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.

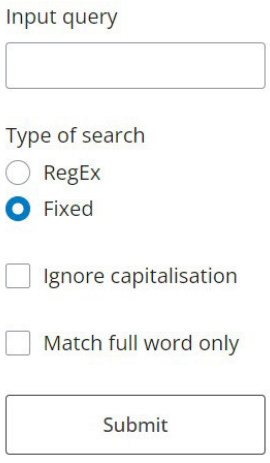

Abb. 2: Die Input-Optionen von ReCKS

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.

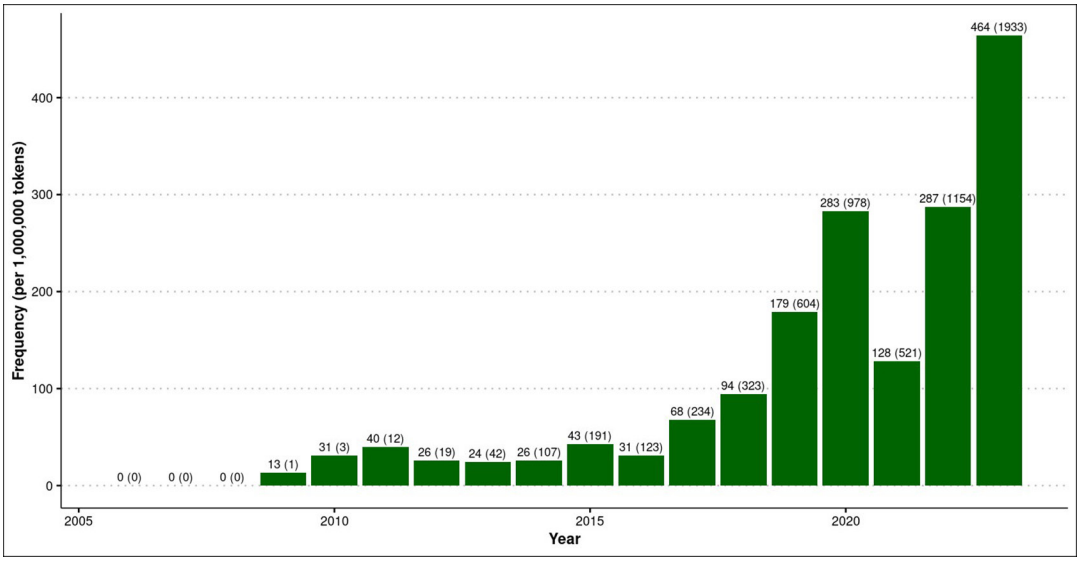

Abb. 4: Histogramm-Visualisierung der Ergebnisse für die RegEx-Suche <\b(k|K)lima[:alpha:]{3,}>

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.

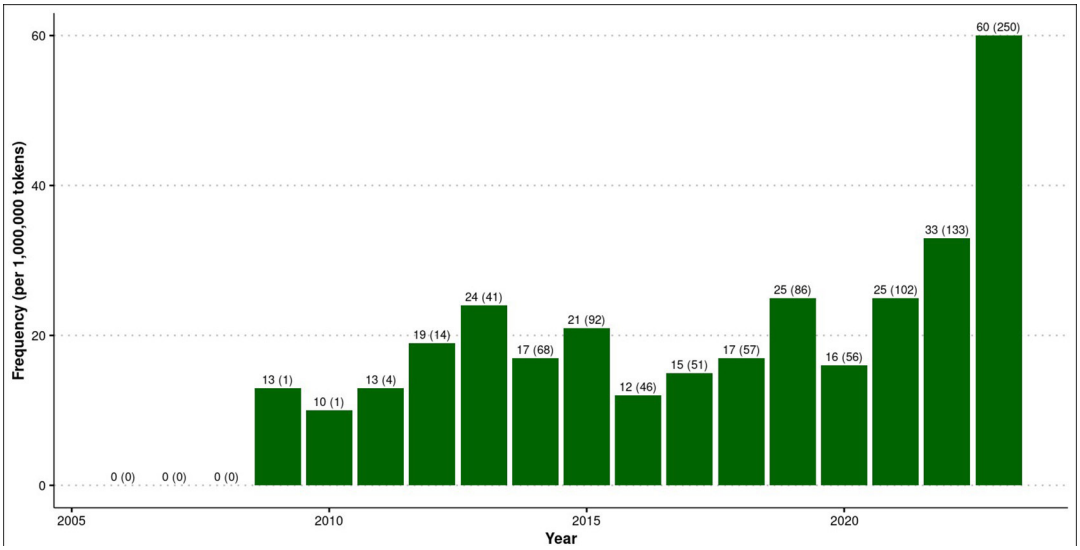

Abb. 6: Diachrone Verteilung der Binnen-I-Variante (normalisierte Häufigkeit pro Jahr)

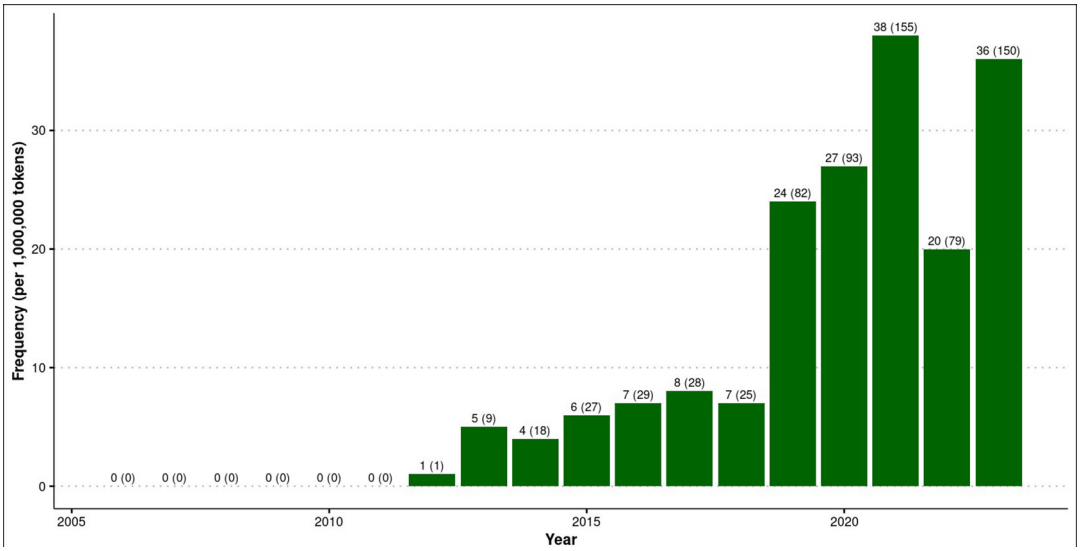

Abb. 7: Diachrone Verteilung der Sternchen-Variante (normalisierte Häufigkeit pro Jahr)

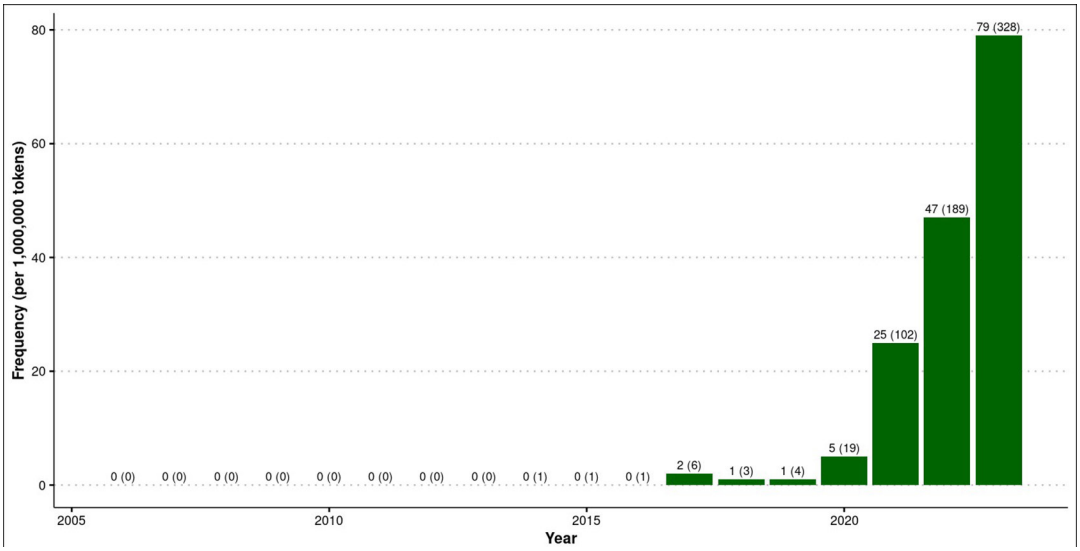

Abb. 8: Diachrone Verteilung der Doppelpunkt-Variante (normalisierte Häufigkeit pro Jahr)

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.