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).

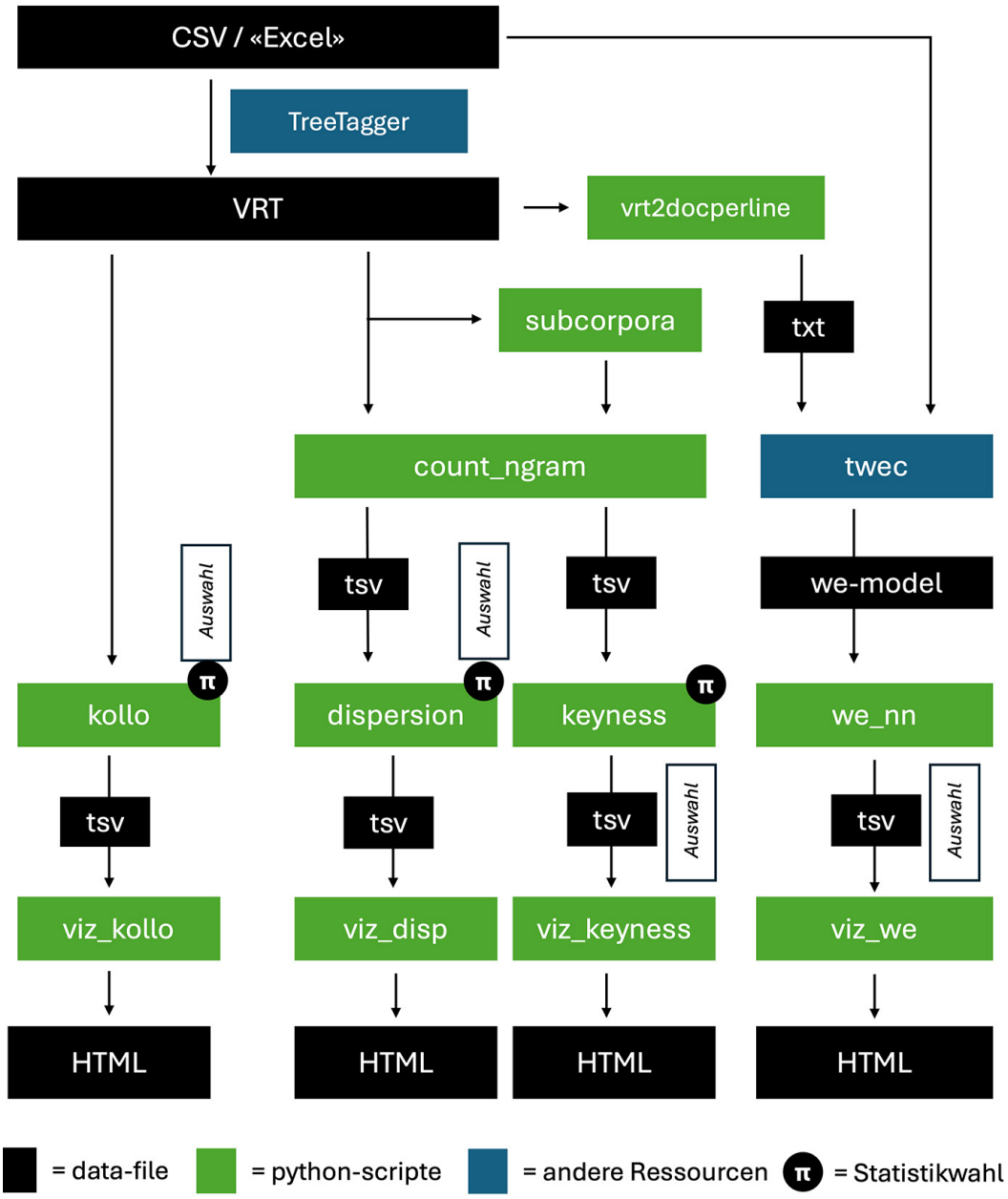

Abb. 1: Schematische Workflows der Toolbox *Visuelle Korpuspragmatik*

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).

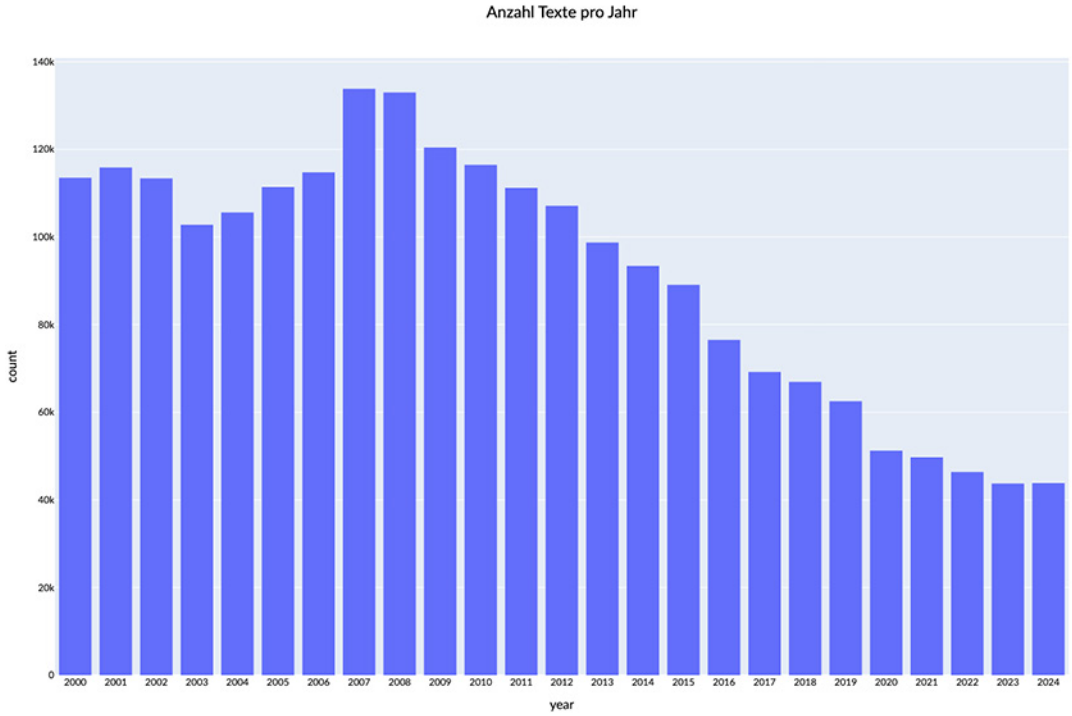

Abb. 2: Anzahl Texte (in Tausend) pro Jahr [Interaktive Version der Abb.][7]

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.

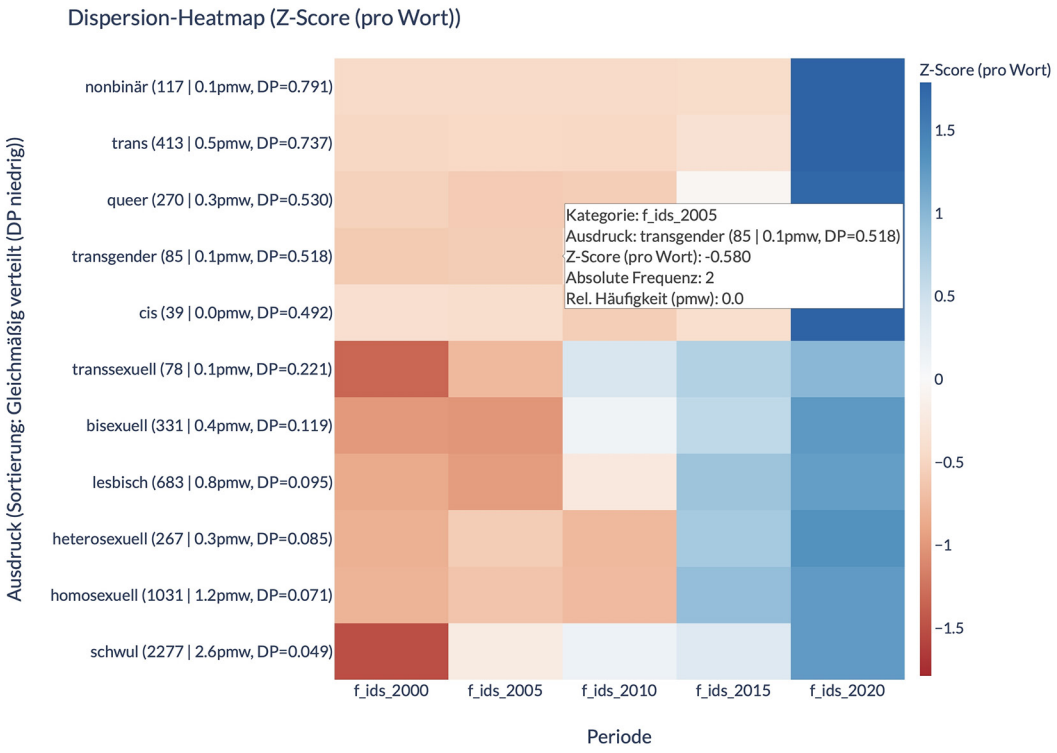

Abb. 3: Dispersionsvisualisierung von Ausdrücken rund um sexuelle Orientierung (Z-Score) [Interaktive Version der Abb.]

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.

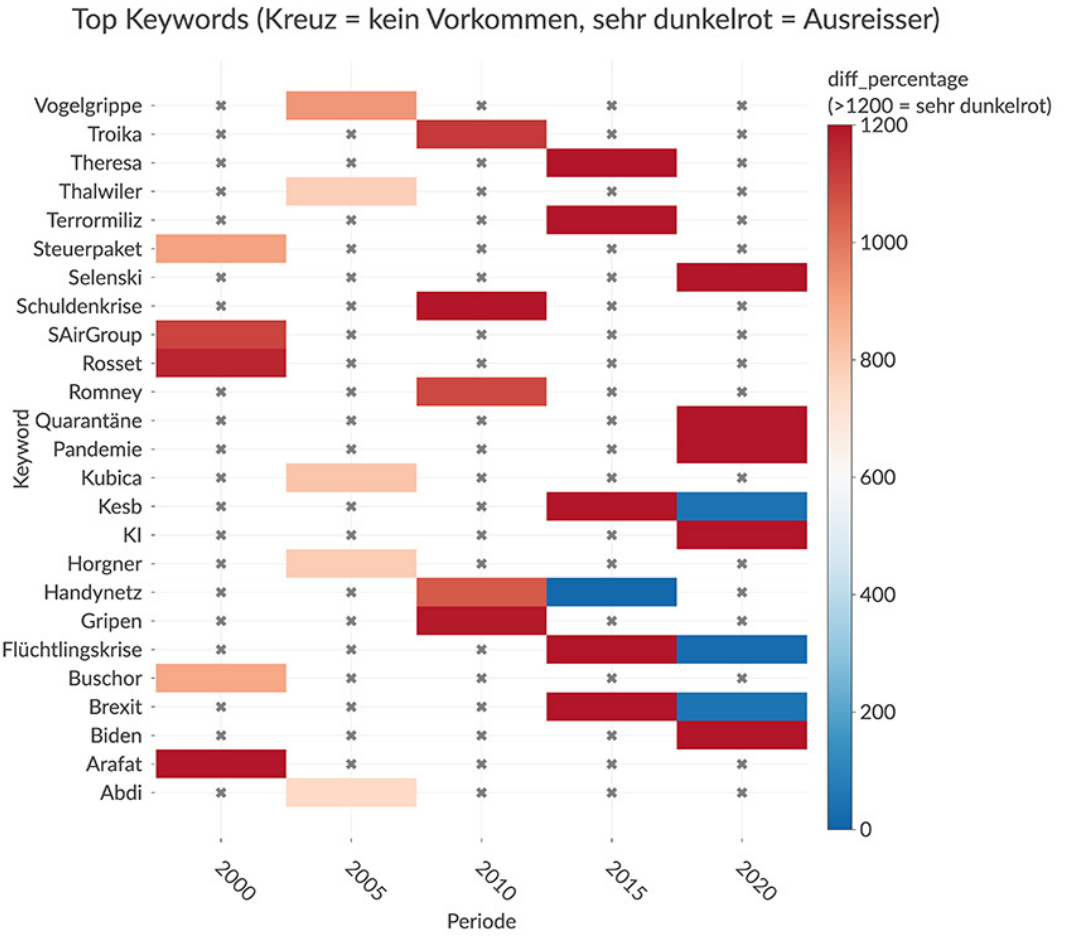

Abb. 4: Keynessvisualisierung als Heatmap (Top 5 Keywords pro Periode) [Interaktive Version der Abb.]

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).

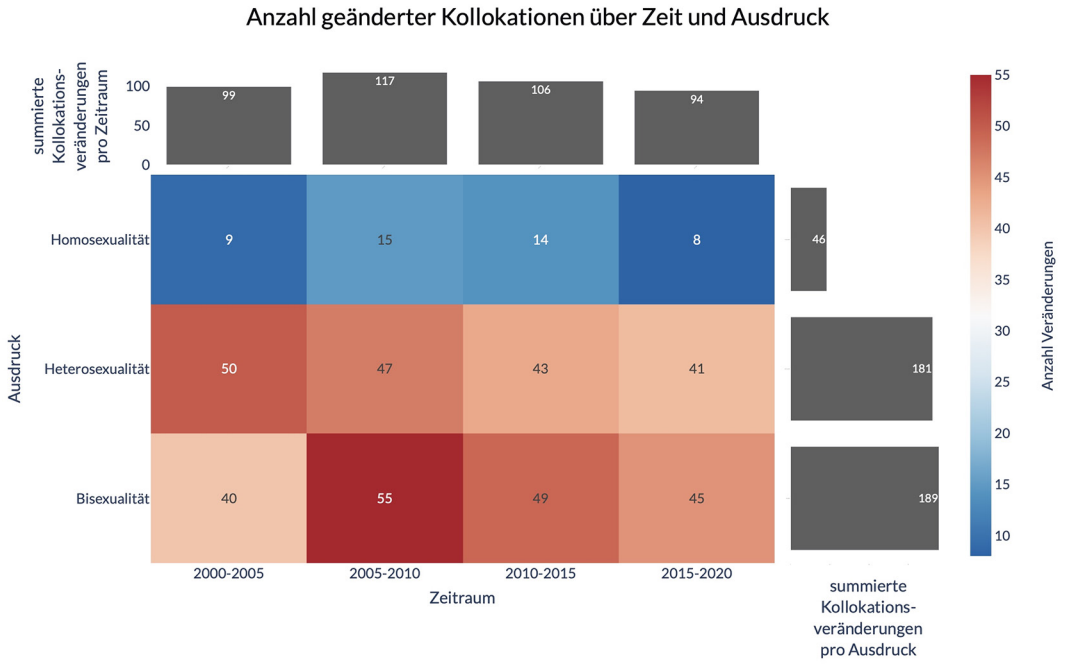

Abb. 5: Anzahl geänderter Kollokationen über Zeit und Ausdruck [Interaktive Version der Abb.]

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.

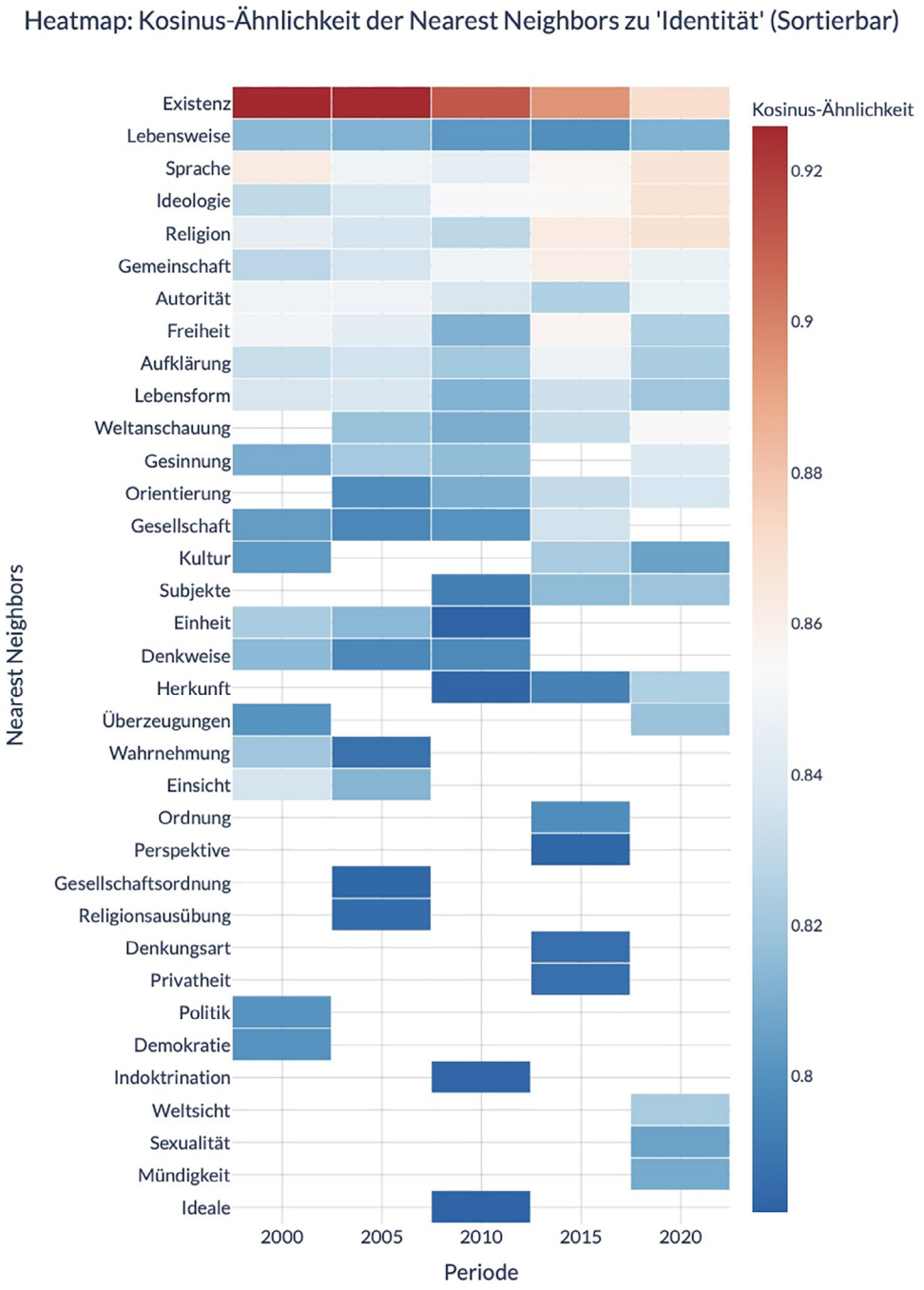

Abb. 6: Word Embedding-Visualisierung Nearest Neighbors zu *Identität* [Interaktive Version der Abb.]

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.