschatz von Einzelsprachen (vgl. Klosa-Kückelhaus/Kernerman 2022; Ivanov et al. 2023). Die Untersuchung pandemiebedingter sprachlicher Veränderungen ist somit immer auch eine Untersuchung von Sprachwandelphänomenen. Am Beispiel der jüngsten Corona-Pandemie lässt sich dieser Zusammenhang besonders gut veranschaulichen: Im Deutschen entstand eine Vielzahl neuer Ausdrücke und neuer Wortbedeutungen, etwa durch Neubildungen oder Bedeutungsveränderungen (vgl. Balnat 2020, S. 3f.; Möhrs 2021, S. 34; Darwish 2020; Depner 2022; Klosa-Kückelhaus 2022; Nossem 2023). So bezieht sich der Ausdruck *Ellenbogengruß* auf eine neu etablierte Begrüßungsform, die der Vermeidung physischer Nähe und damit der Eindämmung von Infektionen diente. *Balkonkonzert* wiederum verweist auf eine während des Lockdowns verbreitete Praktik, musikalische Darbietungen vom Balkon aus für die Nachbarschaft zu veranstalten. Die Wortneuschöpfungen spiegeln spezifische gesellschaftliche Reaktionen und Handlungsweisen in der pandemischen Ausnahmesituation wider.

Auch Entlehnungen wie *Homeoffice*[1] zeigen, dass pandemiebedingter Wortschatz über das akute Krisenereignis hinaus im allgemeinen Sprachgebrauch verbleiben kann (vgl. Klosa-Kückelhaus/Kernerman 2022). Die ursprünglich als temporäre Maßnahme eingeführte Heimarbeit entwickelte sich – sprachlich wie gesellschaftlich – zur etablierten Praxis. Der pandemiebezogene Wortschatz rund um COVID-19 ist mittlerweile gut dokumentiert und erforscht (vgl. Corona-Glossar des IDS 2024; Jakosz/Kałasznik (Hg.) 2023). Zahlreiche sprachwissenschaftliche Studien haben den durch die Corona-Pandemie bedingten Wortschatzwandel dokumentiert und analysiert: Ein zentrales Ergebnis ist dabei die hohe Produktivität im Bereich der Wortbildung, insbesondere der Komposition durch zahlreiche Corona-Komposita wie *Corona-Virus*, *Corona-Maske* oder *Coronakilo* (vgl. Fuchs 2021; Möhrs 2021; Balnat 2020), weiter auch bei Virusbezeichnungen wie *SARS-CoV-2* oder *Wuhan-Virus* (vgl. Klosa-Kückelhaus 2022; Nossem 2023) sowie Kontaminationen wie *Haustralien* (aus *Haus* und *Australien*) für die durch die Quarantäne eingeschränkte Bewegungsfreiheit (vgl. Balnat 2020, S. 3f.). Weiter frequent war der Einsatz von Metaphern, insbesondere aus den semantischen Feldern des Militärischen und der Natur. Beispiele wie *Virusfront*, *Virenbombe* oder *Corona-Welle* betonen dabei die Rahmung der Pandemie als Kampf oder Naturereignis (vgl. Klosa-Kückelhaus 2022; Depner 2022). Weiter belegen quantitative Untersuchungen pandemiebedingte Frequenzverschiebungen im Wortschatzgebrauch. So zeigt sich eine erhöhte Frequenz in zentralen Handlungsfeldern der Pandemie wie: Dynamik der Ausbreitung, betroffene Regionen, gesellschaftliche und wirtschaftliche Auswirkungen sowie Maßnahmen zur Infektionsprävention. Zudem nehmen Wortschatzbereiche wie Emotionen, Medizin und Wirtschaft eine prominente Rolle ein (vgl. Möhrs 2021, S. 34–36). Die Dynamik des Corona-Wortschatzes veranschaulicht so exemplarisch, wie sich gesellschaftliche Krisen in lexikalischen Veränderungen spiegeln. Sprachliche Anpassungen dokumentieren dabei kollektive Reaktionsmuster und machen Krisenbewältigungsstrategien sichtbar (vgl. Klosa-Kückelhaus 2022; Klosa-Kückelhaus/Kernerman 2022). Trotz

1 Das Lexem *Homeoffice* war bereits vor Ausbruch der Corona-Pandemie im Sprachgebrauch präsent. Eine Analyse der Frequenzverläufe im DWDS-Korpus zeigt ab dem Jahr 2012 zunächst einen moderaten Anstieg, der im Jahr 2020 in einen steilen Frequenzzuwachs übergeht. Ab 2022 ist ein erneuter Rückgang zu verzeichnen. Diese Entwicklung verweist auf eine pandemiebedingte Etablierung des Ausdrucks, der sich im Zuge der veränderten Arbeitsrealitäten temporär deutlich im allgemeinen Sprachgebrauch verankert hat. (vgl. DWDS 2025: Homeoffice; www.dwds.de/r/plot/?view=1&corpus=zeitungenxl&norm=date%2Bclass&smooth=spline&genres=0&grand=1&slice=1&prune=0&window=0&wbase=0&logavg=0&logscale=0&xrange=1946%3A2024&q1=%7B%27Homeoffice%27%2C%27Home-Office%27%7D, Stand: 12.6.2025).

der Relevanz solcher Analysen wurde bislang vergleichsweise wenig Forschung zum Pandemiewortschatz historischer Pandemien betrieben (vgl. Lanza et al. 2022). Dabei ist die Dokumentation älterer pandemiebedingter Lexik nicht nur für das Verständnis historischer Sprachzustände von Bedeutung, sondern liefert auch wertvolle Erkenntnisse über langfristige sprachliche Entwicklungen und deren Auswirkungen auf das heutige Deutsch (vgl. Ernst 2021, S. 30–32). Linguistische Untersuchungen zum Wortschatz vergangener Pandemien – etwa der Spanischen Grippe oder der Pest – liegen bislang nur vereinzelt vor. Im Kontext der Spanischen Grippe wurden vergleichende Studien zum Pandemie-Wortschatz angestellt. So analysieren Lanza et al. (2022) englisch- und italienischsprachige Korpora und vergleichen die Sprachverwendung während der Spanischen Grippe mit derjenigen zur COVID-19-Pandemie. Dabei zeigt sich unter anderem, dass die Spanische Grippe häufig durch abschwächende Adjektive beschrieben wurde (vgl. Lanza et al. 2022, S. 945, 948). Thematisch ähneln sich beide Pandemien in ihren lexikalischen Feldern: staatliche Maßnahmen, soziale Distanzierung, Maskenpflicht sowie Metaphern aus dem Bereich von Militär und Krieg sind zentrale semantische Domänen (vgl. ebd., S. 947). Auch zur Pest liegen punktuelle lexikalische Untersuchungen vor. Moulin (2020) analysiert unter anderem historische Komposita mit *Pestilenz* als Erstglied, etwa *Pestilenzbette*, und weist zugleich auf die sprachhistorische Entwicklung hin, bei der *Pestilenz* zur kürzeren Form *Pest* reduziert wurde – ein Beispiel für lexikalische Vereinfachung (vgl. Moulin 2020, S. 269–271). Interessant ist zudem der Befund, dass auch in heutigen Kontexten pandemiebedingte Wortbildungen wie *Corona-Abitur* in einem semantischen Zusammenhang zu früheren Krankheitsbezeichnungen treten können.

Ziel unseres Vorhabens ist es, an linguistische Forschung zum historischen Pandemiewortschatz anzuknüpfen und diese systematisch weiterzuentwickeln. Im Zentrum steht die Analyse pandemiebezogener Lexik in historischen Textkorpora sowie deren lexikografische Aufbereitung in Form eines digitalen Wörterbuchs sowie einer Linked Open Data-Datenbank zu historischen Pandemien. Die Datengrundlage der Analyse bilden daher primär die historischen Korpora des DWDS (2024) sowie das DeReKo-Korpus (HIST-gesamt 2024; Durrell et al. 2012). Ergänzend wurden historische Wörterbücher (u. a. aus dem Trierer Wörterbuchnetz und dem DWDS) sowie pandemiebezogene Sekundärliteratur (z. B. Dettke 1995; Aselmeyer 2015) einbezogen, um eine möglichst breit fundierte Datengrundlage für die lexikografische Erschließung zu gewährleisten. Auf dieser Basis wurde ein themenspezifisches Cholera-Korpus[2] erstellt, das mittels Sketch Engine (Kilgarriff et al. 2004) auf Wortfrequenzen, Schlüsselwörter, Komposita und Kollokationen analysiert wurde. Das daraus entwickelte Cholera-Glossar umfasst derzeit rund 1.300 Lemmata, die lexikografisch bearbeitet und in zwei sich ergänzenden Medienformen veröffentlicht werden: Zum einen im *Pandemictionary*, einem konventionellen digitalen Wörterbuch, das Informationen zu Aussprache, grammatikalischer Kategorisierung, Wortbildung und ausgewählten Korpusbelegen bereitstellt. Zum anderen in der PandeLexBase, einer umfassenden LOD-Datenbank, in der sämtliche Begriffe und deren Kontextinformationen als semantisch strukturierte Tripel im Sinne des Linked-Open-Data-Paradigmas (Chiarcos/Nordhoff/Hellmann 2012; Cimiano/McCrae/Buitelaar (Hg.) 2016) maschinenlesbar aufbereitet und die Korpusdaten auch öffentlich verfügbar gemacht werden.

Im Rahmen dieses Beitrags stellen wir zunächst das an der Universität Trier verortete Forschungsprojekt vor, welches in das durch die Forschungsinitiative Rheinland-Pfalz geförderte Verbundvorhaben Linked Open Data in den Geisteswissenschaften (LODinG) ein-

2 Das Cholera-Korpus ist derzeit nur projektintern zugänglich, wird jedoch mit der Integration der Korpusdaten in die Wikibase auch online öffentlich verfügbar gemacht.

gebunden ist. Wir gewähren dann exemplarische Einblicke in die Untersuchung des historischen Cholera-Wortschatzes, wobei sowohl methodische Aspekte – insbesondere die Korpusgenerierung und die Einbettung in das Linked-Open-Data-Paradigma – als auch lexikalische Besonderheiten im Fokus stehen.

2. Das Projekt: Historischer Pandemiewortschatz als Linked Open Data

Das Trierer Projekt zur Erforschung des historischen Pandemiewortschatzes ist Bestandteil des interdisziplinären Forschungsverbunds *Linked Open Data in den Geisteswissenschaften* (LODinG), der durch die Forschungsinitiative Rheinland-Pfalz gefördert wird. In insgesamt sieben Teilprojekten vereint LODinG verschiedene Fachrichtungen – darunter Germanistik, Romanistik, Geschichte, Sinologie, Rechtswissenschaft, Informatik, Computerlinguistik, Digital Humanities sowie die Universitätsbibliothek – mit dem gemeinsamen Ziel, die Möglichkeiten strukturierter, vernetzter Daten nach den Prinzipien von Linked Open Data (LOD) für die geisteswissenschaftliche Forschung zu erproben.

Das hier vorgestellte lexikografische Teilprojekt reagiert auf eine bislang bestehende Forschungslücke: die systematische Erhebung und Analyse pandemiebezogener Lexik in historischen deutschsprachigen Texten. Im Mittelpunkt der ersten Projektphase steht die Aufarbeitung des Sprachgebrauchs während der Cholerapandemien. Ziel ist es, pandemietypische sprachliche Phänomene in ihren historischen Kontext einzubetten und digital aufzubereiten. Dabei geht das Projekt über eine rein dokumentarische Erhebung hinaus und schafft durch kontextualisierte Aufbereitung eine Basis für diachrone und kontrastive Studien sowie die Rekonstruktion gesellschaftlicher Wahrnehmungs- und Verarbeitungsprozesse auf sprachlicher Ebene.

Die Erfassung und Bereitstellung des Wortschatzes erfolgt in zwei komplementären digitalen Ressourcen: dem *Pandemictionary*, einem benutzerorientierten Online-Wörterbuch in Anlehnung an das Wiktionary, sowie der PandeLexBase, einer Graphdatenbank auf Basis der Wikibase-Technologie. Während das *Pandemictionary* in kompakter Form lexikalische Informationen präsentiert, ermöglicht die PandeLexBase eine feingranulare, LOD-konforme Modellierung der Einträge. Beide Ressourcen richten sich sowohl an wissenschaftlich Forschende – z. B. aus den Bereichen Sprach- und Literaturwissenschaft, Medizingeschichte oder Soziologie – als auch an interessierte Laien (vgl. Kabatnik et al. im Dr.).

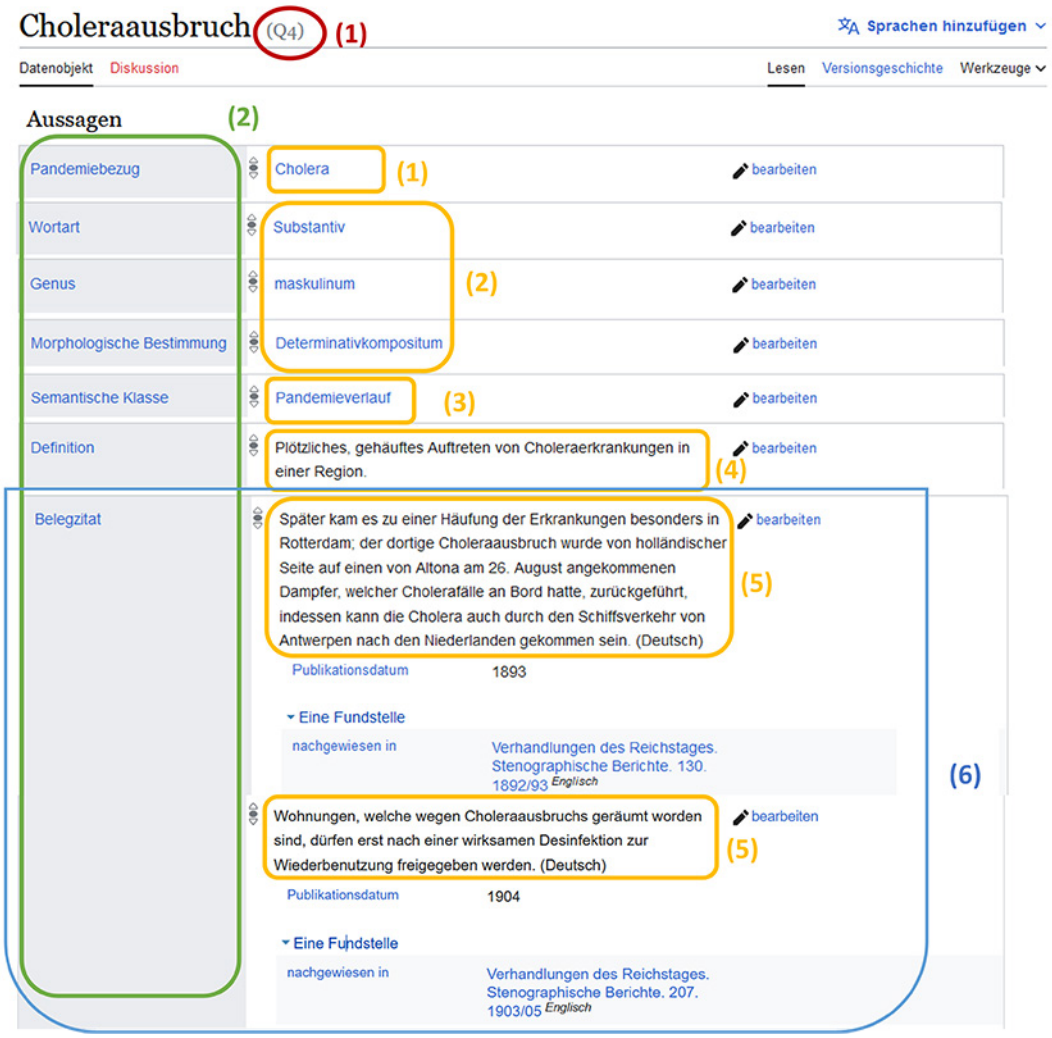

Abb. 1: Das Stichwort *Choleraausbruch* in der PandeLexBase

Als zentrales Element des Projekts fungiert die Wikibaseinstanz des Projekts, die PandeLexBase (siehe Abb. 1). Sie erlaubt eine strukturierte Abbildung pandemierelevanter Begriffe in Form eines graphbasierten Triplestores, einer Datenbank, die Daten in Subjekt-Prädikat-Objekt-Strukturen verwaltet und abfragbar macht. In der PandeLexBase stehen die Pandemiebegriffe an der Subjektposition und werden mit stabilen und eindeutig referenzierbaren Q-IDs modelliert (vgl. (1) in Abb. 1). Mithilfe von sogenannten Statements werden sie durch Properties (2) und zugehörige Objekt-Items (vgl. (1)-(5)) semantisch angereichert.[3] Sie erfassen u. a. semantische Zugehörigkeiten (z. B. zu bestimmten Pandemien oder thematischen Feldern wie „Emotion“ oder „Maßnahme“, vgl. (1) und (3)), morphologische Merkmale (2), sowie weitere Kategorien wie Definitionen (4) und Belege aus dem Projektkorpus (6). Die Objektposition kann dabei durch kontrollierte Vokabulare, Freitexte oder URIs besetzt sein. Zur linguistischen Annotation nutzt das System etablierte Ontologien wie LexInfo (Cimiano et al. 2011) und OliA (Chiarcos 2012); eine normdatenbasierte Verlinkung erfolgt über externe Referenzquellen wie Wikidata (Vrandečić/Pintscher/Krötzsch 2023) oder das grammatische Informationssystem *grammis* (Schneider/Lang 2022). Diese technische Infrastruktur gewährleistet nicht nur semantische Konsistenz und Inter-

3 Ein Beispiel für ein Statement aus der Abbildung 1 ist: Choleraausbruch (Q1) [Subjekt] – Semantische Klasse [Prädikat] – „Pandemieverlauf“ [Objekt].

operabilität, sondern ermöglicht auch die Anbindung des Pandemiewortschatzes an das Semantic Web wie auch andere Ressourcen. So besteht eine Kooperation mit dem IDS, in deren Rahmen der historische Pandemiewortschatz mit dem Wortschatz der Coronapandemie verknüpft wird. Die Vernetzung erfolgt über ein gemeinsam genutztes kontrolliertes Vokabular zur semantischen Kategorisierung des Pandemiewortschatzes. Ein wesentliches Merkmal der PandeLexBase ist ihre querybasierte Erschließbarkeit: Über SPARQL-Abfragen lassen sich gezielt Informationen abrufen – etwa zu Wortbildungsmustern wie „X-maßnahme“, semantischen Kategorien, Synonymgruppen oder genrekontextualisierten Vorkommen. Dadurch eröffnet sich ein breites Anwendungsspektrum für verschiedene Forschungsperspektiven, z. B. für die historische Sprachforschung, Diskursanalysen oder Studien zur Semantisierung von Krankheitsereignissen.

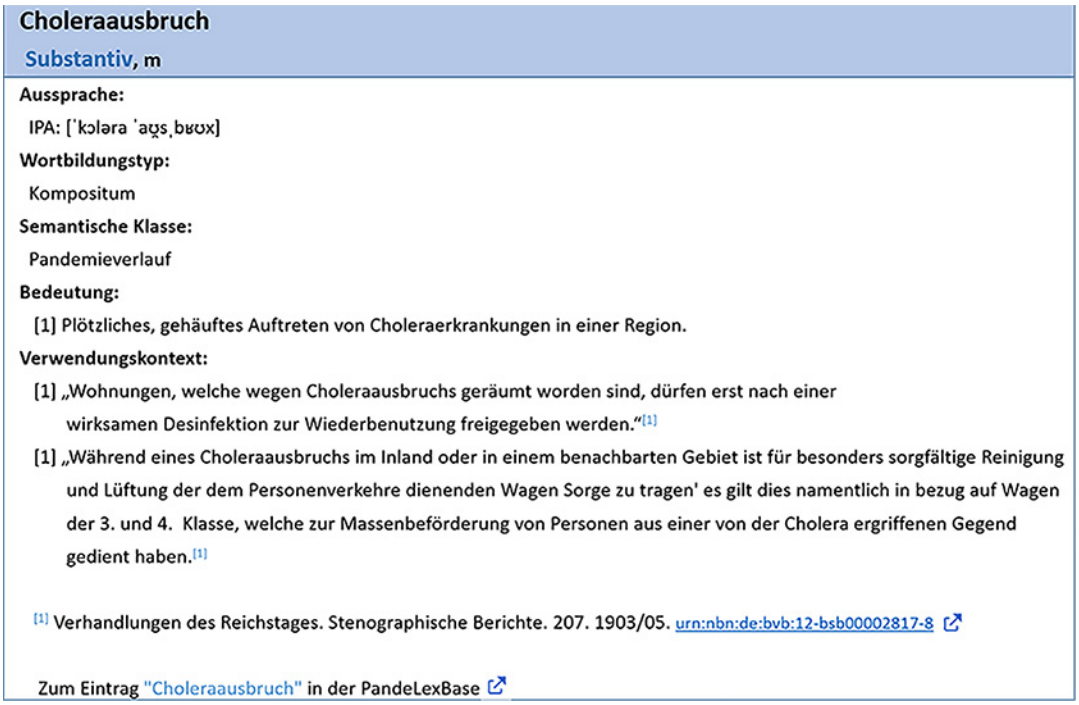

Abb. 2: Das Stichwort *Choleraausbruch* im *Pandemictionary*

Daneben bildet das *Pandemictionary* in Abbildung 2 die zweite zentrale Instanz und dient als leicht zugängliches Online-Wörterbuch. Im Unterschied zur datenbankorientierten PandeLexBase legt es den Fokus auf eine leserfreundliche Darstellung lexikografischer Artikel. Der Aufbau folgt dem Wiktionary-Prinzip: Einträge sind klar gegliedert und umfassen Informationen zu Lemma, Wortart, Genus, Aussprache, Wortbildungsstruktur, semantischer Klassifikation, Bedeutung sowie Belegbeispiele. Farblich und typografisch differenzierte Darstellungselemente unterstützen die schnelle Erfassung der Inhalte. Verlinkungen zu *grammis* erleichtern zudem die grammatische Einordnung. Die Nutzung des *Pandemictionary*s ist sowohl über eine alphabetische Navigation als auch über eine Stichwort-Suchfunktion möglich. Zusätzlich werden thematische Listen von Stichwörtern bereitgestellt, die spezifische komplexe Abfrageszenarien vorwegnehmen, etwa nach spezifischen grammatischen Merkmalen, produktiven Wortbildungsmustern oder semantischen Clustern. Dadurch eignet sich das *Pandemictionary* besonders als Einstiegspunkt für Nutzer*innen mit unterschiedlichem Vorwissen.

Ein entscheidendes Merkmal ist die Verknüpfung beider Ressourcen: Die Wörterbucheinträge im *Pandemictionary* enthalten Belegzitate aus dem Korpus und verweisen direkt auf die weiterführenden Informationen in der PandeLexBase. Dies ermöglicht einen nahtlosen

Wechsel zwischen lexikografischer Übersicht und detaillierter Datenanalyse. Ein exemplarischer Fall ist der Eintrag *Choleraausbruch*: Während im *Pandemictionary* eine benutzerfreundliche Darstellung erfolgt, bietet die PandeLexBase zusätzliche strukturierte Informationen zu Morphologie, Belegen und semantischen Relationen. Diese wechselseitige Anbindung ermöglicht eine Kombination aus wissenschaftlicher Tiefe und zugänglicher Wissensvermittlung.

3. Daten und Methode

Die systematische Erschließung historischen Wortschatzes im Kontext vergangener Pandemien ist mit besonderen methodischen Herausforderungen verbunden: Viele Quellen sind nur fragmentarisch überliefert, schwer zugänglich oder spiegeln sprachhistorische Entwicklungen wider, die eine eindeutige lexikografische Zuordnung erschweren. Um dennoch ein belastbares Glossar pandemiebezogener Lexeme zur Cholerapandemie zu entwickeln, wurde ein mehrstufiges Vorgehen zum Aufbau eines projektspezifischen Korpus gewählt, das eine vielseitige und sorgfältig kuratierte Datenbasis sicherstellt.

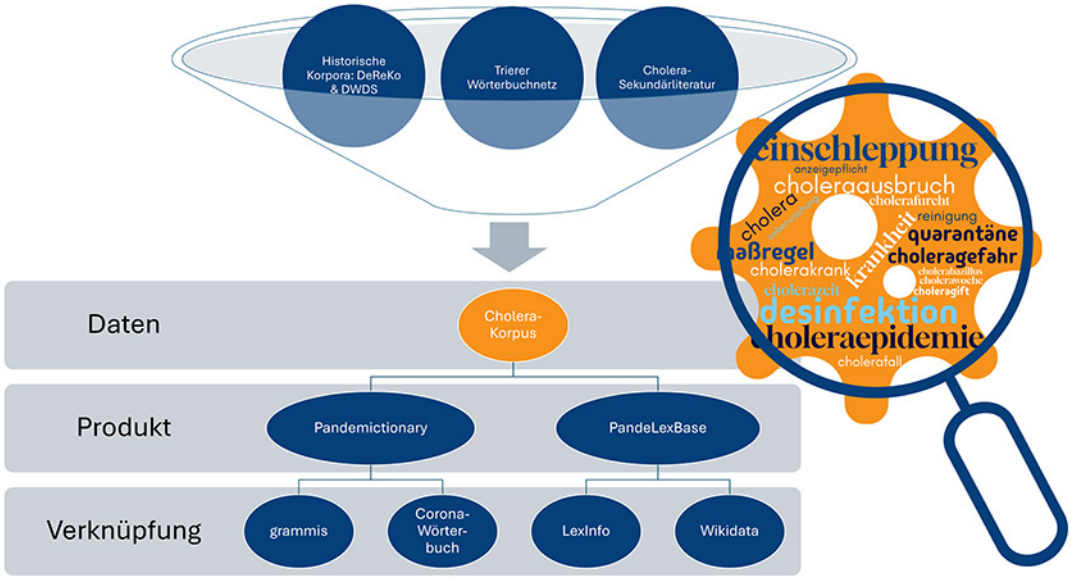

Abb. 3: Aufbau des Trierer Projektes zu historischem Pandemiewortschatz

Im ersten Schritt wurden historische Wörterbücher aus dem *Trierer Wörterbuchnetz*[4] herangezogen, um bereits dokumentierten Wortschatz zu extrahieren (siehe Abb. 3). Ergänzt wurde dieses Material durch einschlägige Sekundärliteratur zur Cholera – insbesondere aus medizin- und kulturhistorischer Perspektive (z. B. Dettke 1995; Aselmeyer 2015). Die in diesen Werken zitierten historischen Quellen – darunter autobiografische Texte, Gedichte, Briefe und Berichte aus dem 19. und frühen 20. Jahrhundert – dienten als Grundlage für die Sammlung authentischer Sprachbelege. Der dritte und zugleich umfangreichste Baustein des Korpus basiert auf zwei großen digitalen historischen Referenzkorpora: dem DWDS-Korpus (2024, ca. 1,53 Milliarden Tokens) sowie dem HIST-gesamt-Korpus des Leibniz-Instituts für Deutsche Sprache (IDS; ca. 70 Millionen Tokens). Die beiden Korpora wurden auf Textstellen mit Cholerabegriffen hin durchsucht, darunter das Wort *Cholera* sowie Kompo-

4 Z. B. Deutsches Wörterbuch von Jacob und Wilhelm Grimm, Goethe-Wörterbuch oder Herders Conversations-Lexicon.

sita mit *Cholera* als Erst- oder Letztglied. Die Suchtreffer selbst sowie ihre Kontexte, welche potenziell weitere pandemiebezogene Lexik enthalten, wurden in das Cholera-Korpus integriert.

Die Auswertung des Cholera-Korpus erfolgte mit Hilfe der Textanalyseplattform Sketch Engine (Kilgariff et al. 2004), die eine quantitative und kontextbezogene Erhebung lexikalischer Einheiten erlaubt. Mithilfe dieser Software wurden Frequenzanalysen, Keyword-Erkennungen, Wortbildungsmuster (insbesondere Komposita) und Kollokationen identifiziert, um typische pandemiebezogene Lexeme systematisch zu erfassen. Die aus diesen Analysen gewonnenen rund 1.300 Stichwörter bilden die Grundlage für das entstehende Cholera-Glossar und dienen gleichzeitig als Grundeinheiten für die lexikografische Erfassung in den beiden Wiki-basierten Ressourcen des Projekts: PandeLexBase und *Pandemictionary*.

4. Analyse: Cholera-Wortschatz, lexikalische Besonderheiten und kulturelle Reaktionen im Spiegel der Cholera

4.1 Korpuslinguistische Befunde

Im Folgenden werden korpuslinguistische Befunde der Analyse des Cholera-Korpus vorgestellt, die mithilfe von Sketch Engine (Kilgarriff et al. 2004) durchgeführt wurde und die insbesondere Frequenzverteilungen, Kookkurrenzen sowie die Ermittlung produktiver Cholera-Komposita fokussiert. Ziel war es, auf Grundlage dieser quantitativen Daten zentrale Lexeme und Wortverbindungen zu identifizieren, die sich als relevant für die lexikografische Aufnahme in das Cholera-Wörterbuch erweisen. Im Anschluss an die Korpusanalyse wird exemplarisch eine Auswahl besonders frequenter Substantive präsentiert. Die dokumentierten Häufigkeitsverteilungen geben dabei nicht nur Aufschluss über die lexikalische Dichte einzelner Nomen, sondern illustrieren zugleich diachrone Verschiebungen in der Verwendung pandemierelevanter Wortschatzbereiche.

Nomen			N-Gramme			Komposita		
Item	Absolute Häufigkeit	Relative Häufigkeit	Item	Absolute Häufigkeit	Relative Häufigkeit	Item	Absolute Häufigkeit	Relative Häufigkeit
cholera	8093	1.027.152.985	asiatische Cholera	230	29.191.299	cholerafall	434	55.082.713
krankheit	1902	241.399.355	ansteckende Krankheit	263	33.379.617	choleraepidemie	382	48.482.941
jahr	1464	185.808.967	Ausbruch der Cholera	135	17.134.024	cholerakranke	312	39.598.632
pest	1200	152.302.432	Verbreitung der Cholera	121	15.357.162	choleragefahr	203	25.764.495
tag	1173	148.875.627	beamteter Arzt	118	14.976.405	cholerazeit	173	21.956.934
zeit	1126	142.910.449	choleraverdächtige Erkrankung	115	14.595.650	choleraerkrankung	88	11.168.845
schiff	1088	138.087.538	gelbes Fieber	116	14.722.568	choleraleichen	73	9.265.065
fall	976	123.872.645	Auftreten der Cholera	112	14.214.894	cholerakranken	66	8.376.634
person	974	123.618.807	Denkschrift über die Choleraepidemie	93	11.803.439	choleraentleerung	63	7.995.878
arzt	836	106.104.028	Choleraepidemie 1892	93	11.803.439	cholerakommission	58	7.361.284
stadt	817	103.692.572	Bekämpfung der Cholera	82	10.407.333	cholera-epidemie	57	7.234.366
herr	742	94.173.670	gemeingefährliche Krankheit	78	9.899.658	cholerabacillen	57	7.234.366
seuche	739	93.792.914	Jahr 1892	95	12.057.276	choleraort	56	7.107.447
ort	714	90.619.947	Pest- oder Cholerafall	68	8.630.471	choleraherd	46	5.838.260
hafen	648	82.243.313	Fall von Cholera	66	8.376.633	cholerabazillus	44	5.584.423
land	619	78.562.671	freier Verkehr	78	9.899.658	cholerabazille	42	5.330.585
kranke	616	78.181.915	deutscher Hafen	67	8.503.552	cholerajahr	33	4.188.317
mensch	610	77.420.403	traurige Verheerung	61	7.742.040	choleraeinschleppung	27	3.426.805
wasser	599	76.024.297	ganze Furchtbarkeit	61	7.742.040	choleragift	26	3.299.886
erkrankung	585	74.247.436	bakteriologische Untersuchung	63	7.995.878	choleraverdacht	26	3.299.886
hamburg	534	67.774.582	verseuchter Hafen	59	7.488.203	choleraanfall	25	3.172.967
typhus	489	62.063.241	schmutzige Wäsche	66	8.376.633	cholerawäsche	24	3.046.049
regierung	466	59.144.111	Förster in der nächsten Stadt	56	7.107.447	cholerakranke|kranker	24	3.046.049
desinfektion	461	58.509.518	entferntes Dorf	58	7.361.284	cholerakranker	22	2.792.211
maßregel	460	58.382.599	Kunde von den traurigen Verheerungen	54	6.853.609	cholerawoche	21	2.665.293

Abb. 4: häufige Nomen, N-Gramme und Cholera-Komposita

Die ermittelten Substantive (siehe Abb. 4) lassen sich verschiedenen Wortschatzbereichen zuordnen.[5] So finden sich etwa Nomen, die Cholera mit anderen Krankheiten vergleichen, wie *Typhus*, oder solche, die eine negative Konnotation transportieren, etwa die Bezeichnung *Seuche*. Dies verweist auf die vielfältigen Herausforderungen während der Cholera-Pandemie (vgl. Kabatnik/Klee 2024, S. 310). Nomen wie *Blut*, *Körper* und *Leiche* stellen dabei einen Bezug zum menschlichen Körper und seinen Flüssigkeiten her und verdeutlichen physische Aspekte der Erkrankung. Spezifisch für die Cholera – mit ihrer Übertragung über verunreinigtes Wasser – sind zudem Nomen wie (*Trink-*)*Wasser* und verschiedene Wasserkomposita. Sie verweisen auf Maßnahmen zur Aufbereitung und Überwachung der Trinkwasserversorgung. Auch Begriffe wie *Desinfektion*, *Quarantäne* und *Mittel* zeigen Praktiken der Prävention, Bekämpfung und Kontrolle der Krankheit auf. Diese betreffen nicht nur medizinische, sondern auch administrative Bereiche: Häufige Nomen wie *Vorschrift*, *Anordnung*, *Maßregel*, *Gesetz*, *Regierung* und *Kommission* verdeutlichen die rechtlich-politischen Steuerungsversuche des Gesundheitsschutzes. Auffällig, wenn auch wenig überraschend, ist das häufige Auftreten von Wörtern wie *Arzt* oder (*Cholera-*)*Kranke*. Bemerkenswert ist außerdem die hohe Frequenz von Nomen wie *Reisende* oder *Auswanderer*, die den mobileren Teil der Bevölkerung betreffen und Rückschlüsse auf die Krankheitsverbreitung ermöglichen. Schließlich weisen nominale Begriffe wie *Furcht*, *Angst* oder *Sorge* auf emotionale Reaktionen der Bevölkerung hin und dokumentieren die psychische Belastung im Verlauf der Pandemie (vgl. Kabatnik/Klee 2024, S. 312).

In der Analyse der N-Gramme (siehe Abb. 4) treten häufige Verbindungen wie *Ausbruch der Cholera*, *Verbreitung der Cholera*, *Auftreten der Cholera*, *Einschleppung der Cholera* oder *Cholerafall an Bord* auf. Diese Konstruktionen betonen die Ausbreitung und transregionale Verbreitung der Krankheit. Besonders der Ausdruck *an Bord* verweist auf die Verbreitung über Seewege, etwa durch Handel oder Migration. Auffällig ist auch die Verbindung *asiatische Cholera*, die einerseits Informationen über Herkunft und Verbreitungsweg der Krankheit nahelegt, andererseits aber auch auf problematische Diskursmuster verweist. Analog zu jüngeren Wortbildungen in der Corona-Pandemie wie *südafrikanisches Virus* oder *Wuhan-Virus* zeigt sich hier eine Tendenz zur Stigmatisierung (Klosa-Kückelhaus 2022; Nossem 2023, S. 113–116): Die geografische Markierung von Krankheiten kann zur Schuldzuweisung beitragen und fungiert als „diskursive Abgrenzungsstrategie“ (Nossem 2023, S. 116), indem sie die Krankheit als Bedrohung von außen konstruiert.

Cholera-Komposita (siehe Abb. 4) treten in unterschiedlichen Wortarten auf: als Nomen (*Cholerafall*), Adjektive (*choleraverdächtig*) oder Verben (*choleraverpesten*) (vgl. Kabatnik/Klee 2024, S. 315). Sie decken unterschiedliche Wortschatzbereiche ab – von Symptomen und Präventionsmaßnahmen bis hin zu Gesundheitseinrichtungen. Substantive wie *Cholerahospital*, *Choleraspital*, *Cholerakrankenhaus*, *Choleralazarett* oder *Cholerabaracke* verdeutlichen, wie Pflege und Isolation institutionell organisiert wurden. Während *-hospital*, *-spital* und *-krankenhaus* als Varianten voneinander gewertet werden können, verweisen *-lazarett* und *-baracke* zusätzlich auf einen militärischen Kontext: Das Militär unterstützte vielerorts bei der Eindämmung, Absperrung betroffener Gebiete und durch Bereitstellung medizinischer Infrastruktur (z. B. Feldbetten) (vgl. z. B. Stamm-Kuhlmann 1989). Darüber hinaus finden sich Komposita wie *Cholerabazillen*, *Cholerakeim*, *Cholerapilz* und *Choleravibrionen*, die auf mikrobielle Krankheitserreger verweisen. Diese Bezeichnungen stammen aus dem medizinischen Fachwortschatz und deuten auf die wissenschaftliche Auseinandersetzung mit der Cholera in medizinischen Texten hin.

5 Für einen detaillierteren Überblick über die Frequenzlisten zu häufigen Nomen, Adjektiven und Verben verweisen wir auf die Ergebnisse der Korpusstudie zum Cholera-Wortschatz in Kabatnik/Klee (2024).

Insgesamt zeigen die Korpusanalysen klar abgegrenzte Wortschatzbereiche mit erhöhter Frequenz, was auf einen möglichen sprachlichen Wandel im Verlauf der Cholerapandemie hinweist. Wenig überraschend ist dabei die hohe Repräsentation der Bereiche Verwaltung und Medizin sowie von Ortsbezeichnungen im Kontext des Pandemiegeschehens. Cholera-spezifisch hingegen ist der lexikalische Fokus auf die Themen Trinkwasserversorgung und -aufbereitung – ein zentrales Element in der Bekämpfung der Krankheit. Für die Aufnahme in das digitale Wörterbuch *Pandemictionary* sind daher der spezifische Pandemiebezug und die inhaltliche Relevanz ausschlaggebend. Allgemeine, auch außerhalb des Cholera-Kontexts häufige Lexeme wie *Arzt* oder *krank* werden hingegen nicht berücksichtigt.

4.2 Lexikalische Besonderheiten während der Cholera-Pandemie am Beispiel von Cholera-Komposita

Cholera-Komposita konnten in der Korpusuntersuchung als produktives Wortbildungsmuster (n=328)[6] identifiziert werden, d. h. Komposita mit dem Erst- oder Zweitglied *Cholera* sind – ähnlich wie bei *Corona-* oder Pest-*Komposita* (Fuchs 2021; Moulin 2020) – häufig und produktiv. Im Rahmen der Cholera-Pandemie lassen sich zahlreiche derartige Zusammensetzungen nachweisen, bei denen *Cholera* als determinierendes oder determiniertes Glied fungiert. Wie bereits oben ausgeführt, treten sie in unterschiedlichen Wortarten auf und gehören überwiegend dem Bereich der Neubildungen an (Donalies 2011). Dabei wird vorhandenes sprachliches Material neu kombiniert, um pandemiespezifische Konzepte zu benennen. Beispiele wie *Cholerakranke*, *choleraverdächtig*, *Cholerazeit*, *Choleraferien* oder *choleraverseucht* belegen diese lexikalische Dynamik. Auffällig ist dabei, dass viele dieser Komposita direkte Entsprechungen in der Corona-Pandemie gefunden haben – etwa *Coronakranker*, *coronaverdächtig*, *Coronazeit*, *Coronaferien* oder *coronaverseucht* (vgl. Möhrs 2021, S. 36). Die meisten dieser Cholerakomposita weisen eine relativ eindeutige Bedeutung auf und lassen sich klar semantisch erschließen. Weitere Cholera-Komposita, deren Bedeutung aus heutiger Sicht deutlich schwerer zu erschließen ist, sind beispielsweise *Cholerablut*, *Hühner-/Geflügel-/Schweine-/Ferkel-Cholera*, *Choleratropfen* oder *Choleraklette*, auf die wir im Folgenden detailliert eingehen möchten.

Das Lexem *Cholerablut* (n=17) eröffnet verschiedene Interpretationsmöglichkeiten, die sich auf Basis der Zusammensetzung sowie des lexikologischen Hintergrundwissens erschließen lassen (vgl. Behrens 1998, S. 1). So könnte *Cholerablut* etwa:

a) im Sinne von „durch Cholera ausgeschiedenes Blut“ verstanden werden, etwa Blut, das beim Erbrechen auftritt („Blut durch/von der Cholera“),

b) analog zu *Halbblut* (DWDS 2025: Halbblut)[7] eine metaphorische Abstammung oder „Verwandtschaft“ zu an Cholera Erkrankten anzeigen,

c) wie *Spenderblut* (DWDS 2025: Spenderblut)[8] ein therapeutisch verwendetes Blutprodukt bezeichnen, das zur Behandlung der Cholera eingesetzt wurde,

6 Für die Berechnung der Produktivität wurden nach Baayen (2009) die folgenden Summen und Formeln herangezogen: V = 328 (unterschiedliche *Cholera-*Komposita), N = 3916 (alle Tokens dieser Komposita im Korpus), V_1 = 106 (Hapax Legomena – Komposita, die nur einmal im Cholera-Korpus vorkommen): 1. Realisierte Produktivität (P_r): V/N = 328/3916 = 0,0837 ⟶ 8,37% der Token entfallen auf jeweils unterschiedliche Typen; 2. Potenzielle Produktivität (P): V1/N = 106/3916 = 0,0271 ⟶ 2,71% der Tokens sind einmalige Bildungen.

7 www.dwds.de/wb/Halbblut (Stand: 9.6.2025)

8 www.dwds.de/wb/Spenderblut (Stand: 9.6.2025)

d) oder – vergleichbar mit *Artistenblut* (DWDS 2025: Artistenblut)[9] – das „Blut von Cholerakranken“ als mit Cholera infiziertes Blut meinen.

Der folgende Korpusbeleg gibt Aufschluss über die korrekte Interpretation im Kontext:

(1) In dem Blute von Gesunden verhalt sich die Menge des Blutwassers zur Menge des Blutkuchens im Durchschnitte, wie 55 zu 45; in dem Blute der Cholerakranken hingegen verhält sich ersteres zu lezterem wie 33,2 zu 66,8; so daß das **Cholerablut** zwei Mal so viel Blutkuchen enthält, als das Blut von Gesunden. (Miszellen. In: Dr. Johann Gottfried Dingler: Polytechnisches Journal. Bd. 44. Stuttgart, 1832. S. 461; Hervorhebung d. V.)

Der Korpusbeleg (1) entstammt einem medizinischen Fachtext aus dem Jahr 1832 und thematisiert den Vergleich der Blutzusammensetzung zwischen Cholerakranken und gesunden Personen. Im Fokus steht dabei insbesondere der sogenannte „Blutkuchen“, dessen Beschaffenheit bei Erkrankten signifikante Unterschiede aufweist. Die Bedeutung des Lexems *Cholerablut* lässt sich in diesem Zusammenhang eindeutig erschließen: Durch die explizite Formulierung „im Blute der Cholerakranken“ erfolgt eine klare Zuschreibung, die das Blut als jenes von an Cholera erkrankten Individuen kennzeichnet. Die semantische Interpretation ergibt sich demnach durch die kontrastive Darstellung sowie die explizite Bezugnahme im Fachdiskurs.

Im Cholera-Korpus fällt eine Gruppe von Determinativkomposita auf, bei denen Cholera als Zweitglied mit Tierbezeichnungen wie *Hühner-*, *Geflügel-*, *Schweine-* oder *Ferkel-* als Erstglied kombiniert wird – etwa in *Hühnercholera* (n=30), *Geflügelcholera* (n=143), *Schweinecholera* (n=16) oder *Ferkelcholera* (n=4). Auch diese Komposita lassen unterschiedliche Interpretationen zu:

a) Sie könnten – analog zu *Hühnerpest* (DWDS 2025: Hühnerpest)[10] – die Bezeichnung einer tierischen Infektionskrankheit darstellen, die spezifisch bei diesen Tieren auftritt.

b) Möglich ist auch eine semantische Nähe zur *Vogelgrippe* (DWDS 2025: Vogelgrippe),[11] also eine Übertragung der Krankheitserreger durch Kontakt mit den genannten Tieren.

c) Weiter denkbar ist ebenfalls ein Bezug zum Konsum von Tierprodukten – etwa eine Infektion durch den Verzehr kontaminierten Fleisches, ähnlich der Übertragung bei Campylobacter-Infektionen.

d) Schließlich könnte auch eine metaphorisch-abwertende Verwendung vorliegen, bei der das Erstglied zur sprachlichen Dramatisierung oder Stigmatisierung dient – im Sinne einer beschimpfenden oder emotional aufgeladenen Benennung wie „elende Schweinegrippe“ (analog siehe Beschimpfungen von Corona oder Krebs in Havryliv 2023, S. 38).

Die Auflösung über die Bedeutungsinterpretation erfolgt durch den folgenden Korpusbeleg:

9 www.dwds.de/wb/Artistenblut (Stand: 9.6.2025)

10 www.dwds.de/wb/H%C3%BChnerpest (Stand: 9.6.2025)

11 www.dwds.de/wb/Vogelgrippe (Stand: 9.6.2025)

(2) Die **Geflügelcholera** befällt namentlich Hühner, Enten und Gänse und endigt in der Regel mit dem Tode. Die Ansteckung eines Geflügelbestandes macht sich zuerst durch ganz plötzlich auftretende Todesfälle bemerkbar; die Tiere sterben oft, ohne daß sie auffällig krank waren. ([N. N.]: Lesebuch für Fortbildungsschulen. Lahr, 1901; Hervorhebung d. V.)

Der Korpusbeleg (2) aus einer Gebrauchsliteraturquelle aus dem Jahr 1901 illustriert die Verwendung des Lexems *Geflügelcholera*. Die Ansteckung von Tieren wird dabei explizit durch das transitive Verb *befallen* (vgl. Schierholz/Uzonyi (Hg.) 2022, S. 738) angezeigt, gefolgt von einer Aufzählung spezifischer Geflügelarten, die als potenziell erkrankte und infolgedessen auch tödlich betroffene Tiergruppen benannt werden. Die semantische Interpretation des Lexems als eine Tierkrankheit wird dadurch eindeutig ermöglicht.[12]

Weiter eröffnet auch das Lexem *Choleratropfen* (n=12) mehrere mögliche Bedeutungen, die je nach Kontext unterschiedlich gedeutet werden können:

a) Im medizinisch-epidemiologischen Sinne könnte es sich um einen Hinweis auf den Übertragungsweg durch Tröpfchen handeln – im Sinne der *Tröpfcheninfektion* (vgl. DWDS 2025: Tröpfcheninfektion).[13]

b) Eine körperlich-symptomatische Deutung wäre der Bezug auf Schweißtropfen infolge von Fieber oder der Anstrengung beim Erbrechen.

c) Möglich wäre auch eine metaphorische oder volkstümliche Verwendung im Sinne eines alkoholischen Kräuterbitters – also Tropfen als Synonym für *Weinbrand* (vgl. DWDS 2025: Tropfen).[14]

d) Schließlich könnte *Choleratropfen* auch ein medizinisches Präparat zur Behandlung der Cholera bezeichnen, vergleichbar mit *Hustentropfen* (vgl. DWDS 2025: Hustentropfen).[15]

Der folgende Beleg aus dem Cholera-Korpus unterstützt bei der Disambiguierung:

(3) Deine Nahrung sei einstweilen eine schleimige Suppe, auch Zwieback oder altbackenes Weißbrod ohne Butter. Hast Du bewährte (nach ärztlicher Vorschrift verfertigte) **Choleratropfen** vorräthig, so nimm davon 20 bis 30 Tropfen auf Zucker! (Verhandlungen des Reichstages. Berlin, 1893; Hervorhebung d. V.)

Der Korpusbeleg (3) entstammt einer Denkschrift zur Cholerapandemie aus dem Jahr 1893, die als Protokoll parlamentarischer Verhandlungen im Reichstag fungierte und konkrete Empfehlungen für den Bevölkerungsschutz im Kontext der Epidemie formulierte. Thematisch im Fokus stehen präventive und therapeutische Maßnahmen bei einer Erkrankung an Cholera. Das Lexem *Choleratropfen* tritt in einer komplexen attributiven Konstruktion auf, in der den Tropfen zum einen eine bewährte Wirksamkeit zugeschrieben, zum anderen aber auch ihre Herstellung „nach ärztlicher Vorschrift“ betont wird. Die semantische Auswertung dieses Kontexts legt nahe, dass es sich bei Choleratropfen um ein ärztlich verordnetes Heilmittel handelt, dessen Zusammensetzung normiert war. Diese Interpretation

12 In analoger Weise lassen sich auch die Belege zu *Schweinecholera* bzw. *Ferkelcholera* deuten, die zwar nicht im Einzelnen präsentiert werden, jedoch dieselbe semantische Struktur und Interpretation nahelegen.

13 www.dwds.de/wb/Tr%C3%B6pfcheninfektion (Stand: 9.6.2025)

14 www.dwds.de/wb/Tropfen (Stand: 9.6.2025)

15 www.dwds.de/wb/Hustentropfen (Stand: 9.6.2025)

wird durch nachfolgende Hinweise auf Dosierung und Einnahmeverhalten zusätzlich gestützt.

Auch das Lexem *Choleraklette* (n=1) lässt sich – völlig losgelöst vom Kontext – auf unterschiedliche Weise interpretieren und verweist auf die semantische Mehrdeutigkeit historischer Komposita. Mögliche Deutungen umfassen:

a) Eine tierische oder parasitäre Assoziation – etwa ein insektenähnliches Wesen, vergleichbar einem Blutegel, das im Rahmen volkstümlicher oder pseudomedizinischer Praktiken gegen Cholera eingesetzt wurde.

b) Eine an der Kleidung angebrachte Kennzeichnung von Cholerakranken mittels eines Klettverschlusses – analog zur Bedeutung von *Klettverschluss* als technisches Befestigungsmittel (vgl. DWDS 2025: Klettverschluss).[16]

c) Eine medizinische oder phytotherapeutische Deutung: *Choleraklette* könnte eine Heilpflanze bezeichnen, etwa die im DWDS beschriebene *Klette* – „eine zu den Korbblütlern gehörende, an Wegrändern wachsende krautige Pflanze mit violetten Blüten und stacheligen, mit Widerhaken versehenen Hüllblättern“ (DWDS 2025: Klette).[17]

d) Schließlich ist auch eine metaphorische Bedeutung denkbar: die *Choleraklette* als Ausdruck für eine durch Cholera ausgelöste extreme Anhänglichkeit, im Sinne der umgangssprachlichen Verwendung von *Klette* als „anhängliche Person“ (vgl. DWDS 2025: Klette).[18]

Welche der skizzierten Bedeutungen dem Lexem im konkreten Fall zugeschrieben werden kann, erschließt sich aus der Analyse des folgenden Korpusbelegs:

(4) Merkwürdige, geheimnißvolle Gewächse sind darunter: die Steppennadel, deren spitzer, mit Widerhaken gezähnter Samen den Thieren durch Haut und Muskeln dringt in ihr Herz und Eingeweide, so daß sie elendiglich erliegen müssen; das „betrunkene Kraut“, dessen Genuß die Pferde toll macht und lähmt, während es den Rindern nicht schadet; die **Choleraklette**, welche zum ersten Male mit der Seuche erschienen und ein specifisches Heilmittel gegen diese und die Hundswuth sein soll, und der Kurai. (Stolle, Ferdinand (Hg.); Diezmann, August (Hg.): Die Gartenlaube. Jg. 7 (1859); Hervorhebung d. V.)

Der Korpusbeleg (4) stammt aus dem Jahr 1859 und erschien in der Zeitschrift „Die Gartenlaube“. Im einleitenden Abschnitt wird durch „Merkwürdige, geheimnißvolle Gewächse“ ein thematischer Rahmen geschaffen, der die Vorstellung neuartiger Pflanzenarten sowie deren Einfluss auf die botanische Vielfalt vorbereitet. Nach der Beschreibung der sogenannten Steppnadel und des „betrunkenen Krauts“ folgt die Einführung der sogenannten Choleraklette. Deren Auftreten wird explizit mit dem zeitgleichen Aufkommen der Choleraepidemie verknüpft. Die semantische Zuschreibung der Choleraklette erfolgt dabei eindeutig über einen Relativsatz, in dem sie als „ein specifisches Heilmittel gegen diese [Cholera]“ bezeichnet wird. Damit lässt sich die Bedeutung der Choleraklette als medizinisches Heilmittel zur Behandlung der Cholera eindeutig erschließen.

16 www.dwds.de/wb/Klettverschluss (Stand: 9.6.2025)

17 www.dwds.de/wb/Klette (Stand: 9.6.2025)

18 www.dwds.de/wb/Klette (Stand: 9.6.2025)

Insgesamt zeigen die präsentierten Beispiele eine interpretative Offenheit historischer Komposita, d. h. insbesondere wie stark sich historische Lexeme mit kulturellen, medizinischen und metaphorischen Deutungsmustern überlagern könn(t)en – v. a. aber auch dann, wenn der ursprüngliche Verwendungskontext fehlt oder nur schwer rekonstruierbar ist, was die Notwendigkeit einer kontextbasierten Bedeutungserschließung aufzeigt und die Relevanz der Bereitstellung von Korpusbelegen in *Pandemictionary* und PandeLexBase hervorhebt.

4.3 Lexikalische Besonderheiten während der Cholera-Pandemie am Beispiel von Metaphern

Abschließend soll auf die deutlich ausgeprägte Metaphorik von Pandemiewortschatz eingegangen werden, die nicht nur in der Corona-Pandemie auffällig war – etwa in der häufigen metaphorischen Rahmung der Bekämpfung der Erkrankung (vgl. z. B. Klosa-Kückelhaus 2022) –, sondern auch durch die Annotation der Stichwortliste im Cholera-Korpus identifiziert werden konnte (vgl. Kabatnik/Klee 2025). Metaphern operieren über eine Bedeutungsübertragung vom Ausgangs- auf einen Zielbereich (vgl. Sanders 2007, S. 119; Ullmann 1970, S. 213) und eröffnen damit weitere Perspektiven auf die (auch subjektive) Wahrnehmung, Verarbeitung und Rezeption von Pandemien innerhalb der Bevölkerung. Sie fungieren als sprachliche Mittel, über die kollektive Haltungen, Ängste und Deutungsmuster artikuliert werden können. Ein markantes Beispiel aus der Corona-Pandemie ist die metaphorische Bezeichnung einer Virusmutation als *Höllenhundvariante*[19], die durch das Aufrufen mythologischer Konzepte – konkret des Höllenhundes der griechischen Mythologie – die besondere Gefährlichkeit der Virusvariante betont. Die metaphorische Bildwelt verweist dabei auf eine biblisch-mythologische Semantik der Verdammnis, in der die Hölle als Gegenpol zum Himmel für ein unentrinnbares, zerstörerisches Verderben steht (vgl. Pommeranz 2015, S. 13; siehe dazu auch Nünning (Hg.) 2024, S. 175–177). Auch in den historischen Belegen zur Cholera lassen sich zahlreiche metaphorische Muster identifizieren. So begegnet die Cholera als *Gespenst*, *Würgeengel*, *unheimlicher Gast* oder *schwarzer Tod*. Diese Metaphern greifen auf religiöse, mystische und farbsymbolische Konzepte zurück und konstruieren die Krankheit als unheilvolle, schwer fassbare und kaum kontrollierbare Macht. In ihnen spiegeln sich kollektive Affekte wie Furcht, Ohnmacht und Ausgeliefertsein angesichts der tödlichen Bedrohung. Die metaphorische Rahmung der Cholera fungiert somit als sprachliches Symptom gesellschaftlicher Krisenverarbeitung und trägt zur diskursiven Konstruktion des Pandemischen bei (vgl. zu Corona Schätz/Kirchhoff 2024).

Die Analyse pandemiebezogener Sekundärliteratur zur Cholera offenbart eine weitere markante semantische Auffälligkeit, nämlich die wiederholte Personifikation der Krankheit als weibliche Figur. Aselmeyer (2015, S. 92) dokumentiert zahlreiche Beispiele dieser sprachlichen Inszenierung, etwa in Formulierungen wie *Frau Cholera*, *Dame aus Asien*, *die unbekannte Feindin*, *eine leibhaftige Hexe* oder *Choleraweib*. Diese Personifizierungen verleihen der pandemischen Bedrohung ein konkretes, personifiziertes Profil, das kulturelle Vorstellungen von Krankheit, Fremdheit und Gefahr spiegelt.

Auffällig ist, dass sich diese Form der Personifikation in den historischen Korpora, die dem Projekt zugrunde liegen, nur vereinzelt nachweisen lässt – konkret konnten lediglich zwei Belege identifiziert werden, vgl. den folgenden Korpusbeleg:

19 Vgl. in Kieler Nachrichten „Corona-Mutation Cerberus: Das sagen Experten aus SH zur „Höllenhund-Variante““ (www.kn-online.de/schleswig-holstein/corona-variante-cerberus-in-sh-das-sagen-experten-zur-hoellenhund-variante-bq-1-1-Y2RH4VHIYII744GIKHDI6QY6VA.html, Stand: 14.7.2025)

(5) Aber ein solcher Winter, der ohne Rücksicht von Anstand den Herbst zur Treppe herabwirft und unangemeldet in jedermanns Stube hereinstürzt, und am Arme noch eine so widerliche Maitresse, wie **Frau Cholera** u'mherführt, ein solch ungeschlachter Winter, der sich weder durch Marie ‚Taglionis süßes Mienenspiel noch durch der Sennoras Pepas und DoloreS warme Gelenksprache erweichen läßt, ein Winter voll kosackischer Tücke und orientalischer Verwesung, ein Protector der Leichenbitter und Todtengräber, ein erklärter Feind aller weltlichen Lust, ein Kriegsschürer und diplomatischer Knotenschürzer; ein solch erbärmlicher Winter ist noch nicht dagewesen. (Die Grenzboten. Jg. 13, 1854, II. Semester. II. Band; Hervorhebung d. V.)

Beispiel (5) stammt aus einem Beitrag der Zeitschrift *Die Grenzboten*, der in satirischem Ton eine Phase gesellschaftlicher Erschütterung beschreibt. Weder kulturelle Ereignisse noch soziale Aktivitäten wie „Marie ‚Taglionis süßes Mienenspiel" oder „Sennoras Pepas und DoloreS warme Gelenksprache" vermochten die negative Grundstimmung in der Bevölkerung zu verbessern – im Gegenteil: Die Rede ist von Leichen, Totengräbern und der alles überschattenden Schuld der Cholera, die in drastischer Sprache als „eine so widerliche Maitresse, wie Frau Cholera" bezeichnet wird. Diese Gleichsetzung von Krankheit und Weiblichkeit bedient sich nicht nur metaphorischer Übertragung, sondern auch expliziter sexualisierter Zuschreibungen. Aselmeyer (2015, S. 92) weist darauf hin, dass hier möglicherweise eine Verwechslung zwischen grammatischem Genus und biologischem Geschlecht vorliege, was aus genderlinguistischer Sicht jedoch kaum haltbar ist. Stattdessen erscheint es plausibler, die Metaphorik im Kontext historischer Weiblichkeitskonstruktionen zu interpretieren, wie Aselmeyer (2015) überzeugend argumentiert, dass hinter solchen „Sprachfiguren" nicht nur „kulturell überlieferte Naturzuschreibungen" (ebd., S. 92), sondern auch ein politisch-patriarchales Programm sichtbar wird (vgl. ebd., S. 92). „Die Feminisierung der Seuche" fungiert dabei als Ausdruck einer „repressiven Rhetorik" (ebd., S. 92), die sich unter dem Deckmantel kultureller Narrative zur Legitimation und Stabilisierung patriarchaler Machtstrukturen bedienen lässt. Verstärkend wirkt hierbei die Weiblichkeitskonstruktion des 19. Jahrhunderts, in der die Frau die Attribute „krankheitsanfällig und krankheitsverbreitend" zugeschrieben werden (ebd., S. 92). Vor diesem Hintergrund wurde Weiblichkeit zur potenziellen Bedrohung des Gemeinwesens stilisiert – ein Diskurs, der mit dem Auftreten der Cholera politische Tragweite gewinnt. In dieser Logik wird die Frau nicht nur zur Metapher der Krankheit, sondern zur ideologischen Projektionsfläche für Krisendiskurse im Dienste der Aufrechterhaltung patriarchaler Ordnung.

5. Methodische Reflexion und Schlussfolgerung: Historische Korpora als Instrument zur Untersuchung von Sprachwandel am Beispiel pandemiebezogener Lexik

Die vorliegende Untersuchung demonstriert exemplarisch das Potenzial historischer Korpora für die Analyse lexikalischen Wandels im Kontext gesundheitshistorischer Krisen. Am Beispiel der Cholera-Pandemie konnte aufgezeigt werden, dass bestimmte Wortschatzbereiche – insbesondere solche, die unmittelbar mit dem Pandemiegeschehen in Verbindung stehen, wie etwa Orte, Maßnahmen der öffentlichen Gesundheitsfürsorge oder Aspekte der Trinkwasserversorgung – während der pandemischen Phase eine signifikante Häufung erfahren. Die beobachteten Frequenzanstiege lassen auf temporäre semantische Relevanz schließen, die sich nach dem Abklingen der akuten Bedrohungslage häufig wieder nivelliert. Ein besonderer Fokus lag auf der Analyse pandemiespezifischer Komposita,

wobei sich das Bildungsmuster *Erkrankung + Zweitglied* als produktive morphologische Konstruktion erwies. Dieses Muster lässt sich nicht nur für die Cholera, sondern ebenso für andere Pandemien – etwa die Pest oder COVID-19 – nachweisen, was auf übergreifende lexikalische Reaktionsmechanismen in Krisensituationen hinweist. Die Wiederkehr bestimmter lexikalischer Einheiten in Verbindung mit pandemischen Bezeichnungen legt nahe, dass pandemiebezogene Wortbildungsmuster einem zyklischen Prinzip folgen: Bestimmte Strukturen und Begriffe werden im Sprachgebrauch reaktiviert und semantisch neu kontextualisiert.

Methodisch hat sich der gezielte Zugriff auf historische Korpora mittels Suchanfragen zu pandemiespezifischen Lemmata (z. B. *Cholera*) als besonders ertragreich erwiesen. Die systematische Kontextanalyse liefert belastbare Ergebnisse zur Frequenzdynamik und semantischen Rahmung relevanter Lexeme und erlaubt Rückschlüsse auf gesellschaftlich bedeutsame Wortschatzbereiche. Damit wird nicht nur der Sprachwandel auf Makroebene sichtbar gemacht, sondern darüber hinaus ein Beitrag zur historisch orientierten Diskursanalyse geleistet. Die Korpusanalyse zeigt jedoch auch die Grenzen einer rein korpusbasierten Annäherung auf. Denn nur durch die Kombination historischer Korpora mit älteren Wörterbüchern und Sekundärliteratur zur Cholera lassen sich einige lexikalische Besonderheiten identifizieren, die in einer rein korpusbasierten Analyse möglicherweise unentdeckt geblieben wären. So zeigen etwa zahlreiche historiografische und literarische Darstellungen eine auffällige Tendenz zur weiblichen Personifizierung der Cholera – ein Phänomen, das in den durchsuchten Korpora lediglich in vereinzelter Form dokumentiert ist. Während solche metaphorischen Zuschreibungen in literarischen und essayistischen Texten durchaus präsent sind, erscheinen sie in den untersuchten historischen Korpora – die überwiegend aus administrativen, medizinischen und journalistischen Quellen bestehen – nur in Einzelfällen. Dies verweist auf eine methodisch bedeutsame Asymmetrie zwischen einzelnen historischen Quellen. Die Ergänzung um weitere Quellen ermöglicht so eine differenziertere Rekonstruktion von Bedeutungskonstitution und lexikalischer Konventionalisierung, die über rein frequenzbasierte Analysen hinausgeht. Die Verwendung von Sketch Engine zur Analyse von Frequenzen, Komposita und Kollokationen erlaubt nicht nur quantitative Auswertungen, sondern auch die Untersuchung von okkasionellen Wortneuschöpfungen, wie *Cholerablut*, *Choleratropfen* oder der *Hühnercholera*, die auf einen pandemiebedingten lexikalischen Wandel hinweisen.

Die aktuelle Anbindung der lexikografischen Ressourcen an Linked-Open-Data-Infrastrukturen ermöglicht eine semantische Vernetzung historischer Begriffe mit zeitgenössischen Datenbeständen – etwa aus Wikidata oder den IDS-Glossaren. Auf dieser Grundlage eröffnen sich neue komparative Analysepotenziale, insbesondere im Hinblick auf diachrone Vergleiche pandemiebezogener Lexik. Dies stellt einen innovativen Zugang zur Erforschung von Sprachwandel dar. Darüber hinaus erweitert die geplante Integration KI-gestützter Verfahren – etwa durch den Einsatz großer Sprachmodelle (LLMs) zur Belegauswahl oder semantischen Clusterbildung – das methodische Spektrum der Untersuchung um eine weitere Komponente. Sprachwandelprozesse können dann algorithmisch rekonstruiert und auf umfangreiche Datenbestände skaliert analysiert werden.

Literatur

Aselmeyer, Norman (2015): Cholera und Tod. Epidemieerfahrungen und Todesanschauungen in autobiografischen Texten von Arbeiterinnen und Arbeitern. In: Archiv für Sozialgeschichte 55, S. 77–106.

Baayen, R. Harald (2009): Corpus linguistics in morphology: Morphological productivity. In: Lüdeling, Anke/Kytö, Merja (Hg.) Corpus linguistics. An international handbook. Bd. 2. (= Handbücher

zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science 29.2). Berlin: De Gruyter, S. 899–919.

Balnat, Vincent (2020): Unter Beobachtung: Corona-Wortschatz im Deutschen und Französischen. In: Nouveaux Cahiers d'Allemand: Revue de linguistique et de didactique 38, 2, S. 1–22. https://hal.science/hal-02931171 (Stand: 9.6.2025).

Behrens, Leila (1998): Ambiguität und Alternation: Methodologie und Theoriebildung in der Lexikonforschung. Diss. München: Ludwig Maximilians Universität.

Chiarcos, Christian (2012): Ontologies of Linguistic Annotation: Survey and perspectives. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Doğan, Mehmet U./Maegaard, Bente/Mariani, Joseph/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12). Istanbul: European Language Resources Association (ELRA), S. 303–310. www.lrec-conf.org/proceedings/lrec2012/pdf/911_Paper.pdf (Stand: 10.12.2025).

Chiarcos, Christian/Nordhoff, Sebastian/Hellmann, Sebastian (2012): Linked data in linguistics: Representing and connecting language data and language metadata. Berlin/Heidelberg: Springer.

Cimiano, Philipp/McCrae, John P./Buitelaar, Paul (Hg.) (2016): Lexicon model for ontologies: Final community group report 10 May 2016. www.w3.org/2016/05/ontolex/ (Stand: 9.6.2025).

Cimiano, Philipp/Buitelaar, Paul/McCrae, John P./Sintek, Michael (2011): LexInfo: A declarative model for the lexicon-ontology interface. In: Journal of Web Semantics 9, 1, S. 29–51. https://doi.org/10.1016/j.websem.2010.11.001.

Darwish, Rasha M. (2020): Nach der Corona-Pandemie: Hat das Virus den deutschen Wortschatz infiziert? In: Research in Language Teaching 1, 13, S. 377–418.

Depner, Shelley C.-Y. (2022): Wortschatz in der Coronavirus-Pandemie im Chinesischen und Deutschen: Lebensmetaphern, Kriegsmetaphern und die sozialen Bedeutungen. In: Interface – Journal of European Languages and Literatures 19, 2, S. 7–44.

Dettke, Barbara (1995): Die asiatische Hydra: Die Cholera von 1830/31 in Berlin und den preußischen Provinzen Posen, Preußen und Schlesien. (= Veröffentlichungen der Historischen Kommission zu Berlin 89). Berlin/Boston: De Gruyter.

Donalies, Elke (2011): Tagtraum, Tageslicht, Tagedieb: Ein korpuslinguistisches Experiment zu variierenden Wortformen und Fugenelementen in zusammengesetzten Substantiven. Mit einem Exkurs und zahlreichen Statistiken von Noah Bubenhofer. (= amades 42). Mannheim: Institut für Deutsche Sprache.

Durrell, Martin/Bennett, Paul/Scheible, Silke/Whitt, Richard J. (2012): The GerManC Corpus. Manchester: University of Manchester.

DWDS (2024) = Digitales Wörterbuch der deutschen Sprache (2024): Historische Korpora: Textkorpus bereitgestellt durch das Digitale Wörterbuch der deutschen Sprache. www.dwds.de/d/korpora/dtaxl (Stand: 4.6.2024).

Epps, Patience (2015): Historical linguistics and socio-cultural reconstruction. In: Bowern, Claire/Evans, Bethwyn (Hg.): The Routledge handbook of historical linguistics. (= Routledge Handbooks in Linguistics). London: Routledge, S. 579–597.

Ernst, Peter (2021): Deutsche Sprachgeschichte: eine Einführung in die diachrone Sprachwissenschaft des Deutschen. 3. Aufl. (= UTB 2583). Wien: facultas.

Fuchs, Julia (2021): Corona-Komposita und ‚Corona'-Konzepte in der Medienberichterstattung in Standardsprache und in Leichter Sprache. In: Zeitschrift für germanistische Linguistik 49, 2, S. 335–368.

Havryliv, Oksana (2023): Expressiv-emotive Corona-Neologismen: morphologische und lexikalisch-semantische Aspekte am Beispiel des Deutschen und Ukrainischen. In: Jakosz, Mariusz/Kałasznik, Marcelina (Hg.): Corona-Pandemie im Text und Diskurs. Bd. 2. (= Fields of Linguistics 2). Göttingen: V&R unipress, S. 27–50.

Innerwinkler, Sandra (2015): Neologismen. (= Literaturhinweise zur Linguistik 1). Heidelberg: Winter.

Ivanov, Andrey/Nikonova, Zhanna/Frolova, Natalia/Muratbayeva, Indira (2023): Linguistic potential of COVID-19 neologisms in the metaphoric language of socio-political discourse. In: Communication and the Public 8, 3, S. 237–253. https://doi.org/10.1177/20570473231186475.

Jakosz, Mariusz/Kałasznik, Marcelina (Hg.) (2023): Corona-Pandemie: Diverse Zugänge zu einem aktuellen Superdiskurs. Göttingen: V&R unipress.

Kabatnik, Susanne/Klee, Anne (2024): Historischer Pandemiewortschatz am Beispiel der Cholera – eine korpusbasierte quantitativ-qualitative Untersuchung Historischer Korpora. In: Sprachwissenschaft 49, 2/3, S. 297–328.

Kabatnik, Susanne/ Klee, Anne (2025): „Choleraausbruch zwang zur Rückkehr" – die Auswahl von Beispielen für das Pandemictionary-Projekt als lexikografisches Linked Open Data-Netzwerk. In: Lexicographica 41, 1, S. 131–148. https://doi.org/10.1515/lex-2025-0007.

Kabatnik, Susanne/Klee, Anne/Bamberg, Claudia/Burch, Thomas/Hinzmann, Maria (im Dr.): Linked Open Data als neues geisteswissenschaftliches Paradigma? Zu einer aktuellen Debatte in den Digital Humanities am Beispiel des „Pandemictionary". In: Wuttke, Ulrike/Seltmann, Melanie/Schröter, Christian/Baillot, Anne/Wachter, Christian/Nunn, Christopher: (Hg.): From global to local? Digitale Methoden in den Geisteswissenschaften im deutschsprachigen Raum: Ein Triptychon. Mainz: Akademie der Wissenschaften und Literatur Mainz.

Kilgarriff, Adam/Rychlý, Pavel/Smrz, Pavel/Tugwell, David (2004): The Sketch Engine. In: Williams, Geoffrey/Vessier, Sandra (Hg.): Proceedings of the 11th EURALEX International Congress. Bretagne: Université de Bretagne-Sud, S. 105–115.

Klosa-Kückelhaus, Annette (2022): Corona und die deutsche Sprachentwicklung. Von Abendlockdown bis Zweitimpfling. In: Wissenschaftsmanagement, S. 1–5.

Klosa-Kückelhaus, Annette/Kernerman, Ilan (2022): Lexicography of Coronavirus-related neologisms. (= Lexicographica – Series Maior 163). Berlin/Boston: De Gruyter.

Lanza, Claudia/Folino, Antonietta/Pasceri, Erika/Perri, Anna (2022): Lexicon of pandemics: A semantic analysis of the Spanish flu and the COVID-19 timeframe terminology. In: Journal of Documentation 78, 4, S. 933–952.

Leibniz-Institut für Deutsche Sprache (2006–): Neologismenwörterbuch. Neuer Wortschatz rund um die Coronapandemie. www.owid.de/docs/neo/listen/corona.jsp# (Stand: 9.6.2025).

Linke, Angelika (2003): Sprachgeschichte – Gesellschaftsgeschichte – Kulturanalyse. In: Henne, Helmut/Sitta, Horst/Wiegand, Herbert E. (Hg.): Germanistische Linguistik: Konturen eines Faches. (= Reihe Germanistische Linguistik 240). Tübingen: Niemeyer, S. 25–66.

Möhrs, Christine (2021): Ein Wortnetz entspinnt sich um „Corona". In: Klosa-Kückelhaus, Annette (Hg.): Sprache in der Coronakrise. Dynamischer Wandel in Lexikon und Kommunikation. Mannheim: IDS-Verlag, S. 34–36.

Moulin, Claudine (2020): Linguistische Kreativität in Pandemiezeiten – eine sprachhistorische Annäherung. In: Aptum 16, 2/3, S. 268–273.

Nossem, Eva (2023): Corona-Diskurse und sprachliche Grenzziehungen in und um Europa. In: Brodowski, Dominik/Nesselhauf, Jonas/Weber, Florian (Hg.): Pandemisches Virus – nationales Handeln. Räume – Grenzen – Hybriditäten. Wiesbaden: Springer, S. 109–122.

Nünning, Ansgar (Hg.) (2024): Grundbegriffe der Literaturtheorie. (= Sammlung Metzler). Berlin/Heidelberg: Springer.

Pommeranz, Johannes (2015): Per monstra ad astra. Eine Annäherung (= Ausstellungskataloge des Germanischen Nationalmuseums, Nürnberg), S. 10–17. [Originalveröffentlichung in: Monster: fantastische Bilderwelten zwischen Grauen und Komik. Ausstellung im Germanischen Nationalmuseum, Nürnberg vom 7. Mai bis 6. September 2015].

Sanders, Willy (2007): Das neue Stilwörterbuch. Stilistische Grundbegriffe für die Praxis. Darmstadt: Wissenschaftliche Buchgesellschaft.

Schätz, Konstantin/Kirchhoff, Susanne (2024): From party to pandemic – Frames and metaphors in the news coverage of the COVID-19 outbreak in Austria. In: Studies in Communication Sciences 24, 1, S. 9–26.

Schierholz, Stefan J./Uzonyi, Pál (Hg.) (2022): Grammatik. Bd. 1: Formenlehre. (= Wörterbücher zur Sprach- und Kommunikationswissenschaft 1.1). Berlin/Boston: De Gruyter.

Schneider, Roman/Lang, Christian (2022). Das grammatische Informationssystem *grammis* – Inhalte, Anwendungen und Perspektiven. In: Zeitschrift für germanistische Linguistik 50, 2, S. 407–427.

Stamm-Kuhlmann, Thomas (1989): Die Cholera von 1831: Herausforderungen an Wissenschaft und staatliche Verwaltung. In: Sudhoffs Archiv 73, 2, S. 176–189.

Traugott, Elizabeth C. (2000): Semantic change. An overview. In: Cheng, Lisa/Sybesma, Rint (Hg.): The first Glot International State-of-the-Article book: The latest in linguistics. Berlin/Boston: De Gruyter, S. 385–406. https://doi.org/10.1515/9783110822861.385.

Trier Center for Digital Humanities (Hg.) (o. D.): Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, Erstbearbeitung (1854-1961). www.woerterbuchnetz.de/DWB (Stand: 17.7.2025). [Digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/25].

Trier Center for Digital Humanities (Hg.) (o. D.): Goethe-Wörterbuch. www.woerterbuchnetz.de/GWB (Stand: 17.7.2025). [Digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/25].

Trier Center for Digital Humanities (Hg.): Herders Conversations-Lexikon (1. Auflage, 1854–1857). www.woerterbuchnetz.de/Herder (Stand 17.7.2025). [Digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/25].

Ullmann, Stephen (1970): Semantics. An introduction to the science of meaning. Oxford: Basil Blackwell. [Reprint].

Vrandečić, Denny/Pintscher, Lydia/Krötzsch, Markus (2023): Wikidata: The making of. In: Ding, Ying/Tang, Jie/Sequeda, Juan/Aroyo, Lora/Castillo, Carlos/Houben, Geert-Jan (Hg.): The ACM Web Conference 2023 (WWW '23): Companion: Proceedings of the ACM Web Conference 2023. (= ACM Digital Library). New York: Association for Computing Machinery, S. 615–624.

Kontaktinformation

JProf. Dr. Susanne Kabatnik
Universität Trier
Abteilung Computerlinguistik/Trier Center for Digital Humanities
Universitätsring 15
54296 Trier
E-Mail: kabatnik@uni-trier.de

Anne Klee
Universität Trier
Trier Center for Digital Humanities
Universitätsring 15
54296 Trier
E-Mail: klee@uni-trier.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.88

Daniel Knuchel/Xenia Bojarski/Davide Ventre/Sonja Huber/Gunilla Kaibel

Diskurswandel korpuspragmatisch aufspüren

Analyse- und Visualisierungsmethoden im Kontext der *data philology*

Abstract This article presents an approach for tracing discursive change in text corpora by combining methods from corpus pragmatics, visual linguistics, and data philology. Using a corpus of more than two million Swiss newspaper articles in German and published between 2000 and 2024, we show how dispersion measures, keyness analysis, collocation profiling, and temporally aligned word embeddings (TWEC) can be brought together within a modular analytical and visualization toolbox which we call Visual Corpus Pragmatics. In our perspective, visualizations serve not only as illustrations but as epistemic instruments which make semantic and pragmatic shifts empirically visible and open for interpretation. The approach is based on an interplay between quantitative procedures and hermeneutic reading, resulting in a reflexive and methodologically transparent framework for investigating discourse in its historical and social context.

Keywords discourse analysis, corpus pragmatics, visual linguistics, data philology, toolbox

1. Einleitung

Diskurse zu untersuchen, hat sich für die germanistische Linguistik als äusserst produktiv erwiesen, wobei spezifisch die Untersuchung von diskursivem Wandel oft unerwartete Ergebnisse zutage bringen kann.[1] Die Diskurslinguistik hat sich dementsprechend im deutschsprachigen Raum als Teildisziplin etablieren können, wie nicht zuletzt unterschiedliche Einführungsbücher belegen (vgl. Spitzmüller/Warnke 2011; Niehr 2014; Bendel Larcher 2015). Methodisch ist die Diskurslinguistik unterdessen breit aufgestellt und hat sich einem Methodenpluralismus verschrieben, wobei gerade diskursiver Wandel in der Regel korpusbasiert untersucht wird (vgl. grundlegend dazu Busse/Teubert 1994). Korpusbasiert bedeutet vor diesem Hintergrund in erster Linie, dass ‚authentische' Sprachdaten im Zentrum der Untersuchung stehen, wobei die Korpusgrösse je nach Untersuchungsgegenstand stark variieren kann. Diese Breite zeigt sich zudem auch in der Diversität der Operationalisierung: Von eher qualitativ-hermeneutischen Textanalysen (vgl. rezent bspw. Happ 2024) bis zu quantitativ-korpuslinguistischen Zugängen (vgl. rezent bspw. Feldmüller i. Dr.) werden mittlerweile verschiedenste Analyseverfahren angewandt. Insbesondere korpuslinguistische Klassiker wie Keyword- und Kollokationsanalysen haben sich für die Untersuchung musterhaften Sprachgebrauchs, der als Effekt von Diskursivität verstanden wird (vgl. grundlegend zur Korpuslinguisik als Methode der Diskursanalyse Bubenhofer 2009), als produktiv und erkenntnisreich erwiesen. Im Zuge des sogenannten „data-driven" Turns (vgl.

1 Der vorliegende Aufsatz wurde in Schweizer Rechtschreibung verfasst.

Bubenhofer/Scharloth 2015) hat sich das Potenzial korpuslinguistischer Methoden für Diskursanalysen nochmals akzentuiert: Korpuslinguistische Methoden fungieren längst nicht mehr nur als „Zettelkästen" – im Sinne einer Belegstellenverwaltung – zur Bestätigung hermeneutischer Vorannahmen, sondern bieten eigenständige, analytische Zugänge, um diskursive Muster und Verschiebungen im Sprachgebrauch empirisch aufzuspüren und interpretativ zu festigen.

Zur Untersuchung diskursiver Dynamiken eröffnet der Einsatz digitaler, datenintensiver Methoden neue Perspektiven: Die Integration von Verfahren wie Keyness-, Dispersions- und Kollokationsanalysen sowie Methoden der distributionellen Semantik (vgl. z. B. Word Embeddings Knuchel/Bubenhofer 2023) ermöglicht es, semantische und pragmatische Verschiebungen in grossen Textkorpora systematisch und diachron zu analysieren. Dabei gewinnen Visualisierungen dieser Analysen eine zentrale Bedeutung: Sie dienen nicht lediglich der Illustration, sondern werden als epistemische Werkzeuge eingesetzt, mit denen komplexe Muster, Brüche und Persistenzen im Diskurs sichtbar und interpretierbar gemacht werden können (vgl. grundlegend zu einer linguistischen Diagrammatik Bubenhofer 2020). Der vorliegende Beitrag greift diesen Paradigmenwechsel auf und diskutiert, wie sich diskursive Dynamiken durch die Kombination korpuspragmatischer Methoden und datengeleiteter Visualisierungen noch gezielter erfassen lassen. Konkretes Ziel ist es, das methodische und heuristische Potenzial eines solchen korpuspragmatischen und visuellen Analyse-Toolsets auszuloten und seine praktische Anwendung exemplarisch zu demonstrieren. Zudem wird das Toolkit zur freien Nutzung zur Verfügung gestellt.[2] Im Fokus steht dabei die Frage, wie quantitative Verfahren und Visualisierungstechniken im Zusammenspiel mit hermeneutischen Interpretationsansätzen genutzt werden können, um diskursiven Wandel datengeleitet aufzuspüren und neue Zugänge zur Deutung von Diskursphänomenen zu eröffnen.

Im Folgenden wird zunächst der theoretische und methodologische Rahmen der Korpuspragmatik, Visual Linguistics und *data philology* skizziert (Kap. 2). Anschliessend werden die methodischen Schritte und visuellen Analyseverfahren anhand eines Korpus aus Schweizer Printmedien exemplarisch erläutert und angewandt (Kap. 3). Abschliessend reflektiert Kapitel 4, welche spezifischen Formen diskursiven Wandels durch diesen visuellen und datengeleiteten Zugang sichtbar werden.

2. Visuelle Korpuspragmatik und *data philology*

Die Analyse diskursiven Wandels in digitalen Korpora stellt spezifische Anforderungen an Theorie und Methodik. Im Folgenden werden drei komplementäre Perspektiven herausgearbeitet, die das methodische Fundament dieses Beitrags bilden: *Korpuspragmatik*, *Visuelle Linguistik* und *data philology*.

2.1 Korpuspragmatik: Sprachgebrauch im Diskurskontext

Korpuspragmatik lässt sich als methodologischer Ansatz verstehen, der an der Schnittstelle von Korpuslinguistik, Pragmatik sowie Sozial- und Kulturwissenschaft operiert (vgl. Felder/Müller/Vogel (Hg.) 2012; Scharloth 2018). Im Mittelpunkt steht damit nicht nur die blosse Erfassung von Worthäufigkeiten oder Kollokationen, sondern insbesondere die Untersuchung von Sprachgebrauch als sozial, medial und diskursiv eingebettete Praxis.

2 Das Toolkit ist unter folgendem Link zugänglich: https://gitlab.uzh.ch/daniel.knuchel/visual-corpus-pragmatics (Stand: 19.12.2025) und wird weiterhin optimiert und ggf. erweitert.

Die Korpuspragmatik begreift Sprache als ein dynamisches Mittel gesellschaftlicher Aushandlung, das in verschiedenen Kontexten unterschiedliche Funktionen und Bedeutungen annimmt. Muster und Wandel in der Verwendung bestimmter Ausdrücke, die Wechselwirkungen zwischen Akteuren, Themen und Diskurspositionen sowie die kontextgebundene Bedeutungsentwicklung werden dabei als zentrale Gegenstände empirisch-korpusbasierter Analysen erschlossen.

Der Ansatz der Korpuspragmatik erweitert die klassische Korpuslinguistik um eine explizit kontextualisierende Perspektive: Wie Scharloth (2018, S. 63) argumentiert, sind die Rehabilitation der sprachlichen Oberfläche sowie die Prämisse, dass Sprachgebrauch den Kontext (mit)herstellt, zentral. Im Mittelpunkt stehen dabei nicht die abstrakten Regelsysteme der Sprache, sondern die soziale Praxis des Sprachgebrauchs in situativen, medialen und thematischen Kontexten. Damit rückt die Analyse typischer, rekurrenter und signifikanter Muster in den Fokus (vgl. dazu aus kulturlinguistischer Perspektive Linke 2011). Dies wiederum setzt Sprachgebrauchsmuster zentral, die sowohl stabile als auch dynamische Aspekte von Diskursen sichtbar machen (vgl. Bubenhofer 2009). Wie Bubenhofer (ebd.) zeigt, lassen sich solche Sprachgebrauchsmuster mit korpuslinguistischen Methoden „berechnen“, so dass Veränderungen, Brüche oder Neujustierungen in Diskursen jenseits einzelner Texte datengeleitet und systematisch erforscht werden können.

Auch Felder/Müller/Vogel (2012, S. 16–21) sehen die Stärke der Korpuspragmatik in der Integration quantitativer und qualitativer Zugänge in das Forschungsdesign: Statistische Verfahren dienen der Identifikation von Mustern, sie werden jedoch nie isoliert angewendet, sondern stets mit interpretativen, diskursanalytischen und sozialtheoretischen Kategorien rückgebunden und argumentiert.

Korpuspragmatische Analysen sind damit durch einen Methodenpluralismus geprägt, der quantitative Empirie, Kontextsensibilität und interpretative Offenheit verbindet (vgl. Scharloth/Bubenhofer 2012; Knuchel 2024, S. 56–60). Die Analyseziele richten sich dabei auf die Schnittstelle zwischen Sprachgebrauch und gesellschaftlicher Wirklichkeit: Diskursiver Wandel wird nicht bloss als Frequenzverschiebung begriffen, sondern als Ausdruck sich verändernder Kommunikationspraktiken, thematischer Relevanzsetzungen und sozialer Aushandlungsprozesse, die durch den Kotext kontextualisiert und herausgearbeitet werden können.

2.2 Visualisierung als epistemisches Werkzeug

Die Visualisierung von Ergebnissen einer Korpusanalyse wird in der visuellen Linguistik als epistemisches Instrument verstanden, das weit über den Status blosser Illustration hinausgeht (vgl. zu einer Grundlegung einer Visuellen Linguistik Bubenhofer 2020). Visualisierungen dienen dazu, in komplexen und vielschichtigen Datenbeständen Muster, Brüche, Trends und diachrone Dynamiken sichtbar zu machen, die in rein textuellen oder tabellarischen Darstellungen unsichtbar bleiben würden. Gleichzeitig sind Visualisierungen immer Ergebnisse diagrammatischer Transformationen, so dass die Visualisierungen nur vor dem Hintergrund der Ausgangsdaten interpretiert werden sollten.

Diagramme, Netzwerkgraphen, Zeitreihen, Heatmaps und weitere Visualisierungsformen übernehmen in der Analyse also sowohl die Funktion des Zeigens als auch des Explorierens der Datengrundlage: Sie machen Relationen zwischen Begriffen, Verschiebungen in semantischen Feldern oder Veränderungen im thematischen Fokus über Zeit und Medien hinweg nachvollziehbar, geben Hinweise darauf oder legen diese gar offen. Wie Bubenhofer (2020) argumentiert, sind Visualisierungen dabei nie neutral, sondern stets Ergebnis

methodischer Entscheidungen – sei es hinsichtlich der gewählten Aggregation, der Granularität der Daten oder der Visualisierungsform. Die methodischen Entscheidungen müssen daher stets reflektiert und mitinterpretiert werden (vgl. exemplarisch zu Keyness Knuchel 2024, S. 83–105).

Die epistemische Stärke visueller Verfahren liegt darin, dass sie neue Hypothesen und Fragestellungen generieren können, die aus der rein quantitativen Analyse heraus nicht oder nur schwer zu erkennen wären. Im Zusammenspiel mit korpuspragmatischen Verfahren entsteht so ein heuristischer Zugang, der explorative, datengeleitete und interpretative Komponenten produktiv miteinander verbindet.

2.3 *Data philology*: Integriertes Lesen, Deuten und Modellieren

Data philology steht für eine forschungspraktische Haltung, in welcher der Umgang mit grossen digitalen Korpora nicht auf numerische Auswertung reduziert wird, sondern stets auf die interpretative Durchdringung und Kontextualisierung der gewonnenen Muster abzielt (vgl. Bubenhofer 2024). Im Zentrum dieses Ansatzes steht ein iterativer Forschungsprozess, der statistische Verfahren, algorithmische Analysen und Visualisierungen als heuristische Werkzeuge für die Generierung und Überprüfung neuer Fragestellungen begreift. Korpusanalysen werden nicht als abschliessende Resultate betrachtet, sondern als Ausgangspunkte für hermeneutische Deutung und theoriebasierte Kontextualisierung unter Einbezug des Kotextes der jeweiligen Analyseergebnisse. Es findet also stets eine Rückbindung an die Datengrundlage statt und macht die Interpretationen so nachvollziehbar und überprüfbar.

Data philology unterscheidet sich damit grundlegend von datengeleiteten Zugängen, die davon ausgehen, dass grosse Datenmengen alleine „für sich sprechen“ und die Analyse gleichzeitig schon als Interpretation darlegt. Vielmehr wird die Verbindung von algorithmischer Mustererkennung, visueller Exploration und theoretisch informierter Interpretation als zentral für die Analyse komplexer Phänomene wie des diskursiven Wandels, von Bedeutungsverschiebungen oder der Herausbildung neuer semantischer Felder verstanden.

Im Sinne einer *data philology* werden Korpusgenerierung, Korpusanalyse, Visualisierung und hermeneutische Reflexion als komplementäre Schritte eines iterativen Forschungsprozesses begriffen, der Offenheit für Ambiguität und die Notwendigkeit methodischer Reflexion gleichermassen einschliesst. Dabei bedeutet Ambiguität nicht, dass jede Interpretation Bestand hat – es geht vielmehr darum, eine Forschungsmethodologie zu etablieren, die Rückbezüge nicht nur erlaubt, sondern sogar erwartet und dabei mithilfe mehrerer Indikatoren Interpretationen zulässt und befürwortet.

Vor dem Hintergrund dieses methodologischen Rahmens bezeichnen wir unseren Zugang als Visuelle Korpuspragmatik. Im folgenden Kapitel werden wir nun die praktischen Schritte, Analyseverfahren und Visualisierungstechniken am Beispiel eines aktuellen Korpus aus Schweizer Printmedien demonstrieren.

3. Visuelle Korpuspragmatik in der Praxis

Um das methodische Vorgehen der visuellen Korpuspragmatik, wie wir sie hier vorschlagen, transparent zu machen, wird im Folgenden zunächst der Ablauf der Analyse als schematischer Workflow dargestellt (vgl. Abb. 1).

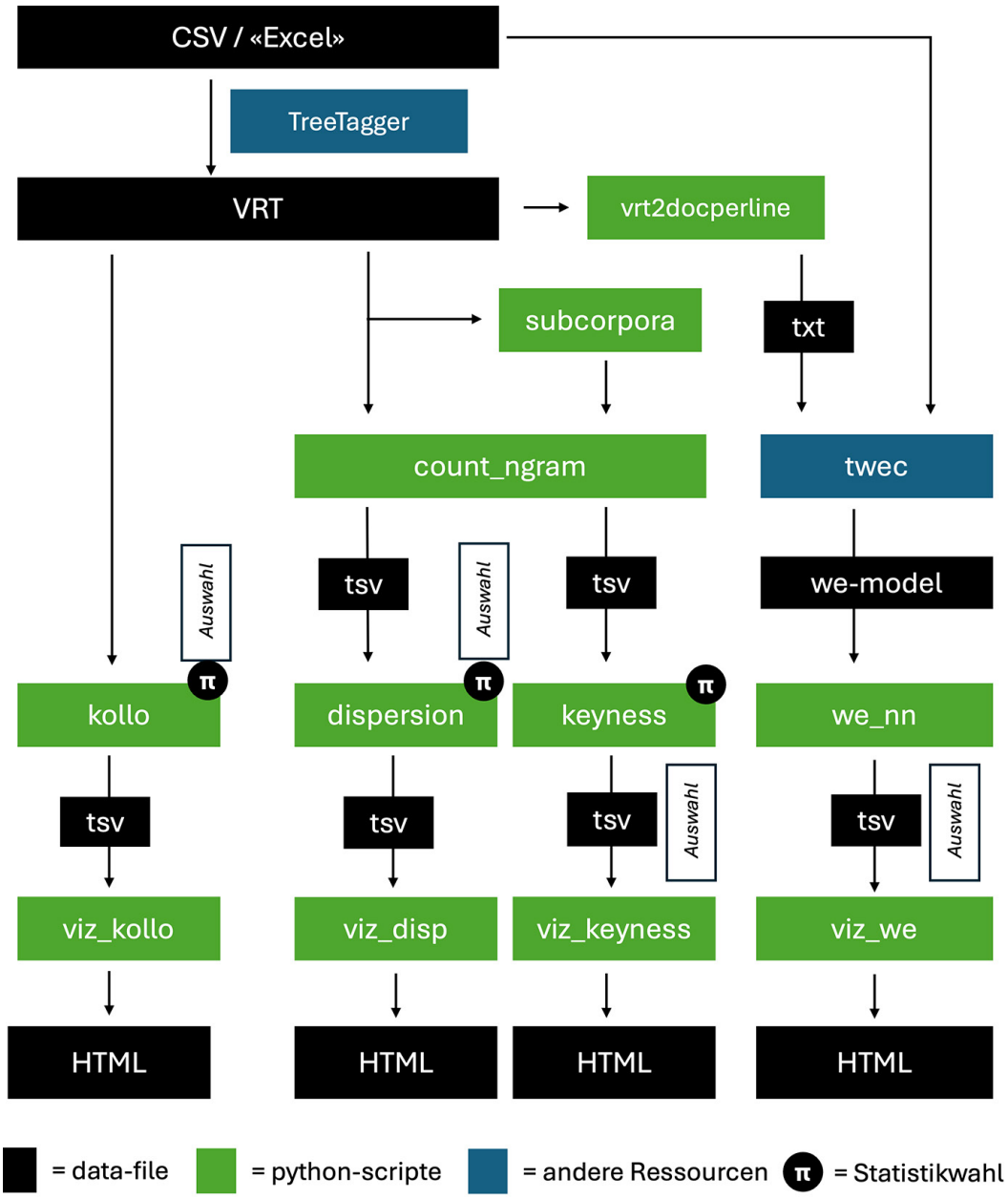

Abb. 1: Schematische Workflows der Toolbox *Visuelle Korpuspragmatik*

Abbildung 1 zeigt den schematischen Workflow unserer Toolbox *Visuelle Korpuspragmatik*.[3] Um die Toolbox nutzen zu können, liegen die Daten idealerweise in vertikalisierter

3 Die entsprechenden Python-Skripte und eine Anleitung sind unter https://gitlab.uzh.ch/daniel.knuchel/visual-corpus-pragmatics/-/tree/main/scripts (Stand: 19.12.2025) verfügbar.

Form[4] vor, so wie sie z. B. vom TreeTagger (vgl. Schmid 1994) generiert werden. In der Toolbox ist auch beschrieben, wie User:innen ein CSV-File – z. B. als Export aus einer Excel-Datei – in ein vertikalisiertes Format (verticalized text format) resp. in eine VRT-Datei (vgl. Fussnote 4) umwandeln können. Alle weiteren Analyse- und Visualisierungsschritte basieren auf der VRT-Datei des Korpus. Ausgehend von diesen vertikalisierten Dateien erfolgt entweder eine direkte Weiterverarbeitung und -analyse auf dem Gesamtkorpus oder die Erstellung von spezifischen Teilkorpora (= subcorpora). Die Toolbox umfasst verschiedene Python-Skripte (grün dargestellt), welche die textuellen Daten mithilfe ausgewählter quantitativer Verfahren analysieren. Dazu gehören Dispersions- (dispersion), Keyness- (keyness) und Kollokationsanalyse (kollibri) sowie Analysen mit einem Word Embedding-Modell (twec). Die Ergebnisse dieser quantitativen Analysen werden wiederum als tsv-Dateien gespeichert. Diese tsv-Dateien werden anschliessend mithilfe weiterer Skripte visualisiert (kol-viz, disp-viz, key-viz und WE-viz). Die daraus entstehenden HTML-Dateien, welche die Visualisierungen enthalten, werden in einer Webapplikation dargestellt, um eine interaktive Exploration der Resultate zu ermöglichen. Die gesamte Toolbox ist in einem Git repository organisiert, bestehend aus klar strukturierten Ordnern und Python-Skripten, was eine einfache Nutzung und Erweiterung gewährleistet.

Um dieses Verfahren praxisnah zu demonstrieren, analysieren wir im Folgenden exemplarisch ein umfangreiches Korpus journalistischer Texte aus der Schweiz. Zunächst stellen wir die Datengrundlage und deren Aufbereitung genauer vor, bevor wir anschliessend ausgewählte Verfahren anwenden und interpretieren.

3.1 Datengrundlage und Aufbereitung

Als Grundlage unserer exemplarischen Analysen nutzen wir ein Korpus mit Texten aus Schweizer Tageszeitungen, die über die Schnittstelle Swissdox@LiRI bei der Schweizer Medien Datenbank AG (SMD) heruntergeladen wurden.[5] Sämtliche Texte, die zwischen dem 1. Januar 2000 und 31. Dezember 2024 in der *Neuen Zürcher Zeitung*, dem *Tages-Anzeiger* und der Boulevardzeitung *Blick* erschienen sind, wurden in eine XML-Struktur überführt und mit dem TreeTagger (vgl. Schmid 1994) wurde eine Tokenisierung, Lemmatisierung und Annotation der Wortarten (Part-of-Speech-tagging) durchgeführt. Die so entstandenen vertikalisierten Textdateien, bildeten, wie im oberen Abschnitt beschrieben, die Grundlage für alle weiteren Schritte (vgl. dazu auch Abb. 1).

Wie bereits erwähnt, umfasst das Korpus Medientexte aus drei bedeutenden Schweizerdeutschen Tageszeitungen, die den grossen Verlagshäusern NZZ-Verlag, Tamedia und Ringier angehören. Es deckt somit einen erheblichen Teil der schweizerischen Presselandschaft ab und eignet sich gut zur Untersuchung diskursiven Wandels in der Schweizer

4 Als Datei in vertikalisierter Form wird eine Datei verstanden, bei der jedes Token in einer separaten Zeile steht. Pro Token werden mehrere Annotationen wie bspw. Lemma und Wortart (Part-of-Speech-Tag), in tabulatorgetrennten Spalten (token \t Part-of-Speech \t Lemma) gespeichert. Strukturelle Informationen wie zum Beispiel Textbeginn, Satzanfang oder Metadaten werden mit XML-artigen Tags markiert. Diese Tags sind in spitzen Klammern notierte Markierungen, die als formale Strukturierungselemente dienen und es ermöglichen zusätzliche Informationen maschinenlesbar zu kodieren. Wird in einem unserer Skripte eine VRT-Datei benötigt, dann ist damit eine in dieser Art strukturierte Datei gemeint.

5 Weitere Informationen zu Swissdox@LiRI: www.liri.uzh.ch/en/services/swissdox.html (Stand: 19.12.2025).

Medienlandschaft im 21. Jahrhundert.[6] Insgesamt enthält das Korpus 2.290.418 Artikel mit 1.058.078.928 laufenden Wortformen (= Tokens). Davon entfallen 42% der Artikel auf die *Neue Zürcher Zeitung* (ca. 497 Millionen Tokens), 38% auf den *Tages-Anzeiger* (ca. 411 Millionen Tokens) und 20% auf den *Blick* (ca. 149 Millionen Tokens). Dies ist aufgrund der Profile der Zeitungen auch zu erwarten. Wie in Abbildung 2 zudem ersichtlich wird, nimmt die Anzahl der publizierten Artikel ab den Jahren 2007/2008 kontinuierlich ab. Dies hängt mit Veränderungen in der Presselandschaft zusammen, was auch als Krise der klassischen Massenmedien infolge von Digitalisierung und damit einhergehenden Änderungen der Medienrezeption diskutiert wird (vgl. Meier (Hg.) 2017).

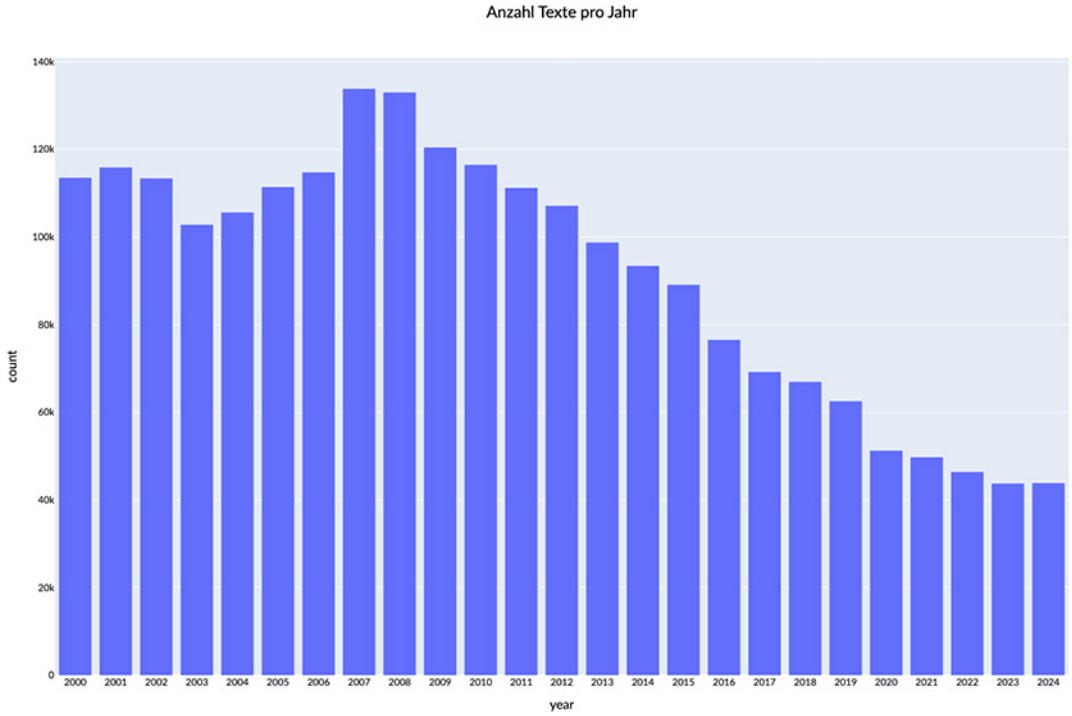

Abb. 2: Anzahl Texte (in Tausend) pro Jahr [Interaktive Version der Abb.][7]

Auffällig ist zudem die erhebliche Spannweite der Artikellängen. Der längste Artikel umfasst 54.794 Tokens, während der kürzeste nur neun Tokens enthält. Im Durchschnitt umfasst ein Artikel 461,96 Tokens, wobei der Median mit 335 Tokens deutlich darunter liegt. Die vergleichsweise niedrige Type-Token-Ratio von 0,0078 deutet auf eine erwartbare hohe lexikalische Redundanz hin, was typisch für umfangreiche journalistische Textkorpora ist.

6 Journalistische Texte bilden für viele diskurslinguistische Arbeiten die Grundlage, was einerseits mit der öffentlichkeitsherstellenden Funktion der Medien und andererseits mit der vergleichsweise einfachen Akquise über Datenbanken wie LexisNexis oder Swissdox@LiRI zusammenhängt. Aus diskurstheoretischer Perspektive ist diese Fokussierung jedoch nicht unproblematisch (vgl. dazu auch Warnke 2013; Linke 2015). Da es uns hier primär um die Funktionalität der Toolbox bei grossen Datenmengen und weniger um spezifisch diskursanalytische Erkenntnisse geht, nutzen wir dennoch ausschliesslich Pressetexte. Unsere Skripte sind für sämtliche schriftbasierten Daten geeignet, sofern diese vertikalisiert vorliegen und über XML-Attribute mit Metadaten ausgezeichnet sind.

7 Die HTML-Dateien der Visualisierungen können über den Link https://idsopen.de/article/view/88/102 „[Interaktive Version der Abb.]“ geöffnet und exploriert werden.

3.2 Dispersion und Keyness

Zur Aufdeckung diskursiver Dynamiken bieten wortschatzbasierte quantitative Analysen einen guten Einstieg im Rahmen der visuellen Korpuspragmatik. Dabei geht es darum, Muster in der Häufigkeit und Verteilung von Wörtern in einem Korpus zu identifizieren, um so diskursiven Wandel zu entdecken. Für das Explorieren der Streuung und des signifikanten Auftretens bestimmter lexikalischer Einheiten eignen sich insbesondere Dispersions- und Keynessanalysen. Basis für beide Verfahren sind Frequenzlisten verschiedener Teilkorpora. In unserem Beispiel fokussieren wir auf die zeitliche Dimension und bilden fünf Teilkorpora, die jeweils fünf Jahrgänge der ausgewerteten Zeitungen umfassen: (1) 2000–2004, (2) 2005–2009, (3) 2010–2014, (4) 2015–2019 und (5) 2020–2024. Für jedes dieser Teilkorpora zählen wir mithilfe eines Python-Skripts die absoluten Häufigkeiten aller Tokens aus. Diese bilden die Grundlage für die anschliessenden Berechnungen.

3.2.1 Dispersionsanalysen und ihre Visualisierung

Dispersionsanalysen untersuchen, wie gleichmässig oder konzentriert ein Ausdruck über die fünf Zeitperioden verteilt ist. Dazu können verschiedene statistische Masse genutzt werden, die jeweils leicht andere Dinge hervorheben (vgl. dazu Brezina 2018, S. 46–53). Für das gesamte Vokabular werden die Werte mit unterschiedlichen Massen berechnet und als Tabelle im tsv-Format gespeichert.[8] So kann einerseits der gesamte Wortschatz aufgrund seiner Streuung sortiert und exploriert werden. Andererseits kann auf der Grundlage der Tabelle eine Auswahl an spezifischen Ausdrücken bspw. zu einem Themenfeld miteinander verglichen werden.[9] Auf der Auswahl von spezifischen Ausdrücken basiert auch die Visualisierung, die wir hier beispielhaft zeigen, um die Exploration diskursiven Wandels zu illustrieren. In Abbildung 3 sind die Z-Scores verschiedener Ausdrücke rund um Sexualität als Heatmap dargestellt. Ein Z-Score gibt an, um wie viele Standardabweichungen die beobachtete Häufigkeit eines Ausdrucks in einer Periode vom Mittelwert über alle Perioden abweicht; positive Werte stehen für überdurchschnittliche, negative für unterdurchschnittliche Häufigkeit.

Die Heatmap (vgl. Abb. 3) zeigt die Verteilung ausgewählter Ausdrücke zum Themenfeld Sexualität und Geschlechtsidentität über die fünf Zeitabschnitte hinweg (jeweils 5 Jahre). Die Farbcodierung gibt den Z-Score der Wortfrequenz pro Periode an: Negative Werte (rot) stehen für eine unterdurchschnittliche Frequenz, positive Werte (blau) für überdurchschnittliche Verwendung im Vergleich zum Mittelwert über alle Zeitperioden. Ab der letzten Periode (2020–2024) treten Ausdrücke wie *queer*, *transgender*, *nonbinär*, *cis* und *trans* stark in den Vordergrund. Dies deutet auf eine diskursive Verschiebung hin, bei der neuere Ausdrücke die älteren teils ablösen oder ergänzen. Die insgesamt stärkere Differenzierung und Sichtbarkeit nicht-binärer Identitäten im jüngeren Zeitraum lässt sich so empirisch und datengeleitet nachvollziehen.

8 Das Skript berechnet zentrale Dispersions- und Häufigkeitsmasse: Mittelwert, Spannweite (Range), Standardabweichung (SD), Variationskoeffizient (CV, CV%), Juilland's D, Deviation of Proportions (DP) sowie die normierte Häufigkeit in „Vorkommen pro Million Wörter" (pmw).

9 Knuchel (2024) nutzt Dispersionsmasse, um Keywords rund um HIV|AIDS zu explorieren und so eine Interpretationshilfe für die Diskursivität der Ausdrücke zu haben.

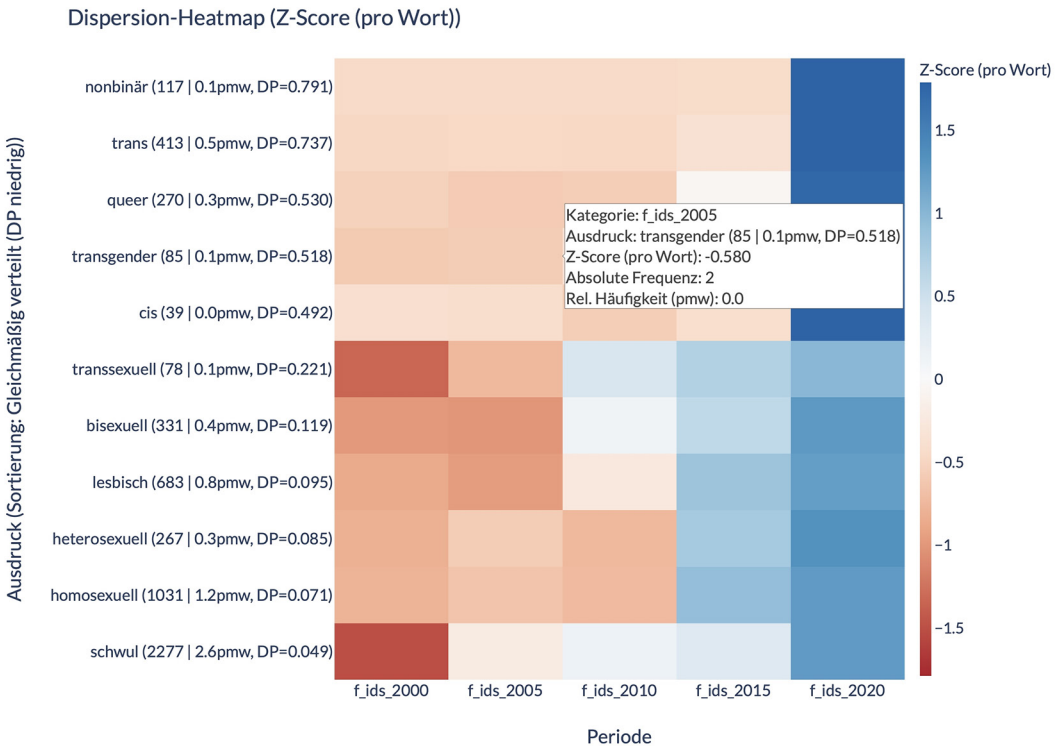

Abb. 3: Dispersionsvisualisierung von Ausdrücken rund um sexuelle Orientierung (Z-Score) [Interaktive Version der Abb.]

Die Heatmap erlaubt eine erste diachrone Rekonstruktion diskursiver Verschiebungen im Themenfeld Sexualität und Geschlechtsidentität. Solche Visualisierungen weisen auf Änderungen und Verschiebungen im Wortschatz hin, was als Hinweis auf diskursive Relevanzsetzung gelesen werden kann. Um solche Muster interpretativ einzuordnen, ist eine weitergehende Analyse mit anderen digitalen Methoden und qualitativer Textarbeit notwendig (vgl. dazu auch das Beispiel der Kollokationsanalyse).

3.2.2 Keynessanalysen und ihre Visualisierung

Keynessanalysen untersuchen, welche Ausdrücke in einem Untersuchungskorpus im Vergleich zu einem Vergleichskorpus signifikant häufiger vorkommen. Die zugrunde liegende Annahme ist, dass solche statistisch auffälligen Wörter Hinweise auf spezifische Themen, Perspektiven oder Diskursverschiebungen geben können. Vergleichskorpora können dabei auf Basis unterschiedlicher Kriterien gebildet werden, wobei die Datenauswahl einen Einfluss auf die Ergebnisse der Keynessanalyse hat (vgl. Goh 2011). Da in unserem Fall das für eine bestimmte Zeitperiode charakteristische Vokabular im Fokus steht, wird die Variable Zeit systematisch variiert. Im zugrunde liegenden Python-Skript sind dazu zwei Vergleichsverfahren implementiert: Zunächst wird jede Zeitperiode mit der Gesamtheit aller übrigen Perioden verglichen, um periodenspezifisches Vokabular zu identifizieren. Anschliessend erfolgt ein direkter Vergleich jeweils zweier aufeinanderfolgender Perioden, um semantische Verschiebungen und diskursive Trends feiner aufzulösen. Für die Berechnung werden verschiedene Masse verwendet, wobei wir aber auf die weitverbreitete Log-Likelihood-Ratio verzichten, da sie zwar statistische Signifikanz, aber keinerlei Aussage über den Grad der Typizität oder Relevanz eines Keywords liefert (vgl. zur Diskussion geeigneter Masse allgemein und den Problemen mit Signifikanztests Gabrielatos 2018; Knuchel 2024, S. 89–95). Stattdessen kombinieren wir mehrere alternative Masse, die unterschiedliche Aspekte lexikalischer Auffälligkeit abbilden. Dazu gehören die Bayesian Information Criterion (BIC)

zur statistischen Signifikanzbewertung sowie die diff_percentage, welche die prozentuale Abweichung zwischen den relativen Häufigkeiten in den beiden Korpora misst (vgl. Gabrielatos/Marchi 2012). Die Ausdrücke, die in beiden Korpora eine minimale relative Häufigkeit überschreiten, werden in einer Tabelle erfasst und nach den genannten Metriken sortierbar gemacht. Analog zur Dispersion lassen sich auch bei der Keyness auf Basis dieser tabellarischen Übersicht gezielt Ausdrucksgruppen – etwa thematisch verwandte Lexeme – explorieren und vergleichen. Darüber hinaus erlaubt eine Visualisierung als Heatmap die dynamische Exploration von Schlüsselwörtern über mehrere Zeitperioden hinweg. Dabei wird sichtbar, in welchen Zeitabschnitten bestimmte Wörter besonders prominent oder marginalisiert auftreten.

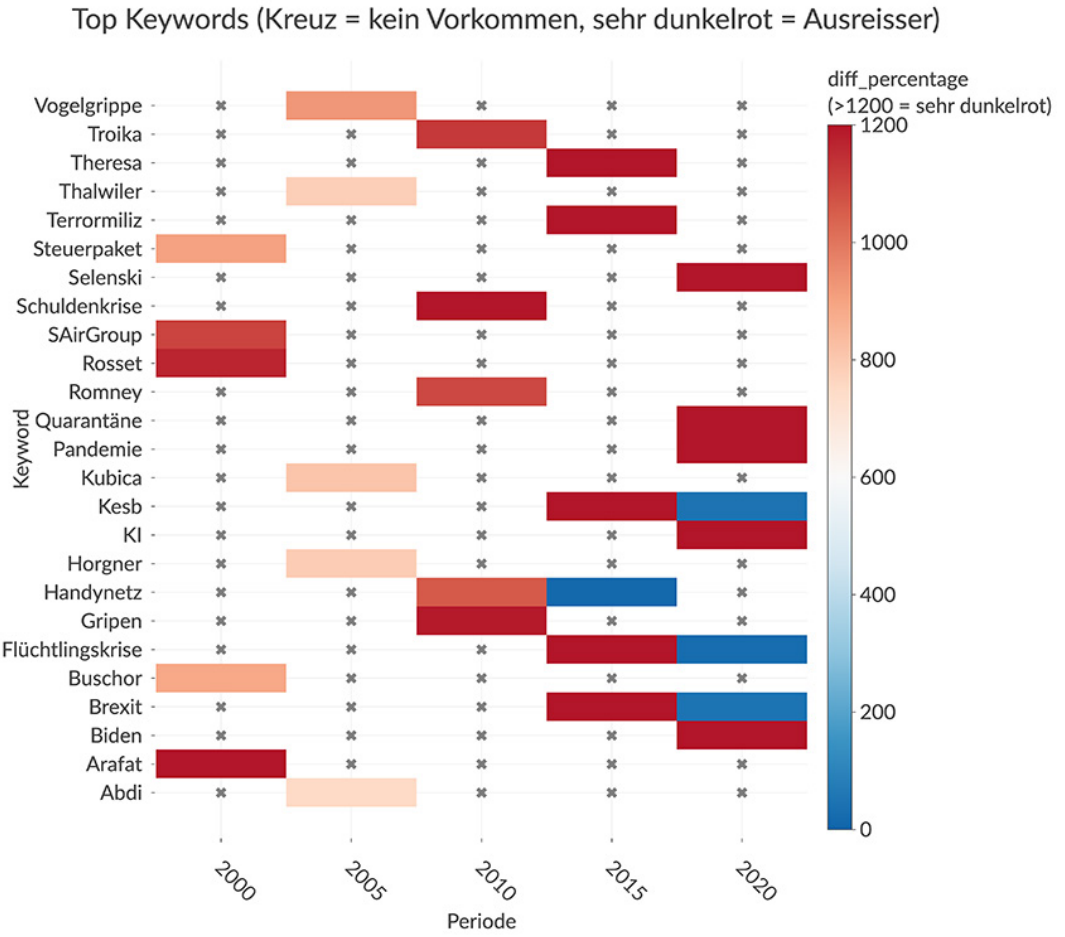

Abb. 4: Keynessvisualisierung als Heatmap (Top 5 Keywords pro Periode) [Interaktive Version der Abb.]

Die Heatmap zeigt jeweils die Ausdrücke mit der höchsten diff_percentage innerhalb der fünf Zeitperioden (2000–2024). Je dunkler das Blau, desto stärker hebt sich ein Ausdruck im Vergleich zu den übrigen Zeiträumen ab. Ein Kreuz (×) signalisiert, dass der Ausdruck im jeweiligen Zeitraum nicht als Keyword vorkommt. Deutlich erkennbar ist, dass bestimmte Ausdrücke stark periodenspezifisch auftreten und mit konkreten historischen Ereignissen korrelieren. In den frühen 2000er-Jahren prägen etwa *SAirGroup* oder *Rosset* das diskursspezifische Vokabular. In der zweiten Dekade erscheinen Ausdrücke wie *Handynetz* oder *Troika*, die auf politische oder technologische Entwicklungen verweisen. Ab 2020 dominieren pandemiebezogene, geopolitische und technologische Ausdrücke wie *Quarantäne*, *Pandemie*, *Selenski* oder *KI*. Die Visualisierung macht somit diskursiven Wandel greifbar, indem sie aufzeigt, welche Themen jeweils die öffentliche Kommunikation dominieren.

Wenn zudem zusätzlich mit thematischen Teilkorpora gearbeitet wird, zeigen sich diskursive Verschiebungen innerhalb eines Themas (vgl. exemplarisch für HIV|AIDS Knuchel 2024).

Die Keynessanalyse bietet damit wertvolle Hinweise auf zeitlich fokussierte diskursive Relevanzen und thematische Setzungen. Besonders bei stark periodenspezifischen Ausdrücken, die mit historischen Ereignissen korrelieren, ist entscheidend, diese in den jeweiligen Texten zu prüfen. Nur durch die gezielte Analyse der Kotexte kann nachvollzogen werden, welche semantischen, pragmatischen oder evaluativen Zuschreibungen mit diesen Keywords einhergehen.

3.3 Kollokationen

Während Dispersion und Keyness globale Verteilungsmuster im Korpus abbilden, erlauben Kollokationsanalysen einen tieferen Einblick in die lokalen semantisch-pragmatischen Kotexte einzelner Wörter. Kollokationen bezeichnen typische Wortverbindungen, also Wörter, die überzufällig häufig gemeinsam auftreten (vgl. Bubenhofer 2017). Die Analyse solcher Ko-Okkurrenzen gibt Aufschluss darüber, wie bestimmte Ausdrücke kontextualisiert, bewertet oder thematisch eingebettet werden – und damit, welche diskursiven Bedeutungsdimensionen ihnen in bestimmten Zeiträumen zugeschrieben werden.

Die Berechnung basiert auf einem Zielwort (sog. node), um das in einem festgelegten Fenster (z. B. ±5 Wörter) das Korpus nach typischen Begleitwörtern durchsucht wird. Diese werden anhand statistischer Masse wie z. B. T-score, MI (Mutual Information) oder LogDice gewichtet (vgl. zur Spezifik unterschiedlicher Masse für Kollokationen Brezina 2018, S. 66–70). Die Auswertung erfolgt je nach Zielsetzung entweder innerhalb eines Teilkorpus (z. B. für eine bestimmte Zeitperiode) oder im Vergleich mehrerer Zeitabschnitte. Im Rahmen der visuellen Korpuspragmatik dient die Kollokationsanalyse dazu, den Bedeutungswandel bestimmter Ausdrücke nachzuzeichnen. Die Resultate werden als tsv-Dateien gespeichert und können mit dem kol-viz-Skript visualisiert werden. In interaktiven Darstellungen lassen sich so semantische Profile der untersuchten Ausdrücke erkunden und vergleichen (vgl. Abb. 5).

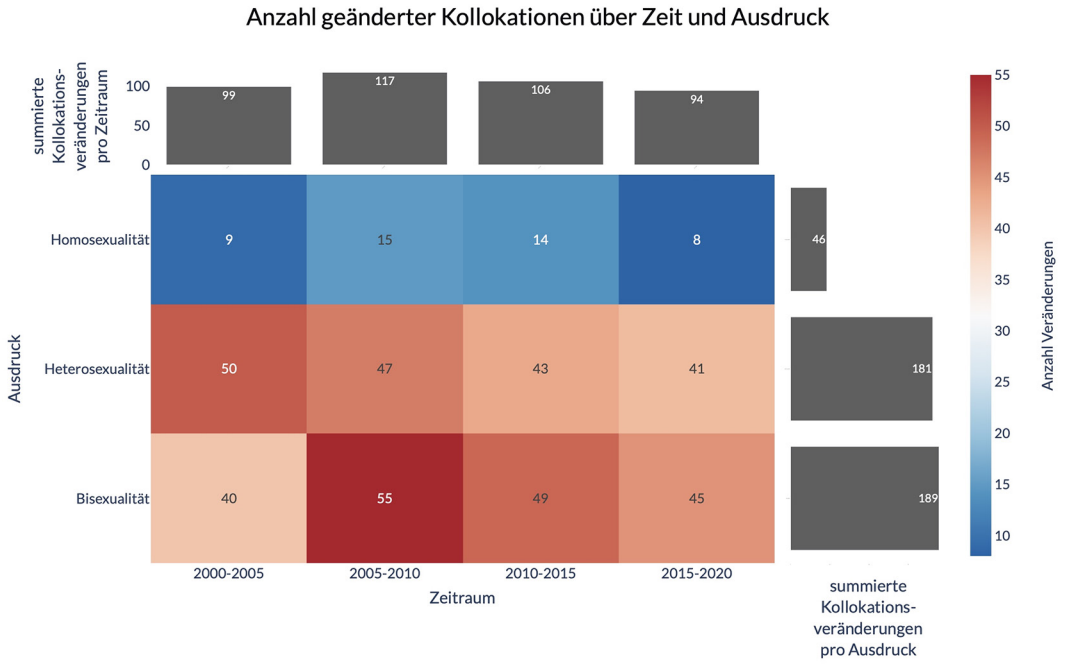

Abb. 5: Anzahl geänderter Kollokationen über Zeit und Ausdruck [Interaktive Version der Abb.]

In Abbildung 5 zeigt sich, wie sich die Kotexte der Ausdrücke *Homosexualität*, *Heterosexualität* und *Bisexualität* über vier zeitliche Intervalle hinweg verändern. Dabei wird zwischen zwei Ebenen unterschieden: Zum einen die Veränderungen pro Ausdruck, die über den gesamten Zeitraum hinweg summiert werden (summierte Kollokationsveränderungen pro Ausdruck), zum anderen die Veränderungen der Ausdrücke pro Zeitabschnitt, welche die Veränderungen aller Ausdrücke zusammenzählt (summierte Kollokationsveränderungen pro Zeitraum). Auf diese Weise wird sichtbar, welche Ausdrücke insgesamt besonders dynamisch sind und in welchen Zeitabschnitten die grössten Verschiebungen stattfinden.

Grundlage der Analyse ist der Vergleich der jeweils 50 bedeutendsten Kollokatoren (verwendetes Mass für die Visualisierung: t-score[10]) pro Ausdruck und Zeitintervall. Gezählt werden jeweils solche Kollokatoren, die im Vergleich zum vorhergehenden Zeitraum entweder neu auftreten oder verschwinden. Auf diese Weise lässt sich quantifizieren, in welchem Ausmass sich die diskursive Einbettung eines Ausdrucks über die Zeit verschiebt. Die interaktive Balkendarstellung im HTML gibt dabei Aufschluss über den Umfang dieser Veränderungen und erlaubt über Mouseover-Funktionen auch den Zugang zu den konkreten Kollokatoren, die sich in einem bestimmten Zeitraum geändert haben.

Auffällig ist zunächst die Stabilität von *Homosexualität*: Die Zahl der veränderten Kollokatoren bleibt durchgehend niedrig. Der Begriff scheint diskursiv etabliert zu sein; seine semantische Einbettung zeigt über die Zeit nur geringe Verschiebungen.

Heterosexualität und *Bisexualität* hingegen weisen eine deutlich höhere Veränderungsdynamik auf. Beide Ausdrücke zeigen über alle Zeitintervalle hinweg ähnlich viele neu auftretende oder verschwindende Kollokatoren. Während *Heterosexualität* dabei mit einem konstanten, leicht abnehmenden Wechsel einhergeht, lassen sich bei *Bisexualität* stärkere Ausschläge beobachten. Hier treten neue Kotexte auf, die auf narrative, identitätsbezogene oder popkulturelle Rahmungen verweisen. Kollokatoren wie *Pansexualität*, *Outing* oder *Prominenz* deuten auf eine wachsende diskursive Differenzierung und Sichtbarkeit hin.

Insgesamt erlaubt die interaktive Balkendarstellung somit die semantische Bewegung einzelner Begriffe sichtbar zu machen. Wo bei *Homosexualität* von einer diskursiven Konstanz ausgegangen werden kann, zeigen *Hetero-* und *Bisexualität* eine vergleichbar höhere Frequenz an semantischen Veränderungen. Diese Hinweise auf Wandel müssen in einem nächsten Schritt durch die Analyse konkreter Textstellen überprüft und kontextualisiert werden.

3.4 Word Embeddings und TWEC

Word Embeddings (vgl. Mikolov et al. 2013) und die darauf basierende Methode zur Darstellung über die Zeit, TWEC (vgl. Di Carlo et al. 2019), kurz für Training Word Embeddings with a Compass, gehören zu den Methoden der distributionellen Semantik. Sie erfassen und modellieren verschiedene semantische Beziehungen von Wörtern innerhalb eines Korpus.

Ein Word Embedding-Modell wird auf der Datengrundlage, dem Korpus, trainiert und gerechnet, wobei dafür jedem Wort im Korpus ein hochdimensionaler Vektor (in der Regel etwa 300 Dimensionen) zugewiesen wird. Dieser Vektor des Wortes im Vektorraum entspricht dem eigentlichen *word embedding*. Alle Wörter des Korpus werden so in den Vektorraum projiziert, dass Wörter, die sich im Korpus semantisch ähnlich sind – das heisst

10 Ein t-score misst, wie sehr sich die beobachtete Häufigkeit von der erwarteten Häufigkeit unterscheidet, unter der Annahme einer Normalverteilung (vgl. Brezina 2018; Evert 2008).

konkret, über einen ähnlichen Kotext verfügen – nah beieinander angesiedelt sind und Wörter, die sich semantisch nicht ähnlich sind, weiter voneinander weg im Raum liegen. Die paradigmatische Beziehung oder Ähnlichkeit lässt sich wiederum über verschiedene Masse bestimmen. Wir verwenden in unserem Beispiel die Kosinus-Similarität als Mass. Diese Kosinus-Ähnlichkeit wird berechnet, indem der Kosinuswert des Winkels der beiden Vektoren der betroffenen Wörter berechnet wird. Die resultierenden Werte liegen dementsprechend zwischen −1 und 1, wobei ein Wert nahe bei 1 einer hohen Ähnlichkeit der beiden Vektoren der Wörter im Vektorraum entspricht.

Da nicht nur semantische Ähnlichkeit modelliert wird, sondern auch andere Arten von semantischen Beziehungen aufgezeigt werden können, lassen sich mit den Vektoren der Wörter verschiedene Rechnungen durchführen, um verschiedene Beziehungen zu erfassen. Einer der häufigsten Verwendungszwecke in der Korpuspragmatik ist die Analyse der semantischen Ähnlichkeit mithilfe der sogenannten *Nearest Neighbors*, also den nächsten Nachbaren, zu einem Zielwort. Es werden also diejenigen Wörter betrachtet, bei denen der Kosinuswert zwischen dem Zielwort und den *Nearest Neighbors* am höchsten ist. Wird das Korpus in mehrere Zeitintervalle aufgeteilt und für jedes der Subkorpora ein Word Embedding Modell ausgehend vom Gesamtmodell (der Kompass zur Alignierung der Modelle[11]) gerechnet (in unserem Fall in Fünfjahresintervallen), so lässt sich über die Veränderungen der Kosinus-Ähnlichkeit verschiedener Wörter zum Zielwort auch hier diskursiver Wandel verfolgen.

Abbildung 6 zeigt beispielhaft die semantische Ähnlichkeit des Zielwortes *Identität* zu seinen verschiedenen nächsten Nachbarn über die Zeit. Wir verwenden für die Visualisierung wiederum eine Heatmap: Die Farbe einer Kachel zeigt die Kosinus-Ähnlichkeit des jeweiligen Wortes zum Zielwort *Identität* im jeweiligen Zeitintervall an. Je dunkelblauer die Kachel ist, desto ähnlicher ist das Wort dem Zielwort im jeweiligen Zeitintervall. Durch die Visualisierung verschiedener nächster Nachbaren in einer Visualisierung wird so nicht nur die Veränderung der semantischen Ähnlichkeit eines Wortes zum Zielwort aufgezeigt, sondern es können auch Wechselwirkungen zu anderen nächsten Nachbarn erkannt werden. Die Visualisierung ist zudem so programmiert, dass eine Sortierung der nächsten Nachbaren nach verschiedenen Parametern möglich ist, was verschiedene Perspektiven ermöglicht und die Visualisierung als Werkzeug noch besser operationalisierbar macht.[12]

Der nächste Nachbar in unserem Beispiel zum Zielwort *Identität* ist über die gesamte Zeit hinweg *Existenz*, wobei die Ähnlichkeit über die Zeit abnimmt und dafür die Ähnlichkeit zu anderen Wörtern wie *Sprache*, *Ideologie* und *Religion*, aber auch *Weltanschauung* und *Herkunft* zunimmt.

Gerade *Weltanschauung*, *Herkunft* und *Gesinnung* erfahren im letzten Zeitintervall eine deutliche Zunahme der Ähnlichkeit im Vergleich zu den vorherigen Zeitintervallen. Diese Auffälligkeiten geben uns Hinweise darauf, dass es sich dabei um diskursiven Wandel handeln könnte, die wir im Anschluss bei einer Rückbindung an die Texte im Korpus interpretativ ausformulieren könnten.

11 Jedes Word Embedding Modell beginnt mit einer zufälligen Initialisierung. Würde man daher vier verschiedene Subkorpora getrennt trainieren, entstünden vier voneinander unabhängige Modelle mit jeweils unterschiedlicher Trainingsbasis. Um dennoch eine Vergleichbarkeit aller Zeitabschnitte zu gewährleisten, wird das Gesamtkorpus als „Kompass" herangezogen, an dem die Modelle der einzelnen Zeitabschnitte ausgerichtet werden.

12 Sortierung hier nach Häufigkeit, in der Browserversion als HTML-File ist die Sortierung nach den verschiedenen Parametern möglich.

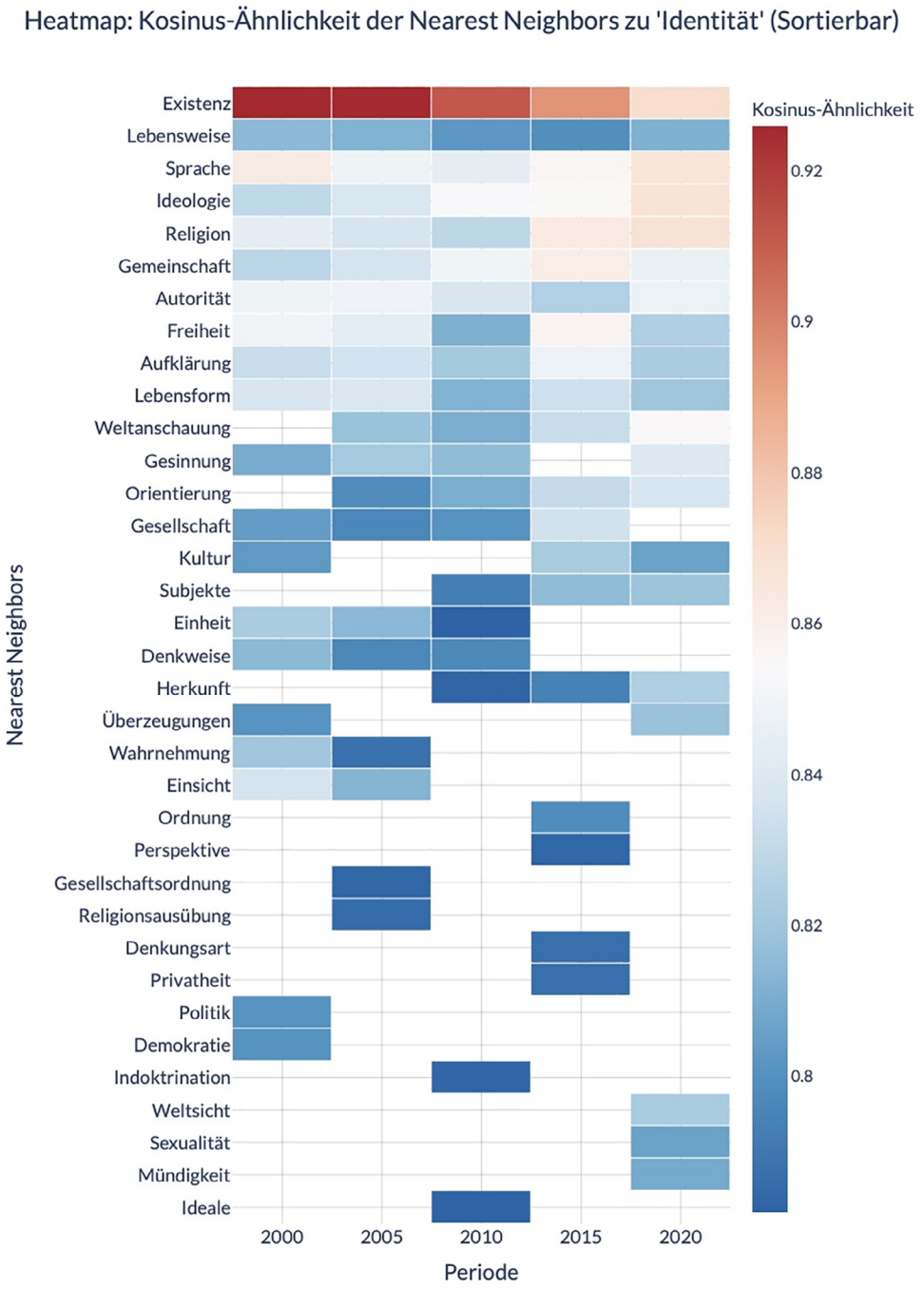

Abb. 6: Word Embedding-Visualisierung Nearest Neighbors zu *Identität* [Interaktive Version der Abb.]

Konstant ähnliche Wörter (mit kleinen Fluktuationen) wie *Existenz*, *Sprache*, *Ideologie*, *Religion*, *Gemeinschaft*, *Autorität* und *Freiheit* geben uns einen Hinweis darauf, welche Bedeutungsnuancen *Identität* innerhalb des Korpus zugeschrieben werden können.

Auffällig und somit überprüfenswert sind auch Wörter, die nur in einem Zeitintervall überhaupt einen Ähnlichkeitswert haben wie hier beispielsweise die folgenden Wörter: *Überzeugungen*, *Mündigkeit*, *Privatheit*, *Perspektive* und *Weltsicht* oder die Wörter *Politik* und *Demokratie*, die nur im ersten Zeitintervall als nächste Nachbaren auftreten. Das einmalige Auftreten ist ein Indikator dafür, dass in den jeweiligen Zeitintervallen ein besonderes Ereignis oder ein bestimmter thematischer Fokus die Datengrundlage so geprägt hat, dass die jeweiligen Wörter als nächste Nachbarn auftreten, was wiederum in den Texten verifiziert werden müsste.

Die mittels Word Embeddings modellierten semantischen Verschiebungen bieten wichtige Indikatoren für potenziellen diskursiven Wandel. Sie müssen jedoch stets im Kontext der konkreten Textverwendung überprüft und ausgedeutet werden. Erst durch die Kontextualisierung der ermittelten semantischen Relationen in den jeweiligen Texten lassen sich belastbare Aussagen über Bedeutungsverschiebungen und ihre diskursive Einbettung treffen.

4. Diskursiven Wandel korpuspragmatisch aufspüren – ein Fazit

Der vorliegende Beitrag hat exemplarisch demonstriert, wie sich diskursiver Wandel mithilfe korpuspragmatischer und visualisierender Verfahren im Paradigma der *data philology* systematisch untersuchen lässt. Ausgehend von einem umfangreichen Korpus Schweizer Pressetexte konnten mit Hilfe der Toolbox *Visuelle Korpuspragmatik* verschiedene Formen diskursiver Dynamiken sichtbar gemacht werden – von lexikalischer Streuung über Schlüsselwörter bis hin zu semantischen Verschiebungen im Vektorraum.

Dabei wurde deutlich, dass quantitative Verfahren wie Dispersions- und Keynessanalysen sowie Word Embeddings zentrale Hinweise auf potenzielle Brüche, Persistenzen und Re-Kontextualisierungen in der diskursiven Struktur eines Textkorpus liefern können. Visualisierungen sind dabei mehr als nur illustrative Ergänzung – sie fungieren als epistemische Werkzeuge, mit denen sprachliche Muster nicht nur sichtbar, sondern auch systematisch vergleichbar und interpretierbar gemacht werden können.

Auch Formen der Variation, etwa zwischen publizistischen Formaten (Tagespresse vs. Boulevard), thematischen Korpora oder institutionellen Sprecherpositionen, können mit den hier vorgestellten Verfahren untersucht und visuell zugänglich gemacht werden.

Entscheidend bleibt dabei, dass quantitative Muster nicht isoliert gelesen werden: Erst im Rückbezug auf die konkreten Texte und in Kombination mit hermeneutischer Analyse entfalten sie ihre analytische Tiefe. Die Toolbox der Visuellen Korpuspragmatik eröffnet so ein methodisch reflektiertes und zugleich explorativ anschlussfähiges Analyseinstrumentarium für die diskurslinguistische Forschung. In ihrer Verbindung von datengetriebenem Zugang und interpretativer Offenheit zeigt sich das Potenzial einer *data philology*, die Sprachgebrauchsmuster nicht nur zählt, sondern verstehen will – und damit den komplexen Aushandlungsprozessen gesellschaftlicher Kommunikation gerecht wird.

Literatur

Bendel Larcher, Sylvia (2015): Linguistische Diskursanalyse. Ein Lehr- und Arbeitsbuch. (= Narr Studienbücher). Tübingen: Narr.

Brezina, Vaclav (2018): Statistics in corpus linguistics: A practical guide. Cambridge: Cambridge University Press.

Bubenhofer, Noah (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. (= Sprache und Wissen 4). Berlin/New York: De Gruyter. https://doi.org/10.1515/9783110215854.

Bubenhofer, Noah (2017): Kollokationen, n-Gramme, Mehrworteinheiten. In: Sven Roth, Kersten/Wengeler, Martin/Ziem, Alexander (Hg.): Sprache in Politik und Gesellschaft. (= Handbücher Sprachwissen 19). Berlin/Boston: De Gruyter, S. 69–93. https://doi.org/10.1515/9783110296310-004.

Bubenhofer, Noah (2020): Visuelle Linguistik: Zur Genese, Funktion und Kategorisierung von Diagrammen in der Sprachwissenschaft. (= Linguistik – Impulse & Tendenzen 90). Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110698732.

Bubenhofer, Noah (2024): Die Lektüre von Texten und Daten: Data Philology statt Data Science. In: Zeitschrift für Literaturwissenschaft und Linguistik 54, 2, S. 269–283. https://doi.org/10.1007/s41244-024-00338-1.

Bubenhofer, Noah/Scharloth, Joachim (2015): Maschinelle Textanalyse im Zeichen von Big Data Und Data-Driven Turn – Überblick Und Desiderate. In: Zeitschrift für germanistische Linguistik 43, 1, S. 1–26. https://doi.org/10.1515/zgl-2015-0001.

Busse, Dietrich/Teubert, Wolfgang (1994): Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In: Busse, Dietrich/Hermanns, Fritz/Teubert, Wolfgang (Hg.): Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik. Opladen: Westdeutscher Verlag, S. 10–28.

Di Carlo, Valerio/Bianchi, Federico/Palmonari, Matteo (2019): Training temporal word embeddings with a compass. In: Proceedings of the AAAI Conference on Artificial Intelligence 33, 1, S. 6326–6334. https://doi.org/10.1609/aaai.v33i01.33016326.

Evert, Stefan (2008): Corpora and collocations. In: Lüdeling, Anke/Kytö, Merja (Hg.): Corpus Linguistics. An International Handbook. Berlin/New York: De Gruyter. https://doi.org/10.1515/9783110213881.2.1212.

Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (2012): Korpuspragmatik. Paradigma zwischen Handlung, Gesellschaft und Kognition. In: Felder/Müller/Vogel (Hg.), S. 3–32. https://doi.org/10.1515/9783110269574.3.

Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hg.) (2012): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. (= Linguistik – Impulse & Tendenzen 44). Berlin/Boston: De Gruyter.

Feldmüller, Tim (im Dr.): Das Framing von Extremismusvarianten im medialen Diskurs der Jahre 1999–2021: Eine corpus-driven Methode zur Erschließung und Visualisierung semantischer Frames. Language Science Press.

Gabrielatos, Costas (2018): Keyness analysis. Nature, metrics and techniques. In: Taylor, Charlotte/Marchi, Anna (Hg.): Corpus approaches to discourse: a critical review. London u. a.: Routledge, S. 225–258.

Gabrielatos, Costas/Marchi, Anna (2012): Keyness: Appropriate metrics and practical issues. Präsentiert auf: CADS International Conference 2012, Bologna, University of Bologna.

Goh, Gwang-Yoon (2011): Choosing a reference corpus for keyword calculation. In: Linguistic Research 28, 1, S. 239–256.

Happ, Alexander M. (2024): Der Einsatz der Bundeswehr im Ausland 1990–2015: Eine diskurslinguistische Untersuchung anhand von Argumentationen. (= Sprache und Wissen 61). Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783111338613.

Knuchel, Daniel (2024): ‚HIV|AIDS' in der Ära der Post-Infektiosität – Korpuspragmatische Analysen zur sprachlichen Konzeptualisierung einer Infektionskrankheit. Diss. Zurich: University of Zurich. https://doi.org/10.5167/UZH-263625.

Knuchel, Daniel/Bubenhofer, Noah (2023): Machine Learning und Korpuspragmatik. Word Embeddings als Beispiel für einen kreativen Umgang mit NLP-Tools. In: Meier-Vieracker, Simon/Bülow, Lars/Marx, Konstanze/Mroczynski, Robert (Hg.): Digitale Pragmatik. (= Digitale Linguistik 1). Heidelberg: Metzler, S. 213–235.

Linke, Angelika (2011): Signifikante Muster – Perspektiven einer kulturanalytischen Linguistik. In: Wåghäll Nivre, Elisabeth/Kaute, Brigitte/Andersson, Bo/Landén, Barbro/Stoeva-Holm, Dessislava (Hg.): Begegnungen. Das VIII. Nordisch-Baltische Germanistentreffen in Sigtuna vom 11. bis zum 13.6.2009. (= Acta Universitatis Stockholmiensis). Stockholm, S. 23–44.

Linke, Angelika (2015): Entdeckungsprozeduren – Oder: Wie Diskurse auf sich aufmerksam machen. In: Kämper, Heidrun/Warnke, Ingo H. (Hg.): Diskurs – interdisziplinär. Zugänge, Gegenstände, Perspektiven. (= Diskursmuster/Discourse Patterns 6). Berlin/München/Boston: De Gruyter, S. 63–86. https://doi.org/10.1515/9783050065281-005.

Meier, Werner A. (Hg.) (2017): Abbruch – Umbruch – Aufbruch. Globaler Medienwandel und lokale Medienkrisen. (= Medienstrukturen 11). Baden-Baden: Nomos.

Mikolov, Tomas/Chen, Kai/Corrado, Greg/Dean, Jeffrey (2013): Efficient estimation of word representations in vector space. In: arXiv:1301.3781 [cs.CL]. http://arxiv.org/abs/1301.3781 (Stand: 19.12.2025).

Niehr, Thomas (2014): Einführung in die linguistische Diskursanalyse. (= Einführung Germanistik). Darmstadt: WBG.

Scharloth, Joachim (2018): Korpuslinguistik für sozial- und kulturanalytische Fragestellungen. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Korpuslinguistik. (= Germanistische Sprachwissenschaft um 2020 5). Berlin/Boston: De Gruyter, S. 61–80. https://doi.org/10.1515/9783110538649-004.

Scharloth, Joachim/Bubenhofer, Noah (2012): Datengeleitete Korpuspragmatik. Korpusvergleich als Methode der Stilanalyse. In: Felder/Müller/Vogel (Hg.), S. 195–230. https://doi.org/10.1515/9783110269574.195.

Schiller, Anne/Teufel, Simon/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Stuttgart, Tübingen. Microsoft Word – STTS-Tagset_m_Bsp.doc (Stand: 19.12.2025).

Schmid, Helmut (1994): Probabilistic Part-of-Speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK. www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf (Stand: 19.12.2025).

Spitzmüller, Jürgen/Warnke, Ingo H. (2011): Diskurslinguistik: eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. (= De Gruyter Studium). Berlin/New York: De Gruyter.

Warnke, Ingo H. (2013): Urbaner Diskurs und maskierter Protest – Intersektionale Feldperspektiven auf Gentrifizierungsdynamiken in Berlin Kreuzberg. In: Sven Roth, Kersten/Spiegel, Carmen (Hg.): Angewandte Diskurslinguistik. Felder, Probleme, Perspektiven. (= Diskursmuster/Discourse Patterns 2). Berlin: Akademieverlag, S. 189–221.

Kontaktinformation

Dr. Daniel Knuchel
Universität Zürich/Deutsches Seminar
Schönberggasse 9
8001 Zürich
Schweiz
E-Mail: daniel.knuchel@ds.uzh.ch

Xenia Bojarski
Universität Zürich/Deutsches Seminar
Schönberggasse 9
8001 Zürich
Schweiz
E-Mail: xenia.bojarski@ds.uzh.ch

Davide Ventre
Universität Zürich/Deutsches Seminar
Schönberggasse 9
8001 Zürich
Schweiz
E-Mail: davide.ventre@ds.uzh.ch

Sonja Huber
Universität Zürich/Deutsches Seminar
Schönberggasse 9
8001 Zürich
Schweiz

Gunilla Kaibel
Universität Zürich/Deutsches Seminar
Schönberggasse 9
8001 Zürich
Schweiz
E-Mail: gunilla.kaibel@ds.uzh.ch

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.90

Jörn Stegmeier/Kevin Kuck/Anna Christina Kupffer/Anne Christine Günther/Angela Hammer/Dario Kampkaspar/Marcus Müller/Thomas Stäcker

Die Digitalisierung des Darmstädter Tagblatts (1740–1986)

Drei Jahrhunderte deutscher Sprachentwicklung

Abstract The Darmstädter Tagblatt (1740–1986) is one of the longest-running regional newspapers in the German-speaking world. Since 2019, it has been digitised, linguistically processed, and made freely accessible as part of a DFG-funded project conducted by the University and State Library Darmstadt (ULB) and Technical University of Darmstadt. The digitisation process poses both technical and legal challenges, particularly regarding layout and script recognition as well as the clarification of rights for individual contributions. The processed data are provided as TEI-XML full texts and made available via platforms such as TUdigit, WDB+, and CQPweb. This enables new applications in diachronic linguistic research. As demonstrated in this paper, the corpus supports analyses of morphosyntactic patterns and lexical developments spanning three centuries. Additional examples highlight the newspaper's transformation from a publication focused on classified advertisements to a modern daily, as well as the geographical reach of Darmstadt in the 18th century.

Keywords digitisation, corpus creation, rights clearance, newspaper, TEI, full texts

Num. III. 1750. den 15. Januarii.
Mit Hochfürstl. Gnädigstem PRIVILEGIO.

Darmstädtisches
Frag- und Anzeigungs-Blättgen/
zu finden bey Gottfried Heinrich Eylau, Hof-Buchdrucker.

AVERTISSEMENT.

Es wird hierdurch angezeiget und bekannt gemacht, daß das in der Langen-Gaß gelegene, bisher gewesene Dreystöckigte Alte Waysen-Hauß an den Meistbietenden soll abgelassen und verkaufft werden. Es stehet daselbige in einem aufgemauerten gantz steinernen Sarg nebst einem Neben- und Seiten-Bau, hat zwey grose Säle, sieben Wohn-Stuben, zwey Küchen, fünff so wol kleine als grose Kammern, einen fürtrefflich guten und grosen Keller, vier grose Böden zu Früchten, fein Wasch-Hauß, Stallung und Holtz-Schoppen, nebst andern Nothwendigkeiten. Wer hierzu Lusten hat, kan sich in dem Hauß

Abb. 1: Titelseite des Darmstädter Tagblatts vom 15.1.1750

1. Projektorganisation

Das hier vorgestellte Projekt[1] wird von der DFG gefördert[2] und ist eine Kooperation zwischen der Universitäts- und Landesbibliothek Darmstadt (ULB, Prof. Dr. Thomas Stäcker) und dem Fachgebiet Digitale Linguistik am Institut für Sprach- und Literaturwissenschaft der TU Darmstadt (Prof. Dr. Marcus Müller). Das Projekt umfasst zwei Phasen: Die erste Projektphase dauerte von 2019 bis 2022 und die zweite Phase begann 2023. In Phase 1 wurden die Jahrgänge 1740 bis 1941 digitalisiert. Die verbleibenden Jahrgänge bis 1986 folgen in Phase 2.

Die ULB ist dabei für die Erstellung der hochauflösenden Digitalisate und des maschinenlesbaren Volltextes zuständig. Der Kooperationspartner versieht zum einen den Volltext auf Token-Ebene mit Annotationen zu Wortart, Lemma und Named Entities (automatisierte Erkennung von z. B. Eigennamen, Ländernamen, Namen von Organisationen). Dabei entwickelt er zum anderen ein Verfahren, Urhebernennungen im Volltext zu erkennen, um damit die Lizenzanfragen im Rahmen der Rechteklärung, die in der zweiten Phase eine erhebliche Rolle spielt, zu unterstützen. Für die Veröffentlichung sind beide Kooperationspartner zuständig, da die Ergebnisse sowohl als durchsuchbarer Volltext als auch als Korpus zugänglich gemacht werden.[3]

2. Ausgangslage und Ziele

Zeitungen eröffnen auf mehreren Ebenen Zugänge zu einer Gesellschaft. Es sind dabei nicht nur die überregionalen Nachrichten, die in den Mittelpunkt des (Forschungs-)Interesses rücken. Regionale Nachrichten und Mitteilungen geben direkten Einblick in bestimmte Aspekte des öffentlichen Lebens, die bei überregionalen Nachrichten keine Rolle spielen. Dies gilt in besonderem Maße für Kleinanzeigen und Artikel über regionales Kultur- und Vereinsleben und zumindest teilweise auch für Werbeanzeigen. Unabhängig von den Ereignissen bieten Zeitungen – und insbesondere Regionalzeitungen – aber auch Zugang zum Sprachgebrauch einer bestimmten Region. Zeitungen, die über einen langen Zeitraum in derselben Region erscheinen, eröffnen all diese Zugänge auch diachron und erlauben somit vergleichende Studien.

Allerdings sind die Originalausgaben historischer Zeitungen nicht immer einfach zugänglich und eher umständlich im Gebrauch. Die in der Regel schlechte Papierqualität insbesondere von Zeitungen des 19. Jahrhunderts schränkt die Benutzbarkeit ebenfalls ein, bis dahin, dass Exemplare nicht mehr ausgehändigt werden können, weil sie durch den Gebrauch zu sehr strapaziert werden. Die Digitalisierung des Zeitungsbestands ist daher eine sinnvolle und auch notwendige Maßnahme, um sowohl interessierten Wissenschaftler:innen als auch Privatpersonen uneingeschränkten und bequemen Zugang bieten zu können. Im Zeitalter von automatischer Layout- (OLR, *optical layout recognition*), Buchstabenerkennung (OCR, *optical character recognition*) und internetbasiertem Zugriff sind frühere Digitalisate wie Filme oder Mikroformate jedoch oft nicht mehr ausreichend. Zum einen erfordert ihre Publikation über das Internet durchaus aufwendige Transformationsschritte und zum anderen reicht die Gesamtqualität meist nicht aus, um daraus maschinenlesbaren Volltext zu erstellen.

---

1 Eine ausführlichere Projektbeschreibung findet sich in Stegmeier et al. (2022).

2 Projektnummer 422840794.

3 Die weiterführenden Links sind über die Projektwebseite erreichbar: www.ulb.tu-darmstadt.de/forschen_publizieren/forschen/darmstaedter_tagblatt.de.jsp (Stand: 27.11.2025).

Das Darmstädter Tagblatt erschien in sich wandelnden Titeln und mit einer Unterbrechung während der 1940er Jahre von 1740 bis 1986, als es von der Echo Medien Gruppe aufgekauft und zugunsten des Lokalkonkurrenten, dem Darmstädter Echo, eingestellt wurde. Während seiner Existenz wandelte sich das Tagblatt von einem ‚Frag- und Anzeigeblatt', das zunächst nur einige Male pro Monat erschien, zu einer modernen Tageszeitung. Während die Anfangsphase primär Anzeigen unter Nennung der Produktpreise, Listen von in Darmstadt ankommenden Reisenden und offizielle Verkündigungen enthielt, wandelte sich der Inhalt ab dem letzten Viertel des 19. Jahrhunderts. Ab diesem Zeitpunkt überwogen Nachrichten und Berichte, während Anzeigen zwar weiterhin existierten, nun aber zunehmend in den Hintergrund rückten. Im Jahr 1941 wurde das Darmstädter Tagblatt durch eine Anordnung der Reichspressekammer eingestellt und erschien nach Ende des Zweiten Weltkriegs ab 1949 wieder. Neben dem Inhalt wandelte sich auch das Format mehrmals und nach dem Krieg fand ebenso ein Schriftwechsel von Fraktur auf Antiqua statt. Insgesamt umfasst das Tagblattprojekt 598 Bände bzw. 600.000 Seiten.

Der Gegenstand des Digitalisierungsprojekts ist daher keineswegs homogen und stellt den Digitalisierungsworkflow des Projekts vor zahlreiche Herausforderungen. Offensichtlich ist zunächst die schiere Menge an Daten, die während des Digitalisierungsworkflows gespeichert, verarbeitet und bereitgestellt werden muss. Allein dies macht das Tagblatt zu einem der aufwändigsten Projekte, das am Zentrum für Digitale Editionen (ZEiD) an der ULB gehostet wird. Weiterhin hat die Heterogenität des Materials Auswirkungen auf den kompletten Digitalisierungsworkflow, etwa beim Scannen oder bei der OCR- und Layouterkennung. So sind beispielsweise Umfang (Zahl der Ausgaben und damit auch der Seitenzahlen), Format und Anzahl bei den einzelnen Bänden, in denen das Tagblatt im Bestand der ULB überliefert wurde, höchst unterschiedlich. Gleiches gilt für die einzelnen Ausgaben. Während das Tagblatt beispielsweise im 18. Jahrhundert einmal wöchentlich mit einem Umfang von bis zu sechs Seiten erschien, wurde es ab dem Ende des 19. Jahrhunderts als Tageszeitung herausgegeben. Je nach Umfang der eigentlichen Zeitungsausgabe sowie etwaigen Beilagen schwankt hier die Seitenzahl zwischen 8 und 36 Seiten. Mit Hilfe der Onlinepräsentation lassen sich zudem die für den eigenen Forschungsschwerpunkt relevanten Details zum Umfang im Sinne von Ausgaben und Seitenzahlen schnell erörtern. Der Arbeitsaufwand bei der Digitalisierung ist also durchaus vergleichbar mit der Digitalisierung mehrerer Zeitungen.

Neben der technischen Umsetzung stellen auch rechtliche Aspekte das Projekt vor große Herausforderungen, da die Ausgaben der zweiten Projektphase (in Ausnahmen sogar noch Ausgaben aus der ersten Phase) dem Urheberrecht unterliegen. Das Projekt konnte in dieser Form überhaupt nur zustande kommen, weil die Eigentümerin des Darmstädter Tagblatts, die Echo Medien GmbH, als Projektpartnerin gewonnen werden konnte und dem Projekt eingeräumt hat, alle Inhalte, für die sie die Rechte besitzt, veröffentlichen zu dürfen. Dennoch ist damit die Rechteklärung noch nicht abgeschlossen. Grundsätzlich ist jeder Beitrag, beispielsweise Berichte, Bilder, Zeichnungen oder Fortsetzungsromane, um nur eine Auswahl zu nennen, als eigenständiges Werk mit eigenem/eigener Urheber:in anzusehen, für den jeweils die Rechtslage mit der/dem betroffenen Verlag, Agentur, Fotograf:in etc. geklärt werden muss.

3. Digitalisierung und Bereitstellung

3.1 Technische Umsetzung

In beiden Projektphasen wurden im Digitalisierungszentrum der ULB hochqualitative Scans angefertigt. Während in der ersten Phase die OCR- und Layouterkennung im ZEiD mittels der Plattform Transkribus (www.transkribus.org/de, Stand: 27.11.2025) anhand eigens trainierter Modelle selbst durchgeführt wurde, wurde in der Projektplanung für die zweite Projektphase zwecks besserer Planbarkeit entschieden, einen Dienstleister mit der automatischen Erstellung der Volltexte zu beauftragen. In jedem Fall steht am Ende die Bereitstellung der durchsuchbaren Volltexte in TEI-XML bzw. der Digitalisate über verschiedene Plattformen, etwa TUdigit (ULB), WDB+ (ULB/ZEiD) und CQPweb (Digitale Linguistik/TU).[4]

3.2 Rechteklärung Phase 1 und Phase 2

Die Vorarbeiten in der ersten Projektphase (1740–1941) haben bestätigt, dass selbst bei einer bis zu den ermittelbaren Einzelurheber:innen reichenden Rechteeinräumung die Möglichkeit verbleibt, dass einzelne Beiträge oder Abbildungen nachlizenziert oder auf Unterlassungsaufforderung wieder aus dem Netz genommen werden müssen bzw. lediglich in den Räumen der Bibliothek gezeigt werden dürfen. Diese Unschärfe ist nicht heilbar, da das Urheberrecht grundsätzlich bei der Person verbleibt, die das Werk geschaffen hat. Wenn man also nicht gänzlich auf die Digitalisierung und Publikation von Zeitungen jüngeren Datums verzichten will, muss trotz Rechteeinräumung von Verlagen ein gewisses, wenn auch geringes Restrisiko in Kauf genommen werden.

Da die Digitalisierung und weltweite Bereitstellung im Internet als Vervielfältigung und Wiederveröffentlichung in neuer Form zu verstehen ist, hat die ULB den rechtlichen Rahmen des Projekts geprüft. Das Tagblatt ist sowohl ein vergriffenes als auch ein verwaistes Werk im Sinne des Urheberrechtsgesetzes (UrhG).[5] Für die Veröffentlichung von vergriffenen Werken ist eine Lizenzvereinbarung zu treffen, für die Publikationsform „Zeitung“ erklärten sich die typischen Lizenzierungspartner wie die VG Wort zu Projektbeginn explizit jedoch als nicht zuständig. Der „Lizenzierungsservice Vergriffene Werke“ der Deutschen Nationalbibliothek (VW-LiS) war zum Zeitpunkt der Prüfung bzw. seit 7.6.2021 für unbestimmte Zeit ausgesetzt.[6] Da die Digitalisate des Tagblatts von Forschenden aber zeitnah benötigt wurden bzw. werden, hatte sich die ULB entschieden, das Tagblatt weiterhin in seiner Eigenschaft als „verwaistes Werk“ zu behandeln.

4 Die verschiedenen Plattformen werden über die Projektwebseite verlinkt. www.ulb.tu-darmstadt.de/forschen_publizieren/forschen/darmstaedter_tagblatt.de.jsp (Stand: 27.11.2025).

5 Das Tagblatt wurde 1986 eingestellt. Rechte an der Marke und am Archiv wurden von der Konkurrenz-Zeitung Darmstädter Echo übernommen. Trotz gezielter Prüfung ist es rechtlich nicht belegt, dass das Darmstädter Echo der vollumfängliche Rechtsnachfolger des Darmstädter Tagblatts ist. Das Darmstädter Tagblatt musste wegen Unterschlagungsdelikten der damaligen Geschäftsführung eingestellt werden (www.spiegel.de/politik/vom-schlag-getroffen-a-41637d60-0002-0001-0000-000013520424, Stand: 27.11.2025). In diesem Zusammenhang sind mündlicher Aussage des Echo-Managements zufolge viele Geschäftsdokumente vernichtet worden, und es lassen sich keine Arbeitsverträge damaliger fester oder freier Mitarbeiter:innen nachweisen.

6 Der „Lizenzierungsservice Vergriffene Werke“ steht mittlerweile wieder zur Verfügung: www.dnb.de/DE/Professionell/Services/VW-LiS/vwlis_node.html (Stand: 28.1.2026).

Für „verwaiste Werke“ sind „besondere gesetzlich erlaubte Nutzungen“ vorgesehen, wenn diese „Bestandsinhalte von Sammlungen öffentlich zugänglicher Bibliotheken sind, […] bereits veröffentlicht (wurden) und deren Rechtsinhaber auch durch eine sorgfältige Suche nicht festgestellt oder ausfindig gemacht werden konnten“ (UrhG §61,2). Die Grundvoraussetzung für die Vervielfältigung und öffentliche Zugänglichmachung in UrhG §61,5 ist insofern erfüllt: Die ULB handelt in Erfüllung ihrer im Gemeinwohl liegenden Aufgaben, indem sie Bestandsinhalte bewahrt und auf Dauer zugänglich macht, um kulturellen und bildungspolitischen Zwecken zu dienen.

Die ULB hat in der ersten Projektphase umfangreiche Vorarbeiten geleistet und zunächst die Expertise eines Fachanwalts für Urheber- und Medienrecht eingeholt. Dieser sieht in der o. g. „sorgfältigen Suche“ den Dreh- und Angelpunkt der Bestimmungen zu den verwaisten Werken. Nur wer die Suche in nachvollziehbarer Weise durchführe, dokumentiere und archiviere, könne sich auf die Privilegierung der Vorschriften des §61 ff. UrhG berufen. Jeder Bestandsinhalt müsse gesucht werden und zwar vor der Veröffentlichung im Internet. Dort, wo kein Name genannt ist oder wo einem Kürzel kein Klarname zugeordnet werden kann, seien die Rechtsverhältnisse an einem Werk nicht nachvollziehbar.

Es scheint auf dieser Grundlage sinnvoll, die Inhalte so zu kategorisieren und zu priorisieren, wie sie im Fall einer Rechtsstreitigkeit ins Gewicht fallen würden. In die oberste Kategorie fällt das Agentur- und Verlagsmaterial (heute noch aktiver Firmen), das im Tagblatt in großer Menge abgedruckt war. Die zweite Kategorie bilden Fotograf:innen und die dritte Gruppe die Textautor:innen. Durch die Vorarbeiten konnten ca. 40% der in den Jahrgängen der Phase II enthaltenen Werke kostenfrei lizenziert werden, vor allem durch die erzielten Vereinbarungen mit den größten Agenturen.

Um auch für den Rest der Werke möglichst vollständig Lizenzabkommen abschließen zu können, wurden bereits in der ersten Phase Listen mit Kürzeln (bestehend aus 1–4 Buchstaben) und Namen der Urheber:innen angelegt. Da die Vertragshistorie des Darmstädter Tagblatts unvollständig ist, sind diese Kürzel und Namen nicht anhand verfügbarer Dokumente auflösbar. Wegen des starken Regionalbezugs der Zeitung konnte die gezielte Internetsuche (Google) in vielen Fällen erste Ergebnisse wie Traueranzeigen Verstorbener oder aktuelle Anschriften liefern. Zudem wurde in Stadtlexika, in der Deutschen Nationalbibliothek (DNB), in HeBIS (Hessisches BibliotheksInformationsSystem) und im jüngsten Darmstädter Adressbuch (2002) gesucht. Wenn mehrere Namensträger in Frage kamen, wurden alle Adressen mit in die Versandliste aufgenommen. Auf Anfrage teilte das Einwohnermeldeamt mit, dass für eine Adressermittlung mindestens drei Datenpunkte wie vollständiger Name, letzte Adresse und Geburtsdatum benötigt werden. Da nur die (Nach-) Namen bekannt sind, ist dieser Weg nicht gangbar. Ähnlich aussichtslos ist ein Ersuchen an das Nachlassgericht zur Ermittlung von Erben. Die Suche wurde detailliert dokumentiert (inklusive Suchdatum, Lebensdaten, Adresse, Notizen zur Art der Tagblatt-Mitarbeit). An alle ermittelten Adressen wurde ein Informationsbrief versendet, der den Urheber:innen das Digitalisierungsprojekt vorstellt und sie um eine kostenlose, einfache, weltweite und zeitlich unbeschränkte Lizenz für ihre im Tagblatt abgedruckten Werke bittet.

In der derzeit laufenden zweiten Phase werden mit fortschreitender Digitalisierung des Tagblatts technische Ansätze für eine (teil-)automatisierte Urhebersuche erprobt und durchgeführt. Diese zielen darauf ab, Kürzel und Namen von Urheber:innen automatisch zu finden, grob zu klassifizieren, zu extrahieren und in einer Datenbank zu speichern. Das Vorgehen ist dabei zweigleisig.

Zum einen erarbeiten die Mitarbeiter:innen des Projekts mögliche Kategorien und reguläre Ausdrücke auf der Grundlage der Lektüre ausgewählter Ausgaben. Die Kategorien dienen

dabei unter anderem der Priorisierung und orientieren sich an den oben genannten Beispielen. Die folgenden Ausführungen dienen der Illustration, sind daher nicht vollständig. Als mögliche Urheber:innen bzw. Lizenzgeber:innen kommen also hauptsächlich Agenturen, Autor:innen, Fotograf:innen und Verlage vor (vgl. Abb. 2). Für die automatische Urhebererkennung ist darüber hinaus noch wichtig zu vermerken, wo im Artikel die Nennung stattfindet und vor allem, wie diese Stellen in der XML-Struktur des Volltexts möglichst eindeutig erreichbar sind. Mögliche Stellen sind z. B. in einem Untertitel oder als vollständiger Untertitel. Dies trifft vor allem auf Nennungen eines vollständigen Namens zu, die oft, aber nicht immer mit „Von“ eingeleitet werden. Sind sie Teil eines Untertitels, stehen sie nach dem Untertitel und sind oft mit einem Schrägstrich vom Untertitel abgetrennt. Autorenkürzel können am Anfang oder am Ende eines Artikels stehen. In der Regel sind sie vor bzw. nach dem Text zu finden und können mit einem Punkt abgeschlossen werden oder Bindestriche enthalten. Agenturkürzel stehen in der Regel nach einem Ortsnamen in der ersten Zeile in runden Klammern. Es können mehrere Agenturen in einer Klammer genannt werden, die dann normalerweise mit Schrägstrichen voneinander abgetrennt sind.[7]

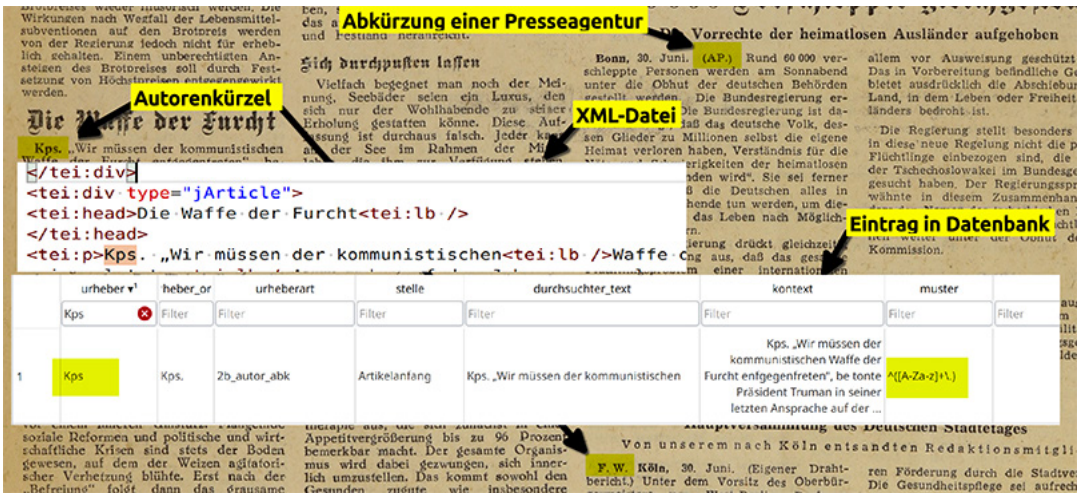

Abb. 2: Verschiedene Formen der Urhebernennung und Extraktionsergebnis

Zum anderen arbeiten zwei Mitarbeiter:innen des Projektes daran, einen Goldstandard der Urhebernennungen auf der Grundlage der bisher gescannten Jahrgänge (1950–1970) zu erarbeiten.[8] In 18 Ausgaben (ca. 4.500 Artikel[9]) annotieren sie manuell alle Urhebernennungen mit den oben beschriebenen Kategorien und vermerken, ob auf eine Agentur, eine:n Autor:in etc. mit Kürzel oder vollständigem Namen am Anfang oder am Ende eines Artikels verwiesen wird. Während der Annotationsarbeiten können neue Kategorien eingefügt werden, die dann von beiden Annotierenden verwendet werden. Nach Abschluss der Annotationen werden Zweifelsfälle vereinheitlicht und die Daten dienen als Referenz für die automatisiert erkannten und extrahierten Stellen. Sobald die Extraktionen per Skript eine akzeptable Trefferquote (angestrebt sind 96%) erreicht haben, werden sie produktiv zur Extraktion eingesetzt und dabei kontinuierlich geprüft und verbessert.

7 Siehe Abbildung 2.

8 Hierfür wird die Software INCEpTION verwendet, mit der beliebige Annotationskategorien erstellt und zugewiesen werden können (vgl. Klie et al. 2018 und https://inception-project.github.io/).

9 In den zu annotierenden Texten wurden 4.224 Texteinheiten von der automatischen Layouterkennung als Artikel und 4.789 als Überschrift ausgezeichnet. Aufgrund des Seitenlayouts werden manche Texteinheiten jedoch fälschlich als Teil eines Artikels erkannt, obwohl es sich inhaltlich um mehrere Artikel handelt. Vorläufige Ergebnisse zeigen ca. 100 Urhebernennungen in einer Ausgabe von 1950 bei 175 als Artikel ausgezeichneten Textheinheiten (184 Überschriften).

Die Einträge in der Datenbank werden dann mit den bereits bekannten Kürzeln und Namen abgeglichen. Noch nicht bekannte Kürzel werden nach Möglichkeit aufgelöst, indem die Impressen und noch vorliegenden Dokumentationen eingesehen werden. Für die aufgelösten Kürzel und eventuell neue vollständige Namen führen die Projektmitarbeiter:innen den oben beschriebenen Prozess durch.

3.3 Linguistische Aufbereitung

Nach der Buchstabenerkennung wurden die resultierenden Texte in einzelne Wortformen zerlegt (Tokenisierung) und für jede Wortform wurden die Grundform (Lemmatisierung) und die Wortart (Part-of-Speech-Tagging) annotiert. Darüber hinaus wurden sogenannte Named Entities, also z. B. Eigennamen und Namen von Institutionen und morphologische Informationen, auf Tokenebene ermittelt. Der Text wurde in Sätze zerlegt und jeder Satz wurde auf seine Struktur hin analysiert (Dependency Parsing).

Die folgenden Werkzeuge kamen für die oben genannten Aufgaben zum Einsatz, wobei die Tagging-Ergebnisse ohne weitere Korrekturen parallel genutzt werden:

- Tokenisierung und Satzsegmentierung: SoMaJo (Proisl/Uhrig 2016)
- Lemmatisierung: TreeTagger (Schmid 1995)
- Part-of-Speech-Tagging: spaCy (de_dep_news_trf; Honnibal et al. 2021) und TreeTagger
- NER: spaCy (de_core_news_lg), Flair (dbmdz/flair-clef-hipe-german-base; Schweter/März 2020)
- Morphologische Informationen und Dependency Parsing: Stanford Stanza (Qi et al. 2020)

3.4 Bereitstellung und Nutzungsmöglichkeiten

Die Daten werden auf unterschiedlichen Wegen und mit unterschiedlichen Nutzungsschwerpunkten zugänglich gemacht.

Die linguistisch aufbereiteten Daten wurden in die Corpus Workbench (CWB) (vgl. Evert/Hardie 2011) importiert und sind über die Plattform DiscourseLab zugänglich.[10] Der Zugang erfolgt über CQPweb (vgl. Hardie 2012), das sowohl eine browserbasierte Nutzeroberfläche als auch Analysefunktionen für Korpora, die in der CWB verwaltet werden, bereitstellt. Eine Veröffentlichung der Faksimiles erfolgt ebenso über die Plattform TUdigit der ULB Darmstadt, die digitale Sammlungen wie die des Darmstädter Tagblatts bereitstellt. Die ebenfalls an der ULB Darmstadt entwickelte Plattform WDB+ ermöglicht neben der Anzeige von Digitalisaten auch textbasierte Recherchen, da zusätzlich digitale Volltexte im TEI-XML-Format bereitgestellt werden.[11] Durch die Bereitstellung dieser Volltexte sind weiterführende linguistische Analysen durch Forschende möglich, etwa zur diachronen Sprachentwicklung, zur Untersuchung lexikalischer oder morphosyntaktischer Muster sowie zur Erforschung sprachlicher Variation.

10 Der Zugang findet sich auf www.discourselab.de/cqpweb/ (Stand: 27.11.2025). Für die Nutzung ist eine Registrierung notwendig.

11 Die jeweils aktuellen Links zu den Plattformen sind auf der Projektwebseite aufgelistet: www.ulb.tu-darmstadt.de/forschen_publizieren/forschen/darmstaedter_tagblatt.de.jsp (Stand: 27.11.2025).

4. Beispiele für Anwendungsmöglichkeiten

4.1 Nominalphrasen zeigen den Wandel vom Anzeigenblatt zur Zeitung

Wie oben in der Beschreibung des Darmstädter Tagblatts dargestellt, entwickelte sich dieses von einem Anzeigenblatt zu einer regionalen Tageszeitung mit wachsendem redaktionellen Teil und verschiedenen Ressorts. Wie in Regionalzeitungen üblich, erschienen weiterhin Klein- und Werbeanzeigen. Es ist zu erwarten, dass die zusätzlichen Themen und Inhalte sich auf das Vokabular auswirken und diese Auswirkungen durch Abfragen des CQPweb-Korpus zu belegen sind.

Listen der häufigsten Substantive sind hierfür nur bedingt geeignet. Hilfreicher präsentieren sich die Ergebnisse, wenn statt einzelner Substantive ganze Nominalphrasen (NP) untersucht werden. In die Abfrage werden die minimal erweiterte NP aus Artikel gefolgt von einer beliebigen Anzahl an attributiv gebrauchten Adjektiven gefolgt von einem Nomen aufgenommen. Die folgende Suchanfrage findet beides: „[pos="ART"] [pos="ADJA"]+ [pos="NN"]". Der Ausdruck „[pos="ADJA"]" findet alle als attributives Adjektiv getaggten Wörter und das Pluszeichen dahinter steht für „einmal bis beliebig oft".

Nr.	1750	1800	1850	1880	1900
1	der armen Seele	dem hiesigen Rathhaus	ein kleines Logis	eine ruhige Familie	ein möbl. Zimmer
2	Ein Maaß-junges[12] Bier	eine ledige Person	Ein vollständiges Logis	der mittlere Stock	eine schöne 3=Zimmer-wohnung
3	das gantze Jahr	der langen Gasse	den evangelischen Gemeinden	ein möbl. Zimmer	eine kleine Wohnung
4	ein ungenannter Freund	der neuen Vorstadt	Ein freundliches Logis	ein schönes Logis	der Technischen Hochschule
5	der Herrnhutischen Secte	der großen Ochsengasse	der mittlere Stock	eine schöne Wohnung	eine schöne 4=Zimmer-wohnung

Tab. 1: Die häufigsten Nominalphrasen in Fünf-Jahres-Abschnitten ab dem im Spaltenkopf vermerkten Jahr. 1750–1900

Wie die Einträge in Tabelle 1 zeigen, sind in den ausgewählten Zeiträumen keine offensichtlichen Hinweise auf nachrichtliche oder redaktionelle Inhalte enthalten. Die meisten der aufgeführten Nominalphrasen zeigen stattdessen deutlich den Anzeigencharakter. Stichproben belegen dies. So ist z. B. *der armen Seele* Teil eines Buchtitels, der in einer Verkaufs-

12 „Maaßjunges" ist ein OCR-Fehler, der auf sehr eng gedruckte Tabellen zurückgeht. Richtig heißt es „Ein Maaß junges Bier".

anzeige genannt wird: „Betrachtung über die Bräutigams=Liebe des HErrn IEsu zu der armen Seele“ (vgl. z. B. Darmstädter Tagblatt vom 4.1.1753, S. 2, https://exist.ulb.tu-darmstadt.de/3/v/pa000059-0001[13]). Dasselbe gilt für den Eintrag der *Herrnhutischen Secte*, dieser ist Teil des Titels „Die Schule der Verständigen, oder das auf Sinnlichkeit gegründete Atheistische Lehr=Gebäude, der heutiges Tages in der Herrnhutischen Secte wieder auflebenden sogenannten Homiuum Intelligentiæ, in einem Gespräch zwischen Cautorius und Berrhoisius ausfindig gemacht und ans Licht gestellt, von Johann Christoph Venator, zweiten Evangelis. Predigern in Friedberg.“ (vgl. z. B. Darmstädter Tagblatt vom 30.4.1750, S. 3, https://exist.ulb.tu-darmstadt.de/3/v/pa000059-0001, Stand: 27.11.2025). Keiner Verkaufsanzeige, sondern einer Art Werbeanzeige gehört *ein ungenannter Freund* an: „AVERTISSEMENT Bei dem HErrn finden unsere Waysen Hülfe , und Errettung aus aller ihrer Noth. Daher schickte ein ungenannter Freund den 20.ten Martii 1.fl. 15 Alb. mit Verlangen vor ihn in seiner Krankheit zu beten“ (vgl. z. B. Darmstädter Tagblatt vom 26.3.1750, Titelseite, https://exist.ulb.tu-darmstadt.de/3/v/pa000056-0013, Stand: 27.11.2025).

Wenn auch die NP *dem hiesigen Rathhaus* ein Hinweis auf Berichterstattung über regionale Politik sein könnte, so zeigen die Konkordanzen doch recht deutlich, dass es sich auch dabei um Teile von Anzeigen handelt, wie in Tabelle 2 gezeigt wird:

Context before	Query item	Context after	year
Mittwoch den 29ten die= ses Nachmittags um 2 Uhr auf	dem hiesigen Rathhaus	unter denen in dem Termin bekannt zu machenden Bedingungen versteigt	1803[14]
den 13ten kunftigen Monats Ottober Nachmittags Um 2 Uhr auf	dem hiesigen Rathhaus	nochmalen versteigt und dem Meistbietenden unwiederruflich zugeschlagen werden. Darmstadt	1802[15]
künftigen Monats August , Nachmittags 2 Uhr , fen auf	dem hiesigen Rathhaus	der dem Feldwebel Fenner im 2ten Bat . Leibregiments gehörige	1801[16]
Monats Februarii , Nachmittags um 2 Uhr , sollen auf	dem hiesigen Rathhaus	nachbeschriebene zur Verlassenschaft des Burger und Metzgermeister Ludwig Haxt gehörige	1802[17]
Erben und Christian Engel, vorbehaltlich der Ratification , auf	dem hiesigen Rathhaus	öffentlich versteigt werden. Darmstadt den 28ten März 1800 .	1800[18]

Tab. 2: Konkordanzen zu „dem hiesigen Rathhaus“ 1800–1805 (zufällige Auswahl)

13 Die digitale Edition des Darmstädter Tagblatts befindet sich noch im Aufbau. Die hier genannten Links auf https://exist.ulb.tu-darmstadt.de (Stand: 27.11.2025) verweisen auf eine Vorschau und sind nicht permanent, also nicht zitierfähig.

14 Vgl. https://exist.ulb.tu-darmstadt.de/3/v/pa000109-0024.

15 Vgl. https://exist.ulb.tu-darmstadt.de/3/v/pa000108-0040.

16 Vgl. https://exist.ulb.tu-darmstadt.de/3/v/pa000107-0031.

17 Vgl. https://exist.ulb.tu-darmstadt.de/3/v/pa000108-0004.

18 Vgl. https://exist.ulb.tu-darmstadt.de/3/v/pa000106-0013.

Die Einträge in Tabelle 3 ab 1910 zeigen jedoch deutlich, dass mehr nachrichtliche Texte abgedruckt werden.

Nr.	1910	1920	1925	1930	1935
1	die höchsten Preise	Die deutsche Regierung	der Deutschen Volkspartei	des deutschen Volkes	Die 22 mm
2	der andern Behörden	der Deutschen Volkspartei	den letzten Tagen	Das deutsche Volk	des deutschen Volkes
3	die trauernden Hinterbliebenen	das deutsche Volk	der letzten Zeit	den letzten Jahren	Das deutsche Volk
4	den politischen Teil	der deutschen Regierung	den letzten Jahren	den letzten Tagen	den letzten Jahren
5	den redaktio-nellen Teil	des deutschen Volkes	der anderen Seite	der letzten Zeit	den letzten Tagen

Tab. 3: Die häufigsten Nominalphrasen in Fünf-Jahres-Abschnitten ab dem im Spaltenkopf vermerkten Jahr. 1910–1935

Die Phrasen „den redaktionellen Teil“ bzw. „den politischen Teil“ in den fünf Jahrgängen ab 1910 referieren direkt auf diesen Wandel, während in den weiteren Fünf-Jahres-Ausschnitten Phrasen wie „das deutsche Volk“, „des deutschen Volkes“, „die deutsche Regierung“ etc. ihn auf inhaltlicher Ebene zeigen.

4.2 Kein Datenschutz im 18. Jahrhundert: Reisende nach Darmstadt

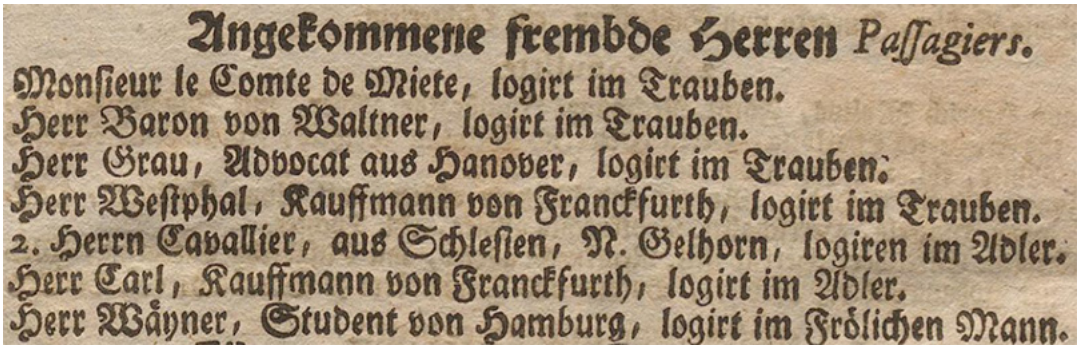
Angekommene frembde Herren Paſſagiers.
Monſieur le Comte de Miete, logirt im Trauben.
Herr Baron von Waltner, logirt im Trauben.
Herr Grau, Advocat aus Hanover, logirt im Trauben.
Herr Weſtphal, Kauffmann von Franckfurth, logirt im Trauben.
2. Herrn Cavallier, aus Schleſien, N. Gelhorn, logiren im Adler.
Herr Carl, Kauffmann von Franckfurth, logirt im Adler.
Herr Wäyner, Student von Hamburg, logirt im Frölichen Mann.

Abb. 3: Liste angekommener Besucher. Darmstädter Tagblatt 15.1.1750, S. 3 unten

Ein Blick in vergangene Ausgaben bringt nicht nur erwartbare Ergebnisse wie z. B. Verkaufsangebote und amtliche Bekanntmachungen. Das Darmstädter Tagblatt veröffentlichte in den frühen Ausgaben bis in das 19. Jahrhundert die Namen, Berufe und Unterkünfte von Reisenden, die nach Darmstadt kamen (vgl. Abb. 3). Eine Auswertung dieser Daten ermöglicht es, einen Eindruck vom Einzugsgebiet Darmstadts im 18. Jahrhundert zu bekommen, wie in Stegmeier/Müller (2024) näher ausgeführt wird. Nach einer semi-automatischen Aufbereitung der Passagierdaten wurden diese als CSV-Dateien gespeichert und über den Geo-Browser (vgl. Kollatz 2016) von DARIAH-DE (https://geobrowser.de.dariah.eu/, Stand: 27.11.2025) verfügbar gemacht. Hierfür wurden die Ortsnamen nor-

malisiert. Im Geo-Browser selbst wurden den Einträgen dann anhand der Ortsnamen Längen- und Breitengrade zugeordnet. In Zweifelsfällen wurde der Darmstadt am nächsten liegende Ort gewählt. Alle so erstellten Datenblätter können über einen persistenten Link von DARIAH-DE heruntergeladen werden.[19] Die folgende Abbildung 4 vergleicht die Ergebnisse für drei Zehn-Jahres-Schritte:

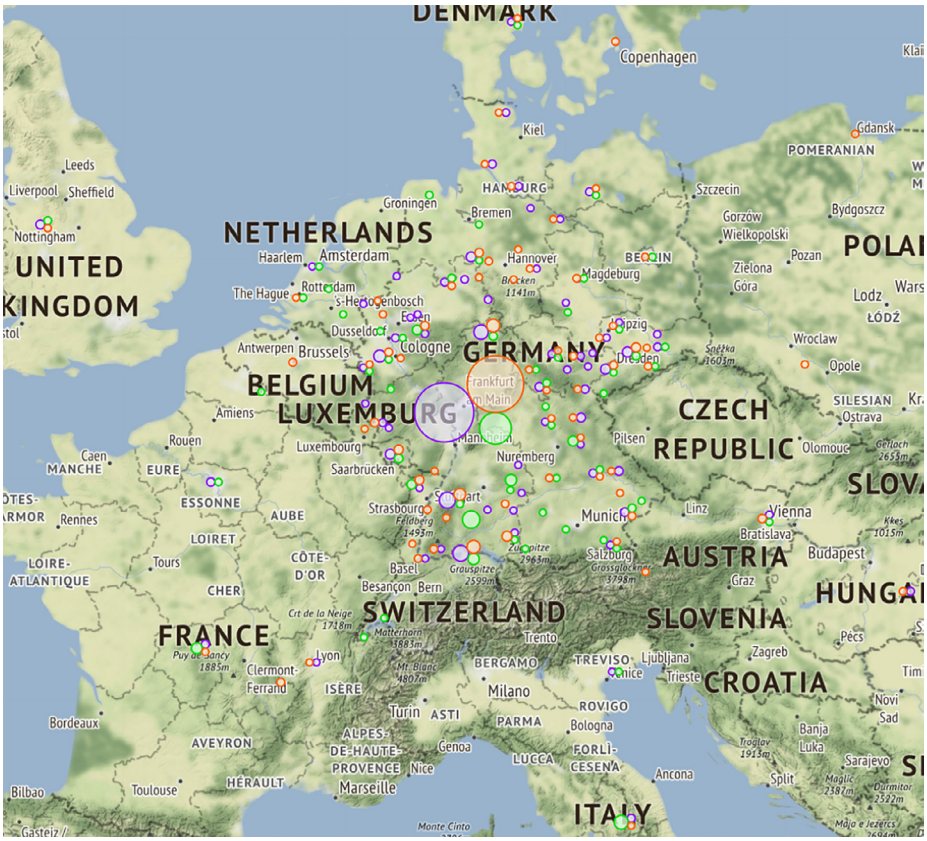

Abb. 4: Zehn-Jahres-Schritte im Vergleich: 1745–1754 (orangefarben), 1755–1764 (violett), 1785–1794 (grün)

Wenig überraschend kommen die meisten Besucher aus der Umgebung. Das Einzugsgebiet insgesamt ist jedoch angesichts der eher beschwerlichen Reisen in der zweiten Hälfte des 18. Jahrhunderts beeindruckend: Vom heutigen Frankreich, England, Italien, Österreich, Ungarn, Polen und Dänemark kamen Besucher nach Darmstadt. Den Berufsbezeichnungen nach waren die meisten davon Kaufleute, aber auch Geistliche und Kulturschaffende zählten zu den Gästen in Darmstadt. Welchem Zweck genau diese Besuche dienten, wird nicht immer vermerkt. Es gibt jedoch Fälle, in denen eindeutig auf den geschäftlichen Charakter des Aufenthalts verwiesen wird:

> Mercier, Kaufmann aus Frankreich, ist hier mit einem Sortiment von ver=schiednen Waaren angekommen, und bestehet solches in Pariser Parfümerien, Confituren und feinen Liqueurs, Dragees von Verdün, Punsch=Cappillair= und Mandelmilch=Sirop, Himbeeren=Essig, Englischen Tabletten von Menthe, Pastillen a la Mamboise von verschiedenem Geschmack, feinem Pariser Senft, Italieni=schen Blumen und mehreren dergleichen. Er wird sich 8 Tage dahier im Gasthaus zum Engel aufhalten, wo man diese Waaren in seinem Zimmer, eine Stiege hoch No. 3, aufgestellt findet. (Darmstädter Tagblatt, 21.3.1791)

19 Beispiel: Die Daten für den Zeitraum 1745–1750 finden sich hier: https://cdstar.de.dariah.eu/dariah/EAEA0-CDE2-C0B0-1BED-0 (Stand: 27.11.2025). Eine vollständige Liste der Links ist in Stegmeier/Müller (2024, S. 75) verfügbar.

5. Resümee

Im vorliegenden Beitrag haben wir das Editionsprojekt des Darmstädter Tagblatts (1740–1986) vorgestellt, das an der Universitäts- und Landesbibliothek Darmstadt (ULB) und der Technischen Universität Darmstadt seit 2019 in einem DFG-geförderten Vorhaben realisiert wird. Wir haben über die Aufbereitung des Datensatzes als Korpus berichtet, die OCR, Layoutanalyse, POS-Tagging, Lemmatisierung, Named Entity recognition und die Erfassung morphologischer Informationen umfasst. Ein besonderer Fokus liegt auf der halbautomatisierten Rechteklärung für Einzelbeiträge. Dabei handelt es sich um einen rechtlich wie technisch komplexen Aspekt, der neue Ideen und Verfahren zur Urhebersuche erfordert.

Das dargestellte Projekt will einen Baustein zur wachsenden Sammlung linguistisch aufbereiteter Ressourcen zur Sprachgeschichte des Neuhochdeutschen bieten, die der empirischen Sprachgeschichtsschreibung zahlreiche neue Perspektiven bietet. Durch die Kombination linguistischer Annotation mit detaillierten Metadaten sind diachrone Analysen über Jahrhunderte hinweg möglich. Die Untersuchung lexikalischer oder morphosyntaktischer Muster – etwa der Wandel von Nominalphrasen im Zuge der Transformation des Tagblatts von einem Anzeigenblatt zur Tageszeitung – bietet exemplarische Einblicke in Medien-, Kultur- und Sprachgeschichte. Die Erfassung von Herkunftsorten reisender Personen im 18. Jahrhundert erlaubt raumzeitliche Rekonstruktionen von Mobilität, Reichweite und Kommunikationsräumen – ein bislang kaum erschlossener Zugang zur historischen Soziolinguistik. Insgesamt bietet das Projekt ein Beispiel dafür, wie digitale Editions- und Korpusarbeit mit linguistischer Analyse, rechtlicher Reflexion und kulturwissenschaftlicher Interpretation verknüpft werden kann.

Das Korpus des Darmstädter Tagblatts eignet sich aufgrund des langen Zeitraums, den es kontinuierlich abdeckt, sehr gut für die Untersuchung semantischen und morphosyntaktischen Wandels, wie Müller am Beispiel von *Risiko* gezeigt hat: Graf/Müller (2025) nutzen das Darmstädter Tagblatt als Quelle für eine begriffsgeschichtliche Darstellung von *Risiko.* Müller (2025) untersucht u. a. am Darmstädter Tagblatt die Dynamik der Morphosyntax von *Risiko* und *Gefahr.* Anhand empirischer Korpusdaten wird dort die Beziehung zwischen Flexion, Wortbildung, Phrasensyntax und Diskursfunktion beim Bedeutungswandel analysiert.

Literatur

Evert, Stefan/Hardie, Andrew (2011): Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. Birmingham: University of Birmingham. www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf (Stand: 26.11.2025).

Graf, Rüdiger/Müller, Marcus (2025): Risiko. In: Müller, Ernst/Picht, Barbara/Schmieder, Falko (Hg.): Das 20. Jahrhundert in Grundbegriffen. Lexikon zur historischen Semantik in Deutschland. Basel/Berlin: Schwabe. https://doi.org/10.31267/Grundbegriffe_91070823.

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

Honnibal, Matthew/Montani, Ines/Boyd, Adriane/Van Landeghem, Sofie/Peters, Henning (2020): spaCy: Industrial-strength natural language processing in Python. https://explosion.ai/blog/spacy-v3 (Stand: 26.11.2025). [Modell: https://spacy.io/models/de#de_dep_news_trf (Stand: 26.11.2025)].

Klie, Jan-Christoph/Bugert, Michael/Boullosa, Beto/de Castilho, Richard E./Gurevych, Iryna (2018): The INCEpTION Platform: Machine-assisted and knowledge-oriented interactive annotation. In:

Zhao, Dongyan (Hg.): Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Santa Fe, NM: Association for Computational Linguistics, S. 5–9.

Kollatz, Thomas (2016): Raum-Zeit-Analysen mit Geo-Browser und Datasheet-Editor. In: Bibliothek Forschung und Praxis 40, 2, S. 229–233. DOI: 10.1515/bfp-2016-0032.

Müller, Marcus (2025): ‚Morphologie aus Sicht der Diskursgrammatik'. Am Beispiel der zeithistorischen Morphosyntax von Risiko. In: Michel, Sascha (Hg.): Diskursmorphologie: Ansätze und Fallstudien zur Schnittstelle zwischen Morphologie und Diskurslinguistik. (= Diskursmuster/Discourse Patterns 36). Berlin/Boston: De Gruyter, S. 55–84. https://doi.org/10.1515/9783111363936-003.

Proisl, Thomas/Uhrig, Peter (2016): SoMaJo: State-of-the-art tokenization for German web and social media texts. In: Cook, Paul/Evert, Stefan/Schäfer, Roland/Stemle, Egon (Hg.): Proceedings of the 10th Web as Corpus Workshop. Berlin: Association for Computational Linguistics, S. 57–62. https://aclanthology.org/W16-2607/ (Stand: 26.11.2025).

Qi, Peng/Zhang, Yuhao/Zhang, Yuhui/Bolton, Jason/Manning, Christopher D. (2020): Stanza: A Python natural language processing toolkit for many human languages. https://nlp.stanford.edu/pubs/qi2020stanza.pdf (Stand: 26.11.2025).

Schmid, Helmut (1995): Improvements in part-of-speech tagging with an application to German. www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Stand: 26.11.2025).

Schweter, Stefan/März, Luisa (2020): Triple E - Effective Ensembling of Embeddings and Language Models for NER of Historical German. In: Cappellato, Linda/Eickhoff, Carsten/Ferro, Nicola/Névéol, Aurélie (Hg.): Proceedings of the Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum. Thessaloniki, Greece, September 22–25, 2020. Padua: Universität Padua. https://ceur-ws.org/Vol-2696/#paper_173 (Stand: 26.11.2025).

Stegmeier, Jörn/Müller, Marcus (2024): „Angekommene frembde Herren Passagiers": Besucher in Darmstadt in drei Dekaden des 18. Jhs.: Ein Werkstattbericht. In: Bartsch, Sabine/Borek, Luise/Hegel, Philipp (Hg.): Computer im Musenhain. Von träumenden Büchern und der Aura des Digitalen. (= Festschriften der Technischen Universität Darmstadt 1.2023). Darmstadt: Technische Universität Darmstadt. DOI: 10.26083/tuprints-00027908.

Stegmeier, Jörn/Günther, Anne-Christine/Hammer, Angela/Müller, Marcus/Stäcker, Thomas (2022): Eine Zeitung in drei Jahrhunderten: Digitalisierung des Darmstädter Tagblatts. In: Information – Wissenschaft & Praxis 73, 2–3, S. 89–96. DOI: 10.1515/iwp-2022-2210.

Urheberrechtsgesetz (UrhG) = Bundesamt für Justiz (2024): Gesetz über Urheberrecht (UrhG) und verwandte Schutzrechte. www.gesetze-im-internet.de/urhg/ (Stand: 8.9.2025).

Kontaktinformationen

Anne Christine Günther
Technische Universität Darmstadt
Magdalenenstraße 8
64289 Darmstadt
E-Mail: anne-christine.guenther@tu-darmstadt.de

Angela Hammer
Technische Universität Darmstadt
Magdalenenstraße 8
64289 Darmstadt
E-Mail: angela.hammer@tu-darmstadt.de

Dario Kampkaspar
Technische Universität Darmstadt
Residenzschloss 1
64283 Darmstadt
E-Mail: dario.kampkaspar@tu-darmstadt.de

Kevin Kuck
Technische Universität Darmstadt
Residenzschloss 1
64283 Darmstadt
E-Mail: Kevin.kuck@tu-darmstadt.de

Anna Christina Kupffer
Technische Universität Darmstadt
Residenzschloss 1
64283 Darmstadt
E-Mail: anna.kupffer@tu-darmstadt.de

Prof. Dr. Marcus Müller
Technische Universität Darmstadt
Residenzschloss 1
64283 Darmstadt
E-Mail: marcus.mueller@tu-darmstadt.de

Prof. Dr. Thomas Stäcker
Technische Universität Darmstadt
Magdalenenstraße 8
64289 Darmstadt
E-Mail: direktion@ulb.tu-darmstadt.de

Dr. Jörn Stegmeier
Universität Trier
54286 Trier
E-Mail: stegmeier@uni-trier.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

https://doi.org/10.21248/idsopen.16.2026.91

Dennis Beitel/Lisa Dücker/Salome Lipfert/Georg Oberdorfer

Hess-*ISCH*

Dialektaler Sprachwandel analysiert mit den Ressourcen von regionalsprache.de

Abstract This article provides an overview of the numerous resources offered by regionalsprache.de and the *REDE SprachGIS* and how, through the material and tools offered on these platforms, dialectal language change can be studied on a large scale from the end of the 19th until the beginning of the 21st century. This is demonstrated by a case study that concerns itself with the pronunciation of the coda of *ich* in Hessian dialects. Through the combination of several corpora that are based on the so-called Wenker sentences, the change in pronunciation can be observed in *real time* from the 19th century onward. In addition, data from the REDE-Neuerhebung corpus that is comprised of speakers of three generations are used to analyze the phenomenon in *apparent time*. Finally, additional survey data regarding the speakers' dialect competence from the *REDE-Infothek* are taken into account to present a multi-faceted view of the change in pronunciation.

Keywords dialectology, language change, audio corpora, phonology, German

1. Einleitung

Die Untersuchung von dialektalem Sprachwandel stellt die germanistische Sprachwissenschaft vor verschiedene Herausforderungen: Dialekte wurden und werden vor allem im Mündlichen gebraucht, was eine diachrone Betrachtung, die in die Zeit vor der Verbreitung von Aufnahmegeräten zurückreicht, erschwert. Doch selbst bei Auswertungen der für das 20. und 21. Jahrhundert vorliegenden Tonkorpora ist Vorsicht geboten. Viele der verfügbaren Korpora sind im Vergleich zu schriftbasierten Korpora nur wenig umfangreich und decken einen kleinen Raum ab. Bei der Kombination dieser kleineren Korpora ist wiederum zu bedenken, dass sie sich in der Auswahl der Gewährspersonen, Aufnahmesituationen, Aufnahmeorte und Erhebungsmethoden teils stark voneinander unterscheiden. Auch das Format, in dem die Daten zur Verfügung stehen, variiert (digitalisiert oder nicht, transkribiert oder nicht, verschiedene Annotationen etc.).

Für eine adäquate Untersuchung dialektalen Sprachwandels bedarf es entsprechend einer Vielzahl an Ressourcen, die mithilfe unterschiedlicher Tools behutsam miteinander kombiniert werden müssen. Das Langzeitprojekt regionalsprache.de (REDE) entwickelt und sammelt solche Ressourcen und stellt sie digital im *REDE SprachGIS* (Schmidt et al. 2020 f.) zur Verfügung, sodass sie u. a. auch gemeinsam ausgewertet werden können. Der Datenbestand, der über das *REDE SprachGIS* abgerufen werden kann, umfasst eine Vielzahl an digitalisierten Sprachatlanten und Tonaufnahmen, die unter anderem mit geografischen Metadaten angereichert sind. Auch die Karten und Erhebungsbögen aus dem *Sprachatlas des Deutschen Reichs* (Wenker 1889–1923) sind dort verfügbar. Anhand der Übersetzung von 40 hochdeutschen Sätzen wurde am Ende des 19. Jahrhunderts für diesen Sprachatlas

die bis heute umfassendste und engmaschigste Dialekterhebung des Deutschen durchgeführt (vgl. Fleischer 2017).

Diese sog. *Wenkersätze* wurden seitdem immer wieder genutzt, um weitere dialektologische Erhebungen durchzuführen. Ein Vergleich der schriftlichen Dialektdaten vom Ende des 19. Jahrhunderts mit jüngeren Tonkorpora, die die Wenkersätze ebenfalls nutzen, ermöglicht es, dialektale Sprachwandelprozesse seit dem ausgehenden 19. Jahrhundert auf der Ebene der Dialektkompetenz flächendeckend abzubilden.[1] Das *REDE SprachGIS* bietet die Möglichkeiten für eine solche Kombination von Ressourcen auf einer einzigen Plattform, was den Vergleich zwischen den unterschiedlichen Korpora erleichtert (vgl. Schmidt/Herrgen 2011, S. 147).

Als Anwendungsstudie für einen solchen Vergleich dient im vorliegenden Beitrag die Aussprachevariation des Auslauts im Pronomen *ich* im heutigen Hessen. Hierbei werden Ergebnisse aus dem *Digitalen Hessischen Sprachatlas* (DHSA), in dem Wenkerkarten aus dem 19. Jahrhundert mit Tonaufnahmen aus dem 21. Jahrhundert kombiniert werden, mit Daten aus den *Tonaufnahmen hessischer Mundarten* (TAHM), dem *Marburger Phonetischen Archiv I* (MrPhA I) und der REDE-Neuerhebung verglichen. Durch die vorsichtige Kombination der verschiedenen Tonkorpora mit den schriftlichen Daten aus den Erhebungsbögen der Wenkererhebung kann das Phänomen umfassend in seiner jüngeren Diachronie analysiert werden.

Über diese *real time*-Analyse hinaus ermöglicht die REDE-Neuerhebung mit der Berücksichtigung von Gewährspersonen aus drei Generationen, das Phänomen auch in *apparent time* zu untersuchen. Die Auswertung wird ergänzt durch die Analyse der Sozialdaten der REDE-Informanten aus der *REDE-Infothek*, in der u. a. Selbsteinschätzungen zur Dialektkompetenz bereitgestellt werden. Mit Blick auf das hiesige Vorhaben kann damit auch der Einfluss außersprachlicher Faktoren auf die Aussprachevariation des Auslauts von *ich* untersucht werden.

Durch die Kombination der Ressourcen im *REDE SprachGIS*, die in diesem Umfang und dieser Vielfalt für die Dialektologie des Deutschen einzigartig ist, wird anhand der Aussprachevariation des Auslauts von *ich* gezeigt, wie die Variation eines regionalen Sprachphänomens durch die kombinierte Auswertung unterschiedlicher Datensätze diachron, diatopisch und intergenerationell analysiert werden kann. Dafür werden zunächst in Abschnitt 2 die Ressourcen des *REDE SprachGIS* sowie die *REDE-Infothek* vorgestellt, bevor in Abschnitt 3 die Anwendungsstudie präsentiert wird. Ein Fazit in Abschnitt 4 schließt den Beitrag ab.

2. Regionalsprachliche Korpora im *REDE SprachGIS*

Das REDE-Projekt stellt zahlreiche Datensätze zur Regionalsprache zur Verfügung. Auf regionalsprache.de können regionalsprachliche Literatur recherchiert, unterschiedliche Sprachatlanten und deren Karten eingesehen, Tonaufnahmen angehört und Wenkerbögen angesehen und transkribiert werden.[2] Außerdem werden die entsprechenden Datensätze im geografischen Informationssystem *REDE SprachGIS* zusammengeführt und zur Verfü-

1 So kann zwar nicht der authentische Sprachgebrauch in seiner Entwicklung erfasst werden, aber dadurch, dass zu unterschiedlichen Zeitpunkten Gewährspersonen die gleiche Aufgabe (Übersetzung der gleichen Sätze in den intendierten Ortsdialekt) gestellt wurde, lassen sich dennoch Sprachwandelprozesse nachzeichnen.

2 Das Ansehen und Transkribieren der Wenkerbögen geschieht in der Wenkerbogen-App (https://apps.dsa.info/, Stand: 3.12.2025), die über eine Schnittstelle mit dem *REDE SprachGIS* verknüpft ist (Engsterhold/Glaser 2023).

gung gestellt. Somit lassen sich regionalsprachliche Korpora neben der Ansteuerung über die verschiedenen Kataloge auf regionalsprache.de auch direkt im *REDE SprachGIS* aufrufen, dort miteinander kombinieren sowie, dank der Georeferenzierung der Daten, im Raum verorten.

Die verschiedenen Datensätze im *SprachGIS* decken einen Zeitraum von knapp 130 Jahren ab und lassen so neben der synchronen Einsicht auch diachrone Zugänge zu. Zur Datenanalyse sind dabei neben den Wenkerbögen als historischer Ausgangspunkt vor allem die Tonkorpora bestens geeignet. Durch eine Kombination der verschiedenen Suchfunktionen, die das *SprachGIS* bereitstellt, ist es beispielsweise möglich, Sprachkarten aus einem Atlas mit Daten zu ausgewählten Ortspunkten aus weiteren Korpora anzureichern. Mit den Zeichentools des *SprachGIS* können dann komplexe Sprachkarten erstellt werden, die sowohl synchrone Verhältnisse als auch diachronen Wandel auf einen Blick sichtbar machen. Dieses Vorgehen wird beispielsweise bei der Erstellung des DHSA verwendet, bei dem vor dem Hintergrund von Wenkerkarten aus dem ausgehenden 19. Jahrhundert Tonaufnahmen aus dem 21. Jahrhundert dargestellt werden (Schmidt et al. 2023). Insbesondere diese Kombination von Sprachkarten und Tonaufnahmen aus verschiedenen Quellen eignet sich sehr gut, um diachronen Wandel eindrücklich darzustellen.

2.1 Regionalsprachliche Tonkorpora in *REDE* und seinem *SprachGIS*

Auf der REDE-Plattform werden Sprachaufnahmen zur Regionalsprache des Deutschen im überwiegend binnendeutschen, aber auch übrigen europäischen Gebiet versammelt. Dabei handelt es sich zumeist um Korpora, die Aufnahmen der Wenkersätze (siehe Abschn. 1) in einem bestimmten Raum enthalten (Schmidt/Herrgen 2011, S. 147). Teils entstammen sie Sprachatlasprojekten und können im *SprachGIS* gemeinsam mit den entsprechenden Karten eingesehen werden (insbesondere darunter „sprechende“ Atlanten wie der *Sprechende Sprachatlas von Unterfranken*, Fritz-Scheuplein/Klein 2021, oder der *Sprechende Norddeutsche Sprachatlas*, Elmentaler (Hg.) 2023), teils entstammen die Aufnahmen aber auch räumlich abgegrenzten Vorhaben mit anderem Impetus (z. B. Wörterbucherhebungen) oder gar geografisch weitreichenden Aufnahmeserien wie jenen aus dem Zwirner-Korpus (Zwirner 1956) und den umfangreichen Aufnahmen des MRPhA.

Abseits der Wenkersatzaufnahmen stellen REDE und das *SprachGIS* auch Tonaufnahmen, die auf den Abfragen aus Fragebüchern unterschiedlicher Sprachatlasprojekte beruhen, zur Verfügung. Inhaltlich handelt es sich dabei zumeist um lebensnahe Begrifflichkeiten und Tätigkeiten. Darüber hinaus sind noch in geringerem Umfang Ausschnitte freier Erzählungen und Aufnahmen von der Vorleseaussprache enthalten.

Auch die Aufnahmen, die im Zuge der REDE-Neuerhebung entstanden sind, befinden sich auf der REDE-Plattform. Dieses Korpus wurde von 2008 bis 2015 im Zuge des Langzeitprojekts erhoben und umfasst Aufnahmesettings, die von der standardintendierten Vorleseaussprache über Wenkersätze im intendierten Standard und intendierten Dialekt bis zu freien Gesprächen reichen.

Die verschiedenen Datensätze bilden im *SprachGIS* unterschiedlich dichte und unterschiedlich weitläufige Ortsnetze, indem sie sich überlagern, überschneiden oder auch – mehr oder minder – aneinandergrenzen. Allerdings können nicht alle regionalsprachlichen Korpora und Datenklassen, die dem *Forschungszentrum Deutscher Sprachatlas* (DSA) zur Verfügung stehen, ohne Weiteres vollständig auf der REDE-Plattform bereitgestellt werden, da im Kontext einer öffentlichen Plattform bestimmte Nutzungs- und Persönlichkeitsrechte

gewahrt werden müssen. Dies gilt insbesondere für freie Gesprächsdaten, die auf REDE bzw. im *SprachGIS* deshalb seltener vorkommen.[3] Eine Auflistung der Tonkorpora, die am Deutschen Sprachatlas zur Verfügung stehen, mit Stand 2023 findet sich in Fischer/Ganswindt/Oberdorfer (2023).

Mit den Tonkorpora und ihren teils ähnlichen Erhebungssettings lassen sich im *REDE SprachGIS* nicht nur räumliche Sprachbilder besonders gut erstellen. Aufgrund der unterschiedlichen Entstehungszeiträume und der diversen Geburtsjahre der Sprecher*innen lassen sich auch diachrone Studien durchführen. Wenn dann noch, wie oben erwähnt, die schriftlichen Sprachdaten aus den Wenkerbögen hinzugezogen werden, spannt sich hier ein Zeitrahmen von 130 Jahren auf, in dem die sprachliche Entwicklung analysiert werden kann. In den folgenden Kapiteln wird dies anhand einer Anwendungsstudie dargestellt. Zuvor wird die *REDE-Infothek* vorgestellt, in der verschiedene Informationen zur REDE-Neuerhebung gespeichert sind und die für die Auswertung mit den angesprochenen Tonaufnahmen kombiniert werden können.

2.2 Die *REDE*-Infothek

Die Plattform *REDE-Infothek* führt unter rede-infothek.dsa.info Informationen zur Erhebung, Aufbereitung und Auswertung der im REDE-Projekt erhobenen Daten zusammen (Schmidt et al. 2020 f.). Hierzu gehören unter anderem umfassende Informationen zum Erhebungsablauf und dem verwendeten technischen Equipment sowie Beschreibungen einzelner Arbeitsschritte bei der Aufbereitung der Sprachaufnahmen. Darüber hinaus werden verschiedene Analysemethoden wie die Formantmessung und die Dialektalitätswertmessung vorgestellt und Ergebnisse präsentiert.

Über Erläuterungen zu verwendeten Materialien und Arbeitsabläufen hinaus stellt die *REDE-Infothek* auch umfassende Datensätze bereit, wie zum Beispiel das REDE-Ortsnetz und eine Datentabelle mit sprachbiografischen Daten der REDE-Informanten, die im Zuge der REDE-Neuerhebung mittels eines Fragebogens erhoben wurden (vgl. Lipfert 2024).

In diesem wurde unter anderem nach dem Alter, dem Beruf und der Schulausbildung gefragt. Zudem wurden die REDE-Informanten gebeten, sich mithilfe von Antwortskalen (0 = „gar nicht“ bis 6 = „perfekt“) hinsichtlich der folgenden Fragen zur Dialektkompetenz selbst einzuschätzen:

1. „Wie gut können Sie das [Platzhalter][4] Ihres Heimatortes verstehen?“
2. „Wie gut können Sie das [Platzhalter][5] Ihres Heimartortes sprechen?“

3 Bei wissenschaftlichem Interesse ist ein Zugriff auf diese Materialien aber auf anderem Wege über Kontaktaufnahme möglich.

4 In die durch eckige Klammern gekennzeichneten Platzhalter wurde die von den Informanten genannte Bezeichnung für die Sprechweise der Alteingesessenen am Ort eingetragen. Beispielsweise gaben die REDE-Sprecher aus Fulda neben „Platt“ auch „Schlitzer Dialekt“ (FDJUNG5) oder „Rhöner Platt“ (FD7) an, in Homberg (Efze) bezeichneten die REDE-Sprecher die Sprechweise ausschließlich als „Platt“.

5 In die durch eckige Klammern gekennzeichneten Platzhalter wurde die von den Informanten genannte Bezeichnung für die Sprechweise der Alteingesessenen am Ort eingetragen. Für Beispiele von Sprechweisenbezeichnungen vgl. Fußnote 4.

Mit der ersten Frage wird die Selbsteinschätzung zur Verstehenskompetenz und mit der zweiten Frage die Selbsteinschätzung zur aktiven Sprachkompetenz ermittelt. Diese und weitere Zusatzmaterialien stehen auf der Homepage und im Linguistischen Repositorium des DSA zur Verfügung.[6] Diese Daten ermöglichen einerseits eine informierte Nachnutzung der Daten aus der REDE-Neuerhebung und bieten andererseits ergänzendes Material zur Auswertung der Korpusdaten.

Nach diesem Überblick über die verschiedenen Ressourcen, die von REDE zur Verfügung gestellt werden, folgt im nächsten Abschnitt die Anwendungsstudie, anhand derer die Kombination von verschiedenen Ressourcen aus dem REDE-Kontext demonstriert und mittels der im *SprachGIS* erstellten Sprachkarten eine effektive Darstellung der umfangreichen Daten präsentiert wird.

3. Anwendungsstudie: Aussprache des Auslauts bei *ich* in Hessen in der diachronen Entwicklung

Die Anwendungsstudie beschäftigt sich mit der Verteilung der Aussprachevarianten des Auslauts in *ich* seit dem ausgehenden 19. Jahrhundert. Dabei ist besonders das Aufkommen der Realisation mit auslautendem [ʃ], hier und im Folgenden auch umschrieben als *isch*, von Interesse. Diese Realisierung wird heute im allgemeinen Sprachbewusstsein mit dem hessischen Sprachraum in Verbindung gebracht (Vorberger 2019, S. 128–134). Die Karte des DHSA zum Auslaut von *ich* (vgl. Abb. 1) kontrastiert die Daten der Wenker-Erhebung von 1880 (farbige Flächenkarte im Hintergrund) mit den Ergebnissen einer modernen Tonaufnahmeserie von 2014 (farbige Lautsprechersymbole). Dadurch werden diachrone Veränderungen der Lautverteilung unmittelbar sichtbar: Im südlichen Teil des Untersuchungsareals ist 1880 überwiegend die Variante *ich* belegt, lediglich in Griesheim findet sich ein Nachweis für die Variante *isch* (kleiner orangener Punkt). Demgegenüber weist die Erhebung von 2014 in demselben Raum eine klare Dominanz von [ʃ]-Belegen für den Auslaut von *ich* auf. Für den nördlichen Teil Hessens, der dem niederdeutschen Sprachraum zugeordnet wird, ist hingegen sowohl 1880 als auch 2014 die Auslautvariante [k] belegt.

Dieser Befund verweist auf einen kurzzeitdiachronen Sprachwandel in einem Zeitraum von etwa 130 Jahren. Die Analyse dieser Entwicklung zwischen den hier repräsentierten Zeitschnitten kann durch die Ressourcen des *REDE SprachGIS* aufgrund der gleichlautenden Abfragekontexte sinnvoll ergänzt werden.

6 Die Daten zur *REDE-Infothek* stehen im LinguRep unter https://hdl.handle.net/20.500.14450/19355 (Stand: 3.12.2025) zur Verfügung.

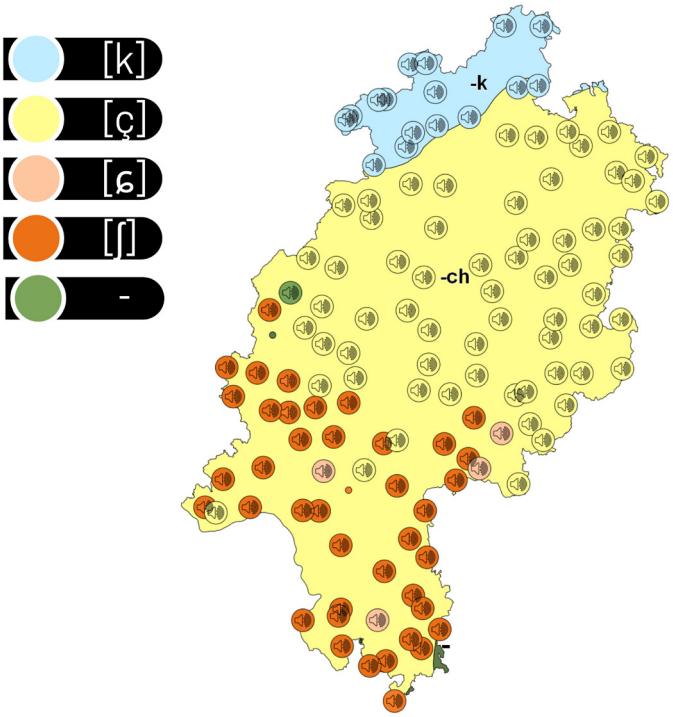

Abb. 1: Karte zur Aussprache des Auslauts von *ich* aus dem DHSA (regionalsprache.de/Map/MiRMoPQU, Stand: 3.12.2025). Die Flächenkarte im Hintergrund ist eine Nachzeichnung der Wenkerkarte 143 und illustriert damit die Lautverhältnisse im ausgehenden 19. Jahrhundert. Farbige Punkte ohne Lautsprechersymbol stellen Ausreißer aus der Wenkererhebung in einem ansonsten einheitlichen Gebiet dar

Für die Untersuchung dieses Phänomens erweist sich das Bundesland Hessen aus zwei wesentlichen Gründen als besonders geeignet:

1. Dialektale Vielfalt: Das Bundesland Hessen zeichnet sich durch eine hohe dialektale Heterogenität aus. Neben den hessischen Basisdialekten – Nord-, Ost- und Zentralhessisch – gehört auch das Rheinfränkische zum hochdeutschen Sprachraum. Darüber hinaus sind im Norden Hessens mit dem Westfälischen und Ostfälischen auch niederdeutsche Dialekte vertreten (vgl. Schmidt et al. 2023, S. 104).
2. Diachrone Abdeckung: Die umfassende Dokumentation durch Tonaufnahmen aus unterschiedlichen Erhebungszeiträumen wie dem Korpus der TAHM (1982) und dem DHSA (2014) ermöglicht eine systematische Gegenüberstellung und damit die kleinschrittige Nachzeichnung des ausgewählten Sprachwandelphänomens.

Die konkrete Auswahl des verwendeten Datenmaterials sowie dessen Aufbereitung wird im folgenden Abschnitt vorgestellt.

3.1 Auswahl der Tonkorpora und Erhebungsorte

Für die Anwendungsstudie wurden Tonkorpora ausgewählt, die in Kombination eine möglichst große Zeitspanne abdecken und einen räumlichen Schwerpunkt in Hessen haben. Neben dem DHSA, der Wenkererhebung und der REDE-Neuerhebung wurden die Aufnahmen aus dem Korpus TAHM (Kehrein/Vorberger 2018, S. 138) und dem MRPhA I (Göschel (Hg.) 1977) ausgewählt (vgl. auch Fischer/Ganswindt/Oberdorfer 2023, S. 192–194).

Das TAHM umfasst insgesamt 499 Tonaufnahmen mit dialektalen Übersetzungen der Wenkersätze an 298 Orten und wurde 1982 im *Forschungsinstitut für deutsche Sprache* (später DSA) an der Philipps-Universität Marburg erhoben. Das MRPhA umfasst Tonaufnahmen, die vom DSA ab Anfang der 1960er-Jahre durchgeführt wurden. Dessen erste Serie (MRPhA I), die hier verwendet wird, umfasst ca. 550 Aufnahmen. Es handelt sich dabei überwiegend um gesprochene Aufnahmen der Wenkersätze im Dialekt aus dem Zeitraum von 1964 bis 1993. Für beide Korpora liegen Angaben zu den Geburtsjahren der Gewährspersonen vor, die zwischen dem Ende des 19. und der Mitte des 20. Jahrhunderts liegen. Das Ortsnetz für das MRPhA I ist dabei nicht gleichmäßig über Hessen verteilt, sondern hat verschiedene regionale Schwerpunkte.

Da der DHSA als Vorbild für die Anwendungsstudie dient, wurden die darin enthaltenen 135 Aufnahmen aus 132 Erhebungsorten als Grundlage für das Ortsnetz genutzt. Da es zwischen dem DHSA und der REDE-Neuerhebung keine Überschneidung in den Erhebungsorten gibt, ergänzen die 14 hessischen Orte mit insgesamt 79 Aufnahmen das Ortsnetz (vgl. Abschn. 2.1).[7] Die Auswahl der Orte aus den übrigen Korpora orientiert sich an diesen kombinierten Ortsnetzen aus insgesamt 146 Orten. Entsprechend gehen aus dem TAHM 133 Aufnahmen aus 89 Orten in die Untersuchung ein[8], aus dem MRPhA I sind es 83 Aufnahmen aus 46 Orten[9]. Dazu kommen 142 Wenkerbögen. Durch dieses Verfahren entsteht ein dichtes Ortsnetz mit Erhebungspunkten in ganz Hessen und obwohl sich die Ortsnetze der verschiedenen kombinierten Korpora am Ende noch leicht voneinander unterscheiden, ermöglicht die Dichte der Ortspunkte für Hessen die Analyse der diachronen Entwicklung.

Aus diesem Datensatz wurde jeweils die Realisierung des Auslauts von *ich* im Wenkersatz 10 *Ich will es auch nicht mehr wieder thun!* nach den REDE-Transkriptionskonventionen (Lipfert 2023) transkribiert[10] und annotiert, die in allen untersuchten Aufnahmen von den Gewährspersonen im intendierten Ortsdialekt erfragt wurde.[11] Die Anzahl der Gewährspersonen und damit die Anzahl der im Folgenden analysierten Datenpunkte beträgt 418, wobei pro Ort unterschiedliche Anzahlen an Gewährspersonen vorliegen.

Die folgenden Abschnitte zeigen, wie der nach den genannten Gesichtspunkten zusammengestellte Datensatz dazu verwendet werden kann, um die Variation in der Realisierung dieses Phänomens auf verschiedenen Ebenen zu analysieren. Da hierbei schriftliche Laientranskriptionen und mündliche Erhebungsdaten miteinander verglichen werden, gilt stets zu berücksichtigen, „dass die phonische Interpretation der graphischen Zeichen immer nur auf einer bestimmten Abstraktionsebene verbleibt“ (Elmentaler/Rosenberg 2022, S. 85). Nichtsdestoweniger zeigen u. a. Schmidt/Herrgen (2011, S. 108–127), Elmentaler/Rosen-

7 Aus der REDE-Neuerhebung wurden jeweils die dialektintendierte Übersetzung der Wenkersätze aller vorhandenen Sprecher für die vorliegende Analyse ausgewählt.

8 Aus dem TAHM sind teilweise abweichende, aber benachbarte Orte aufgenommen worden, wenn keine Übereinstimmung mit den Erhebungsorten des DHSA oder der REDE-Neuerhebung bestand.

9 Teilweise enthält das MrPhA I auch Aufnahmen von Kleinkindern an den ausgewählten Erhebungsorten. Diese wurden von der vorliegenden Untersuchung ausgeschlossen.

10 Die hier untersuchten Daten aus der REDE-Neuerhebung lagen zum Teil bereits nach den REDE-Transkriptionskonventionen transkribiert vor. Für Sprachaufnahmen ohne phonetische Transkription wurden diese – unter Berücksichtigung der REDE-Transkriptionen – speziell für die vorliegende Analyse angefertigt.

11 Alle für die folgenden Auswertungen verwendeten Datensätze mit den Daten aus den Wenkerbögen, TAHM, MrPhAI und der REDE-Neuerhebung stehen unter https://hdl.handle.net/20.500.14450/133794 (Stand: 3.12.2025) zur Verfügung.

berg (2022), Schmidt et al. (2023) und Frank/Rohloff/Peters (im Ersch.)[12], dass aus den schriftlichen Daten der Wenkererhebung durchaus auf lautliche Verhältnisse der zeitgenössischen Dialekte geschlossen werden kann.

3.2 Analyse der Auslautvariation

Durch die Kombination von Daten, die zu unterschiedlichen Erhebungszeitpunkten gewonnen wurden, kann die Auslautvariation von *ich* in *real time* untersucht werden. Die Berücksichtigung von drei Generationen aus der REDE-Neuerhebung ermöglicht zudem eine Auswertung in *apparent time*. Zusätzlich können die Selbsteinschätzungen der REDE-Probanden zur Dialektkompetenz über die *REDE-Infothek* herangezogen werden, um mögliche Korrelationen zwischen den Auslautvarianten und den Selbsteinschätzungen zu eruieren.

3.2.1 Interpretation des Wandels in *real time*

Für die Analyse des Aussprachewandels in *real time* werden mehrere Erhebungszeitpunkte zwischen 1880 und 2014 herangezogen, um diachrone Veränderungen des Phänomens in Kartenform sichtbar zu machen. Die Auswertung der Wenkerbögen zeigt zunächst das Fehlen der Auslautvariante <sch> im ausgehenden 19. Jahrhundert (vgl. Abb. 2).[13] Im nördlichen Hessen ist die niederdeutsche Variante <k> vorherrschend, im restlichen Gebiet dominiert <ch>. Viermal ist im westlichen Teil des Untersuchungsgebiets auch die Schreibung mit auslautendem <g> belegt. Diese kann als alternative Schreibung für [k] interpretiert werden, allerdings ist es auch möglich, dass die Schreibung auf eine Aussprache mit Spirans [ç] verweisen soll (vgl. <König> [kø:nɪç]), weshalb hier auf eine lautliche Interpretation verzichtet wird. Abbildung 2 stellt lediglich die schriftlichen Realisierungen des Auslauts in den Wenkerbögen dar.[14]

Grundsätzlich ist bei der Interpretation der schriftlichen Wenkerdaten zu berücksichtigen, dass die Lehrer, die als Gewährspersonen dienten, zwar dazu angehalten waren, die Übersetzungen mit größtmöglicher Sorgfalt und Genauigkeit anzufertigen und zugleich die lokale Mündlichkeit möglichst exakt zu repräsentieren (vgl. Schmidt/Herrgen 2011, S. 100). Eine verzerrende Orientierung an der Schriftnorm ist aber natürlich nicht auszuschließen, zumal für bestimmte Laute keine üblichen lateinischen Schriftzeichen vorliegen, bspw. alveolo-palatales [ɕ] (s. u.).

12 Für eine zeitgenössische kritische Auseinandersetzung mit der Repräsentativität der schriftlichen Wenkerdaten vgl. Bremer (1895), Mitzka (1952) und Wrede (1895). Zur wissenschaftlichen Auseinandersetzung mit verschrifteten Dialektdaten allgemein vgl. Auer (1990) und Kleiner (2006). Speziell zur Kombination der Wenkerdaten mit anderen Arten von Dialektdaten vgl. Schmidt/Herrgen (2011, S. 147).

13 Die Auswahl der Orte für die Auswertung der Wenkerbögen orientiert sich an den in Abschnitt 3.1 genannten Kriterien. Deshalb ist der Ort Griesheim trotz des oben erwähnten Belegs der geschriebenen Variante <isch> nicht in die Auswertung eingeflossen. Der Wenkerbogen 27255 Griesheim kann unter https://doi.org/10.57712/lingurep-29205 (Stand: 3.12.2025) abgerufen werden.

14 Ganswindt (2017, S. 205–207, S. 211) und Lenz (2007, S. 182–186) weisen auf Herausforderungen bei der phonetischen Interpretation der Wenkerdaten in Bezug auf g-Spirantisierung und die Velarisierung von [ç] hin (wenn auch nicht bezogen auf den Auslaut von ich). So ist es durchaus möglich, dass im Korpus keine Belege für <sch> zu finden sind, weil die Gewährspersonen „schriftsprachorientierten Fehlschreibungen“ verwenden (Ganswindt 2017, S. 202).

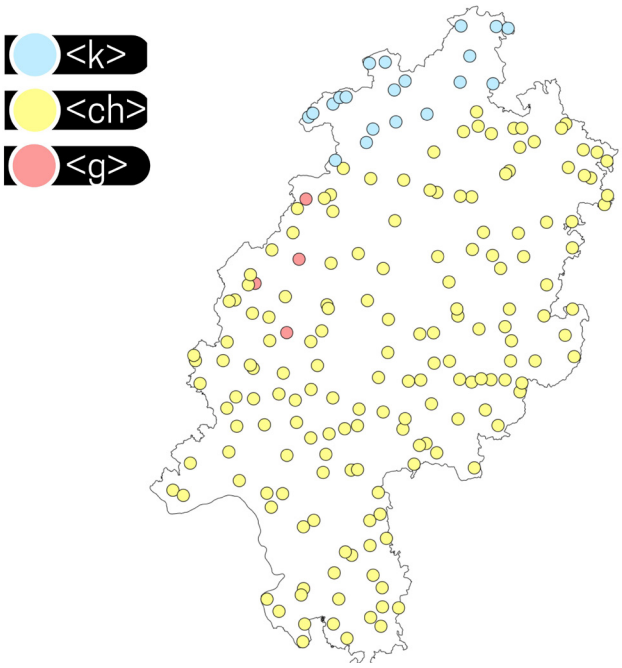

Abb. 2: Der Auslaut von *ich* in Wenkersatz 10 in den ausgewählten Wenkerbögen (n = 142)

Im Vergleich zu den Wenkerdaten dokumentieren die Tonaufnahmen aus dem TAHM und dem MRPhA I aus der Mitte des 20. Jahrhunderts zahlreiche Belege für die alveolare [ʃ]-Variante (vgl. Abb. 3). Besonders im Süden Hessens wird auslautendes [ʃ] in dieser Erhebungsphase zur dominanten Aussprachеform, während die Belegdichte für diese Variante Richtung Norden kontinuierlich abnimmt. Zudem treten in der Mitte Hessens vereinzelt Belege für alveolo-palatales [ɕ] auf, das eine phonetische Zwischenstellung zwischen palatalem [ç] und alveolarem [ʃ] einnimmt. Zwischen den Daten aus dem TAHM und dem MrPhA I zeigen sich dabei keine Unterschiede bzgl. der untersuchten Variable.

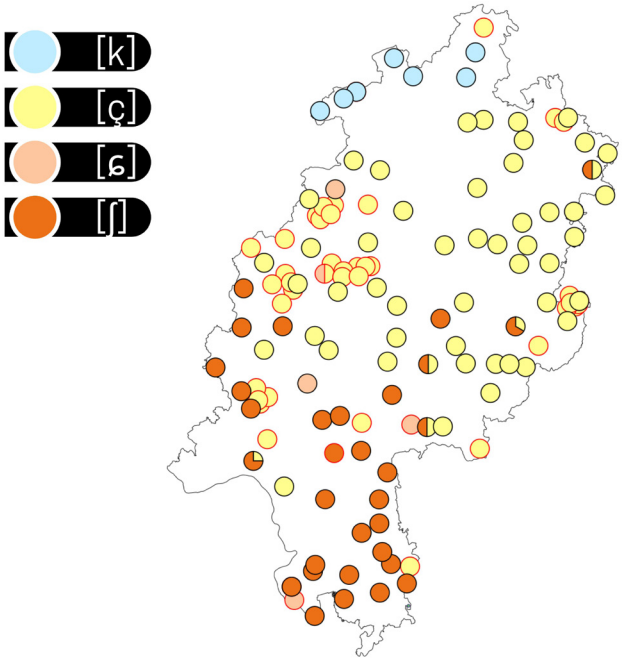

Abb. 3: Auslaut von *ich* in den Korpora TAHM und MRPhA I. Die Punkte mit roter Umrandung stammen aus dem MRPhA I (86 Aufnahmen an 46 Orten), alle übrigen Erhebungsorte sind aus dem TAHM (133 Orte an 89 Orten). Punkte mit mehreren Farben symbolisieren verschiedene Realisierungen an einem Ortspunkt

Die hier beobachtete Verteilung kontrastiert mit den Wenkerdaten, insbesondere bzgl. der Verbreitung der Varianten mit [ʃ] und [ɕ] in der südlichen Hälfte des Untersuchungsgebiets. Es ist aufgrund der bereits genannten Faktoren nicht gänzlich auszuschließen, dass diese Auslautvarianten bereits zum Zeitpunkt der Erhebung verbreiteter waren, in den schriftlichen Daten jedoch keine bzw. keine systematische Berücksichtigung fanden, da die Gewährspersonen sich zu sehr an der Standardschreibung orientierten oder das ihnen zur Verfügung stehende Buchstabeninventar nicht zur adäquaten Wiedergabe des verwendeten Lauts ausreichte.

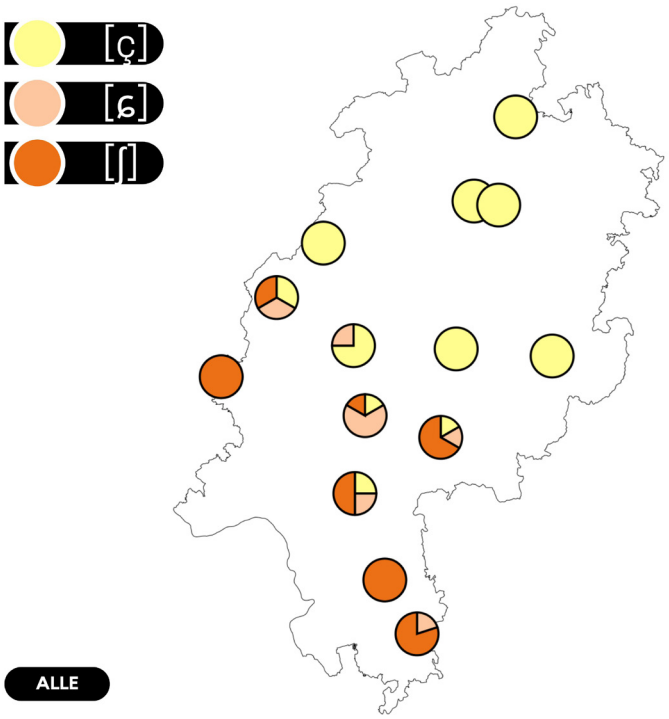

Abb. 4: Karte zur Aussprache des Auslauts von *ich* bei allen REDE-Sprechern (79 Aufnahmen an 14 Orten)[15]

Ein Vergleich der Ausbreitungsmuster aus den TAHM- und MRPhA-I-Daten mit den neueren Tonaufnahmen des DHSA (Abb. 1) und der REDE-Neuerhebung (Abb. 4) zu Beginn des 21. Jahrhunderts zeigt eine Stabilisierung der Variante mit auslautendem [ʃ]. In den DHSA-Daten sowie den aggregierten Daten der REDE-Neuerhebung lässt sich die Verbreitung dieser Variante mittlerweile vom Süden bis in die Mitte des Bundeslandes nachweisen.

Während in Abbildung 3 noch keine deutliche Ballung der [ʃ]-Variante im südwestlichen Hessen zu erkennen ist, zeigt sich in den neueren Erhebungen eine ausgeprägte Konzentration in diesem Raum. Insgesamt lässt sich für den Erhebungszeitraum des DHSA und der REDE-Neuerhebung eine nahezu flächendeckende Ausbreitung des [ʃ]-Auslauts für die südliche Hälfte des Untersuchungsgebiets feststellen. Dies deutet darauf hin, dass sich die [ʃ]-Variante in der gesprochenen Sprache durchgesetzt hat und im Laufe des späten 20. und frühen 21. Jahrhunderts zunehmend regional stabilisiert wurde.

15 Unter den 14 hessischen Orten aus der REDE-Neuerhebung sind keine aus dem niederdeutschen Gebiet enthalten. Das Fehlen der niederdeutschen Form auf dieser und weiteren REDE-Karten ist also nicht als Nachweis für den Rückgang des plosivischen Auslauts zu verstehen.

3.2.2 Interpretation des Wandels in *apparent time*

Als weiteren Zugang ermöglicht die REDE-Neuerhebung mit drei Alterskohorten eine Analyse in *apparent time*: eine ältere Generation über 65 Jahre, eine mittlere zwischen 45 und 55 Jahren und eine junge Generation um die 20 Jahre.[16] Pro Ort und Alterskohorte liegen Aufnahmen von mindestens einem Sprecher der älteren und der jüngeren Generation und von zwei Sprechern der mittleren Generation vor.[17]

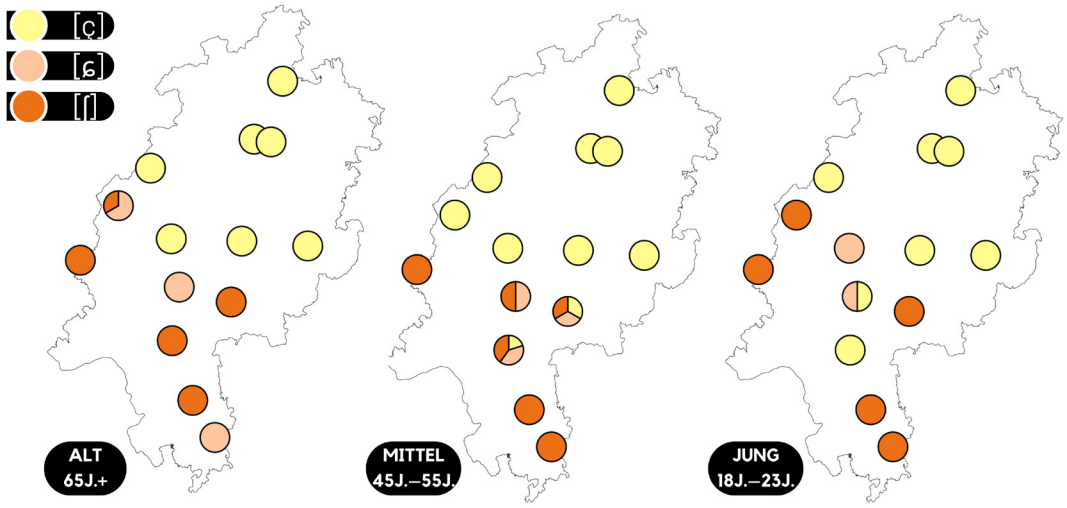

Abb. 5: Karte zur Aussprache des Auslauts von *ich* in der alten Generation der REDE-Sprecher

Die Analyse der Tonaufnahmen zeigt nur geringe intergenerationelle Unterschiede in der Realisierung des Auslauts von *ich*. Die Variantenverteilung in Abbildung 5 zeigt die bereits beobachtete Nord-Süd-Distribution in allen drei Generationen mit dominierendem [ç] im Norden. In den beiden südlichsten Erhebungsorten zeigt sich in allen drei Generationen ein Nebeneinander von [ʃ] und [ɕ]. In weiter nördlich gelegenen Orten sind darüber hinaus [ç]-Realisierungen in der mittleren und jüngeren Generation belegt, während diese Variante in der älteren Generation ausschließlich in der nördlichen Hälfte des Untersuchungsgebiets auftritt.

Für die drei Generationen der REDE-Neuerhebung ist also, abgesehen von leicht erhöhten Anteilen für die [ç]-Variante im Süden, kaum ein Wandel in *apparent time* festzustellen. Stattdessen herrscht im Süden Hessens ein Nebeneinander von verschiedenen Varianten, während im nördlichen Teil die Aussprache mit [ç] dominiert. Dies könnte darauf hindeuten, dass Angehörige der mittleren und der jungen Alterskohorte stärker zur standardnahen Aussprache tendieren und möglicherweise weniger dialektkompetent sind als Angehörige der alten Alterskohorte. Eine Korrelation der Aussprachevariation mit den Selbsteinschätzungen der REDE-Sprecher erfolgt im nächsten Abschnitt.

16 Altersangaben beziehen sich auf den Erhebungszeitraum der REDE-Neuerhebung in den Jahren 2008–2015. Alle Gewährspersonen waren Männer.

17 Teilweise liegt die Anzahl der Gewährspersonen höher, da in manchen Orten zusätzliche Personen aufgenommen wurden. Um keine verfälschende Auswahl der Sprecher zu treffen, wurden jeweils alle vorliegenden Aufnahmen für die vorliegenden Analysen verwendet. Bad Nauheim n = 6, Biedenkopf n = 4, Borken (Hessen) n = 6, Büdingen n = 6, Dillenburg n = 6, Erbach n = 5, Frankfurt am Main n = 8, Fulda n = 12, Gießen n = 4, Homberg (Efze) n = 5, Kassel n = 4, Reinheim n = 4, Thalheim n = 4, Ulrichstein n = 5.

3.3 Auslautvariante und Selbsteinschätzung

Abschließend werden die Selbsteinschätzungen der REDE-Informanten mit der von ihnen realisierten dialektintendierten Auslautvariante in *ich* im Wenkersatz 10 in Beziehung gesetzt. Dabei wird sowohl die Selbsteinschätzung zur aktiven Sprechkompetenz als auch die zur Verstehenskompetenz für die von den Informanten genannten Sprechweise der Alteingesessenen am Ort berücksichtigt (vgl. Abschn. 2.3).

Abbildung 6 und Abbildung 7 setzen die Werte der Selbsteinschätzungen der REDE-Informanten mit der von ihnen realisierten Auslautvariante in *ich* in Beziehung. Die Werte der Selbsteinschätzungen sind auf der X-Achse in Säulen von 0 („gar nicht“) bis 6 („perfekt“) angegeben. Die Auslautvarianten in *ich* sind als verschiedenfarbige Abschnitte innerhalb der Säulen inklusive ihrer absoluten Belegzahl veranschaulicht.[18]

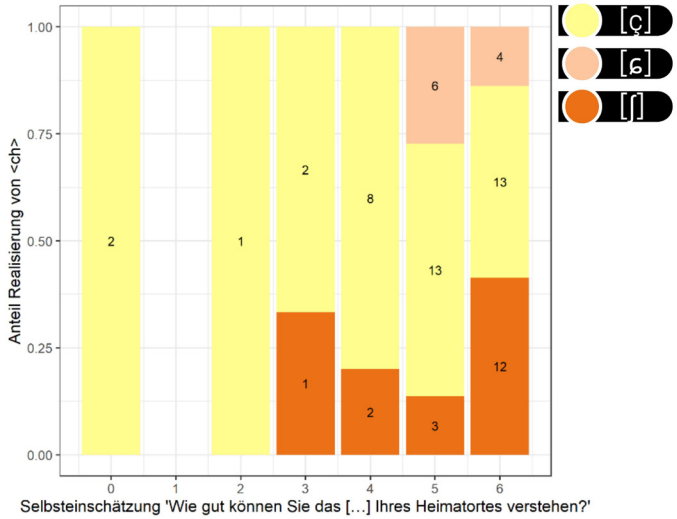

Abb. 6: Selbsteinschätzung der REDE-Informanten zur Verstehenskompetenz für die Sprechweise der Alteingesessenen am Ort mit den Auslautvarianten in *ich*

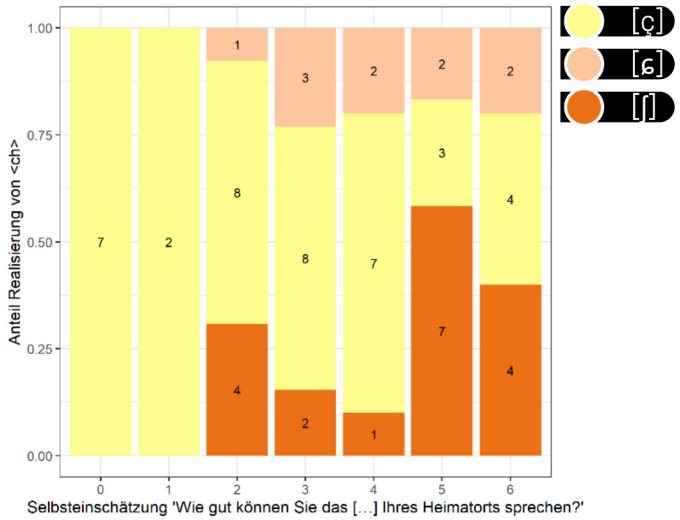

Abb. 7: Selbsteinschätzung der REDE-Informanten zur Sprechkompetenz für die Sprechweise der Alteingesessenen am Ort mit den Auslautvarianten in *ich*

18 Die niederdeutsche Variante [k] ist, wie in Abschnitt 3.2.2 beschrieben, bei den REDE-Informanten nicht belegt.

Für diejenigen Informanten, die ihre Kompetenz in Verständnis oder Sprechen mit 0 bis 2, also eher niedrig, einschätzen, ist ausschließlich auslautendes [ç] belegt. Die vorverlagerten Varianten [ɕ] und [ʃ] treten erst ab einer Selbsteinschätzung von 2 auf. Stärker nach vorne verlagertes [ʃ] ist dabei insgesamt häufiger bei Sprechern zu verzeichnen, die ihre Kompetenzen mit mindestens 5, das heißt „gut" bis „sehr gut", bewerten.

Damit ist festzustellen, dass die vorverlagerten Varianten [ɕ] und [ʃ] grundsätzlich häufiger bei höheren Selbsteinschätzungen zu verzeichnen sind. Schlussfolgernd ist bei vorverlagerten Varianten von Merkmalen auszugehen, die eine erhöhte Selbsteinschätzung voraussetzen, um mit der Sprache der Alteingesessenen am Ort assoziiert zu werden. Unter Berücksichtigung dieser Annahme ist demnach nicht auszuschließen, dass es sich bei den vorverlagerten Varianten [ɕ] und [ʃ] um Erweiterungen im Lautinventar des intendierten Ortsdialekts handelt.

4. Fazit

Der vorliegende Beitrag hat gezeigt, wie das *REDE SprachGIS* mit seinen vielfältigen Ressourcen dazu genutzt werden kann, um regionalen Sprachwandel vom Ende des 19. bis zum Beginn des 21. Jahrhunderts (hier die Aussprachevariation im Auslaut von *ich* in Wenkersatz 10) durch die Kombination verschiedener Tonkorpora mit einer historischen Dialekterhebung in schriftlicher Form zu untersuchen. Eine Herausforderung besteht hierbei darin, dass Daten unterschiedlicher Medialität miteinander verglichen werden. Dies gilt umso mehr, als das untersuchte Phänomen in den Audiodaten Varianten wie [ɕ] aufweist, für die im Schriftdeutschen kein gängiges Schriftzeichen existiert.

Trotz allem konnte gezeigt werden, dass sich in einem Zeitraum von knapp 130 Jahren die Aussprachevariante mit auslautendem [ʃ] vor allem im Süden des Untersuchungsgebiets stark ausgebreitet hat. Unter Einbezug der Daten aus der REDE-Neuerhebung konnte das Phänomen außerdem anhand von drei Generationen in *apparent time* analysiert werden. Die zusätzlichen Informationen aus der *REDE-Infothek* ermöglichten es zudem, die Auslautvarianten den Selbsteinschätzungen zur Dialektkompetenz gegenüberzustellen.

Das Potenzial der Plattform ist damit noch lange nicht ausgeschöpft: Mithilfe der hier verwendeten Ressourcen und Tools wurden verschiedene Datensätze zu einer diachronen Studie kombiniert, die ein sprachliches Phänomen aus mehreren Perspektiven analysiert. Für zukünftige Analysen könnten noch zusätzliche Ortspunkte aus mehr Tonkorpora ausgewertet werden, zusätzliche Sozialdaten aus der *REDE-Infothek* wie das Geburtsjahr und die sprachliche Prägung herangezogen werden. Das Korpus könnte außerdem um weitere Ressourcen wie Ortsgrammatiken, die über die in REDE integrierte Georeferenzierte Online-Bibliografie zur Areallinguistik (GOBA) geografisch präzise abgerufen werden können, ergänzt werden. Die verschiedenen hier diskutierten Einflussfaktoren in statistischen Analysen könnten außerdem miteinander korreliert werden.

Literatur

Auer, Peter (1990): Phonologie der Alltagssprache. Eine Untersuchung zur Standard/Dialekt-Variation am Beispiel der Konstanzer Stadtsprache. (= Studia Linguistica Germanica 28). Berlin/New York: De Gruyter.

Bremer, Otto (1895): Beiträge zur Geographie der deutschen Mundarten. In Form einer Kritik von Wenkers Sprachatlas des Deutschen Reichs. Leipzig: Breitkopf & Härtel.

Elmentaler, Michael (Hg.) (2023): Sprechender Norddeutscher Sprachatlas (SNOSA), auf Grundlage des „Norddeutschen Sprachatlas (NOSA). Bd. 1: Regiolektale Sprachlagen“ von Michael Elmentaler und Peter Rosenberg. Unter Mitarbeit von Vivian Lucia Hansmann, Johanna Falkson sowie Martin Wolf (Erstellung der Tonproben) und Tuarik Buanzur (technische Bearbeitung). In: REDE, bearbeitet von Robert Engsterhold und Marina Frank. In: Schmidt/Herrgen/Kehrein/Lameli (Hg.).

Elmentaler, Michael/Rosenberg, Peter (2022): Norddeutscher Sprachatlas (NOSA). Bd. 2: Dialektale Sprachlagen. (= Deutsche Dialektologie). Hildesheim u. a.: Olms.

Engsterhold, Robert/Elvira Glaser (2023): Ein Jahr Wenkerbögen-App. In: Sprachspuren: Berichte aus dem Deutschen Sprachatlas 3, 5. https://doi.org/10.57712/2023-05.

Fischer, Hanna/Ganswindt, Brigitte/Oberdorfer, Georg (2023): Die regionalsprachlichen Tonkorpora des Forschungszentrums Deutscher Sprachatlas. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 11). Tübingen: Narr, S. 189–202.

Fleischer, Jürg (2017): Geschichte, Anlage und Durchführung der Fragebogen-Erhebungen von Georg Wenkers 40 Sätzen. Dokumentation, Entdeckungen und Neubewertungen. Hildesheim/Zürich/New York: Olms.

Frank, Marina/Rohloff, Tio/Peters, Jörg (im Ersch.): Wandel und Stabilität der Sprachkompetenz in Übersetzungsdaten. Eine Variablenanalyse niederdeutscher Wenkersätze. In: Zeitschrift für Dialektologie und Linguistik 92, 2, S. 179–223.

Fritz-Scheuplein, Monika/Klein, Wolf P. (2021): Sprechender Sprachatlas von Unterfranken (SprSUF). Bearbeitet von Verena Diehm und Monika Fritz-Scheuplein. Würzburg. In: Regionalsprache.de (REDE), bearbeitet von Robert Engsterhold, Vanessa Lang, Georg Oberdorfer und Anna Wolańska. In: Schmidt/Herrgen/Kehrein/Lameli (Hg.).

Ganswindt, Brigitte (2017): Landschaftliches Hochdeutsch: Rekonstruktion der oralen Prestigevarietät im ausgehenden 19. Jahrhundert. (= Zeitschrift für Dialektologie und Linguistik – Beihefte 168). Stuttgart: Steiner.

Göschel, Joachim (Hg.) (1977): Die Schallaufnahme deutscher Dialekte im Forschungsinstitut für deutsche Sprache. Bestandsbeschreibung und Arbeitsbericht. Marburg: Universität Marburg, Abteilung Phonetik.

Kehrein, Roland/Vorberger, Lars (2018): Dialekt- und Variationskorpora. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Korpuslinguistik. (= Germanistische Sprachwissenschaft um 2020 5). Berlin/Boston: De Gruyter, S. 125–150.

Kleiner, Stefan (2006): Sprachvariation bei Polizeinotrufen in Südbaden. Eine Fallstudie im Rahmen des Notruf-Pilotprojekts. In: Klausmann, Hubert (Hg.): Raumstrukturen im Alemannischen. Beiträge der 15. Arbeitstagung zur alemannischen Dialektologie. Schloss Hofen Lochau (Vorarlberg) vom 19.–21.9.2005. (= Schriften der Vorarlberger Landesbibliothek 15). Graz-Feldkirch: WN-Verlag, S. 189-193. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/5684/file/Kleiner_Sprachvariation_bei_Polizeinotrufen_in_Suedbaden_Eine_Fallstudie_im_Rahmen_des_Notruf_Pilotprojekts_2006.pdf.

Lenz, Alexandra N. (2007): Zur variationslinguistischen Analyse regionalsprachlicher Korpora. In: Kallmeyer, Werner/Zifonun, Gisela (Hg.): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. (= Jahrbuch des Instituts für Deutsche Sprache 2006). Berlin/New York: De Gruyter, S. 169–202.

Lipfert, Salome (2023): REDE-Infothek. Phonetische Transkriptionskonventionen. In: Schmidt/Herrgen/Kehrein/Lameli (Hg.). https://doi.org/10.57712/lingurep-62373.

Lipfert, Salome (2024): REDE-Infothek: Sprachliche Prägung der REDE-Informanten. In: Schmidt/Herrgen/Kehrein/Lameli (Hg.). www.rede-infothek.dsa.info/?page_id=960.

Mitzka, Walther (1952): Handbuch zum Deutschen Sprachatlas. Marburg: Elwert.

Schmidt, Jürgen Erich/Beitel, Dennis (2020ff.): Digitaler hessischer Sprachatlas (DHSA). Unter Mitarbeit von Vanessa Lang u. a. Erstellt in: Schmidt, Jürgen Erich/Herrgen, Joachim/Kehrein, Roland/

Lameli, Alfred/Fischer, Hanna (Hg.): Regionalsprache.de (REDE III). Marburg: Forschungszentrum Deutscher Sprachatlas.

Schmidt, Jürgen Erich/Herrgen, Joachim (2011): Sprachdynamik. Eine Einführung in die Regionalsprachenforschung. Berlin: ESV.

Schmidt, Jürgen E./Beitel, Dennis/Frank, Marina/Gerstweiler, Luisa/Lang, Vanessa (2023): Der Digitale hessische Sprachatlas (DHSA). Einführung und sprachgeschichtliche Interpretation am Beispiel des Zusammenfalls von Alt- und Neudiphthongen. In: Zeitschrift für Dialektologie und Linguistik 90, 1, S. 97–139.

Schmidt, Jürgen E./Herrgen, Joachim/Kehrein, Roland/Lameli, Alfred (Hg.) (2020f): Regionalsprache.de. Forschungsplattform zu den modernen Regionalsprachen des Deutschen. Bearbeitet von Robert Engsterhold, Hanna Fischer, Marina Frank, Heiko Girnth, Simon Kasper, Juliane Limper, Salome Lipfert, Georg Oberdorfer, Tillmann Pistor, Anna Wolańska. Unter Mitarbeit von Dennis Beitel, Lisa Dücker, Lea Fischbach, Milena Gropp, Heiko Kammers, Maria Luisa Krapp, Vanessa Lang, Salome Lipfert, Jeffrey Pheiff, Bernd Vielsmeier. Studentische Hilfskräfte. Marburg: Forschungszentrum Deutscher Sprachatlas.

Vorberger, Lars (2019): Regionalsprache in Hessen. Eine Untersuchung zu Sprachvariation und Sprachwandel im mittleren und südlichen Hessen. (= Zeitschrift für Dialektologie und Linguistik – Beihefte 178). Stuttgart: Steiner.

Wenker, Georg (1889–1923): Sprachatlas des Deutschen Reichs. Handgezeichnet von Emil Maurmann, Georg Wenker & Ferdinand Wrede. Marburg: Forschungszentrum Deutscher Sprachatlas.

Wenkerbogen-App (2023 f.): Herausgegeben vom Forschungszentrum Deutscher Sprachatlas, Marburg. https://apps.dsa.info/wenker/ (Stand: 3.12.2025).

Wrede, Ferdinand (1895): Über richtige Interpretation der Sprachatlaskarten. Vortrag gehalten in der germanistischen Section der 43. Versammlung deutscher Philologen und Schulmänner zu Köln am 26. September 1895. In: Wenker, Georg/Wrede, Ferdinand (Hg.): Der Sprachatlas des Deutschen Reichs. Dichtung und Wahrheit. Marburg: Elwert, S. 31–52.

Zwirner, Eberhard (1956): Lautdenkmal der deutschen Sprache. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 9, 1, S. 3–13.

Kontaktinformation

Dennis Beitel
Forschungszentrum Deutscher Sprachatlas
Philipps-Universität Marburg
Pilgrimstein 16
35027 Marburg
E-Mail: dennis.beitel@uni-marburg.de

Lisa Dücker
Forschungszentrum Deutscher Sprachatlas
Philipps-Universität Marburg
E-Mail: lisa.duecker@uni-marburg.de

Salome Lipfert
Forschungszentrum Deutscher Sprachatlas
Philipps-Universität Marburg
E-Mail: salome.lipfert@uni-marburg.de

Georg Oberdorfer
Forschungszentrum Deutscher Sprachatlas
Philipps-Universität Marburg
E-Mail: georg.oberdorfer@uni-marburg.de

Bibliografische Angaben

Dieser Text ist Teil der Publikation: Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag. https://doi.org/10.21248/idsopen.16.2026.63.

Bibliografische Informationen

Angaben zur Zitierung dieser Publikation:

Brunner, Annelen/Hansen, Sandra/Lang, Christian/Tu, Ngoc Duyen Tanja/Wolfer, Sascha (Hg.) (2026): Deutsch im Wandel – Tools, Methoden, Ressourcen. Beiträge zur Methodenmesse der IDS-Jahrestagung 1. (= IDS*open* 16). Mannheim: IDS-Verlag.

DOI https://doi.org/10.21248/idsopen.16.2026.63

Impressum

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

IDS-Verlag · Leibniz-Institut für Deutsche Sprache
R 5, 6–13 · 68161 Mannheim
www.ids-mannheim.de

IDS-Verlag

Schriftenreihe: IDS*open*: Online-only Publikationen des Leibniz-Instituts für Deutsche Sprache
Reihenherausgeber/-innen: Norman Fiedler, Katrin Hein-Antonioli, Siegwalt Lindenfelser, Beata Trawiński
Redaktion: Melanie Kraus
Satz: Carolin Häberle

ISBN: 978-3-948831-80-6 (PDF)

ISSN: 2749-9855