

VOM FUNDUS ZUM KORPUS: Reddit als Medium und digitale Sprachressource

Daniel Pfurtscheller
Universität Innsbruck

Abstract

Der Beitrag untersucht die Nutzung von Reddit als Medium und digitale Sprachressource im Kontext von DaF/DaZ und stellt verschiedene Ansätze zur Nutzung von Reddit als Korpus vor. Anhand einer Fallstudie, die sprachliche Variationen und Diskurspraktiken in den deutschsprachigen Subreddits *r/Austria* und *r/de* vergleicht, werden thematische und sprachliche Unterschiede herausgearbeitet und hinsichtlich ihrer Implikationen für den DaF/DaZ-Unterricht diskutiert. Insgesamt unterstreicht der Beitrag die Bedeutung einer systematischen Analyse und didaktischen Aufbereitung der auf Reddit verfügbaren sprachlichen Ressourcen, um das Verständnis für sprachliche Variation und kulturelle Unterschiede im digitalen Zeitalter zu fördern.

Keywords: Social Media; digitale Sprachressource; Korpuslinguistik; internetbasierte Kommunikation; sprachliche Variation

Abstract

This paper explores the use of Reddit as a medium and digital language resource in the context of DaF/DaZ and outlines different approaches to the use of Reddit as a corpus. Based on a case study comparing linguistic variations and discourse practices in the German-language subreddits *r/Austria* and *r/de*, thematic and linguistic differences are highlighted and discussed with regard to their implications for DaF/DaZ teaching. Overall, the article underlines the importance of a systematic analysis and didactic elaboration of the linguistic resources available on Reddit in order to promote the understanding of linguistic variation and cultural differences in the digital age.

Keywords: social media; digital language resource; corpus linguistics; internet-based communication; linguistic variation

1. Einleitung

Die Relevanz authentischer Sprachmaterialien wird im Lehr-Lernkontext Deutsch als Fremd- und Zweitsprache (DaF/DaZ) seit geraumer Zeit diskutiert (eine kritische Übersicht gibt Riedner 2018). Bemerkenswert ist, wie das Internet und andere (als ‚neu‘ titulierte) Medien als Materialquellen in den Vordergrund gerückt werden. So schreiben etwa Nodari / Steinmann (2010: 1160): „Insbesondere durch das Internet steht heute eine große Zahl leicht zugänglicher, authentischer zielsprachlicher Materialien zur Verfügung.“ Für den DaF-/DaZ-Unterricht erscheint eine solche Vorstellung, die das Internet als eine Art digitaler Textfundus konzipiert, aus mehreren Gründen problematisch: Neben rechtlichen und ethischen Fragen, die die Verfügbarkeit und Nutzung von Online-Materialien einschränken, vernachlässigt eine solche Vorstellung die medialen und sprachlichen Besonderheiten internetbasierter Kommunikation (z.B. Chats, Forendiskussionen, Social-Media-Postings etc.). Zudem kann man Authentizität als Konstrukt selbst in Frage stellen, da es letztlich von der Perspektive und den Zielen der Lernenden und Lehrenden abhängt, was als authentisch angesehen wird (vgl. Gilmore 2007).

Ausgangspunkt meines Beitrags ist daher ein scheinbarer Widerspruch: Einerseits steht außer Frage, dass man das Internet – genauer gesagt Sprachdaten, die bei der Nutzung webbasierter Anwendungen entstehen (vgl. Ebersbach et al. 2016: 30) – als Korpusressourcen nutzen kann. Die Korpuslinguistik kennt das Prinzip des Web-as-a-Corpus seit langem (vgl. z.B. Gatto 2014) und unterscheidet verschiedene Arten der Korpuserstellung und -nutzung (vgl. Bernardini / Baroni / Evert 2006), die von umfangreichen Web-Korpora bis hin zu kleineren, fokussierten Textsammlungen reichen (sogenannte *disposable corpora*, vgl. Varantola 2003). Andererseits wäre es aus medienlinguistischer Sicht jedoch verfehlt, sprachliches Handeln im Internet zu untersuchen, ohne Aspekte der Medialität (vgl. Schneider 2018; Pfurtscheller i.E.) und Verankerung in spezifischen Online-Communities und Diskursgemeinschaften (vgl. Spitzmüller / Warnke 2011: 181; Naef 2021) zu reflektieren. Ebersbach et al. (2016: 30) stellen fest: „Social-Web-Anwendungen ohne zugehörige Community sind nicht denkbar, sie gehört als zwingende Voraussetzung dazu, um diesen Titel tragen zu können“.

Die Forschung zur internetbasierten Kommunikation hat auf die Charakteristika und Anforderungen von Social-Media-Kommunikation aufmerksam gemacht und die Entwicklung spezifischer Korpuswerkzeuge und Bewertungsmaßstäbe gefordert. Storrer (2018) hat auf Besonderheiten der interaktionsorientierten Schreibhaltung hingewiesen, die Social-Media-Daten zu einem „Korpustyp der dritten Art“ (Storrer 2014: 189) machen, der sich in zentralen Eigenschaften von Text- und Gesprächskorpora unterscheidet. Beißwenger (2020) hat in der Folge dafür plädiert, internetbasierte Kommunikation nicht länger als Sonderfall, sondern als dritte Organisationsform sprachlichen Handelns zu konzipieren, „die die Qualitäten von Textformen adaptiert, um situationsentbundene, sequenziell intendierte Kommunikation zu ermöglichen“ (Beißwenger 2020: 297). Für die internetbasierte Kommunikation sind eigene Sprachnormen anzusetzen, die auch im Unterricht bei der Beurteilung von stilistischer Angemessenheit und Textqualität berücksichtigt werden sollten. So haben Storrer (2020) und Linthe (2020) zuletzt Vorschläge gemacht, wie man das traditionelle Zürcher Textanalyseraster erweitern kann, um „die Konformität zu Wert und Qualitätsvorstellungen für das Schreiben in Sozialen Medien“ (Storrer 2020: 104) und die „Plattformspezifik“ (Linthe 2020: 135) sprachlichen Handelns in Social-Media-Umgebungen zu berücksichtigen.

Die Möglichkeiten und Herausforderungen, die sich bei Nutzung von Social-Media-Daten im Sprachunterricht ergeben können, fokussiere ich im Folgenden am Beispiel der Online-Plattform Reddit, einer der weltweit größten Websites für Nutzer:innen-generierte Inhalte. Reddit bietet im DaF-/DaZ-Kontext interessante Nutzungsmöglichkeiten, setzt aber auch spezifisches Wissen voraus. Damit ein sinnvoller und nicht überfordernder Unterricht erfolgen kann, müssen aus den scheinbar ‚authentischen‘ Sprachdaten auf Reddit ‚didaktisierte‘ oder ‚didaktische‘ Texte (vgl. Thonhauser / Hufeisen 2016) werden. Dafür ist nicht nur Korpusarbeit, sondern auch Textarbeit (vgl. Thonhauser 2020) erforderlich. Dafür stellen sich eine Reihe von Fragen:

- Wie lassen sich relevante und repräsentative Texte und Diskussionen auf Reddit identifizieren und auswählen, die für den DaF-/DaZ-Unterricht geeignet sind?
- Welche methodischen Ansätze und Analysewerkzeuge können eingesetzt werden, um sprachliche Merkmale und Besonderheiten der auf Reddit gefundenen Texte systematisch zu erfassen und zu beschreiben?
- Wie können die auf Reddit identifizierten sprachlichen Phänomene und Variationen didaktisch aufbereitet und in den Unterricht integriert werden, um die Lernenden beim Erwerb von kommunikativen Kompetenzen im digitalen Kontext zu unterstützen?
- Inwiefern müssen Lehrkräfte und Lernende für die spezifischen Kommunikationsnormen und -stile von Reddit sensibilisiert werden, um eine angemessene Einschätzung von Sprachgebrauch und Textqualität im Internet zu ermöglichen?
- Welche Rolle spielen die sozialen und kulturellen Kontexte der auf Reddit gefundenen Diskussionen und Texte für das Verständnis und die Analyse der sprachlichen Ressourcen, die auf der Plattform verfügbar sind?

- Wie können die auf Reddit gefundenen sprachlichen Ressourcen genutzt werden, um das Bewusstsein der Lernenden für Sprachvariation, kulturelle Unterschiede und die Dynamik des Sprachwandels im digitalen Zeitalter zu fördern?

In diesem Beitrag konzentriere ich mich vor allem auf die medien- und korpusbezogenen Grundlagen, gehe aber auch auf einige fachdidaktische Aspekte der aufgeworfenen Fragen ein. Ziel des Beitrags ist es, die Möglichkeiten und Herausforderungen der Verwendung von Reddit im DaF/DaZ-Unterricht aufzuzeigen. In den folgenden Abschnitten werden dafür zunächst die Grundlagen von Reddit und seine Relevanz als Lern- und Lehrmedium für den DaF-/DaZ-Unterricht erörtert (Abschnitt 2). Anschließend stelle ich im Überblick dar, wie Reddit als Korpus für sprachliche Untersuchungen und Übungen im DaF-/DaZ-Unterricht genutzt werden kann (Kap. 3). Mit einer Fallstudie demonstriere ich dann exemplarisch, wie sich Korpora nutzen lassen, um Sprach- und Diskurspraktiken auf Reddit zu untersuchen, und diskutiere die Frage, welche Implikationen sich daraus für den DaF/DaZ-Unterricht ergeben (Abschnitt 4). Das abschließende Fazit diskutiert Herausforderungen und Lösungsansätze im Umgang mit Reddit-Korpora im DaF-/DaZ-Unterricht und gibt einen Ausblick auf künftige Entwicklungen (Abschnitt 5).

2. Reddit als Lehr- und Lernmedium

Reddit.com ist 2005 gestartet und gilt vor allem in den USA als „Frontpage of the internet“ (so der eigene Wahlspruch). Weltweit zählt es zu den zehn meistgenutzten Websites (vgl. Semrush 2023). Im Jahr 2021 hat das Unternehmen in Berlin seine erste nicht-englischsprachige Niederlassung gegründet (vgl. Reddit 2021). Auch wenn die Zahl der deutschen Nutzer:innen im Vergleich zu Facebook, Instagram und Co. eher gering ist (vgl. Koch 2022), verfügt Reddit doch über eine beträchtliche deutschsprachige Community. Laut eigenen Angaben tragen deutschsprachige Nutzer:innen rund 160 Millionen eigene Posts, Kommentare und Interaktionen bei – pro Monat wohlgermerkt (vgl. Reddit 2022).

Versteht man Lehr- und Lernmedien in einem weiten Sinn als Oberbegriff für „alle in irgendeiner materiellen Form vorliegenden Dinge [...], die zum Fremdsprachenlehren und/oder -lernen eingesetzt werden“ (Würffel 2021: 282), kann man Reddit als digitales Medium betrachten, das Lernende und Lehrende im Sprachunterricht auf verschiedene Art und Weise unterstützen kann. Im Folgenden gebe ich zunächst einen Überblick über die Medienumgebung von Reddit, indem ich die medialen Affordancen der Website und seine „stoffliche Gestalt“ (Würffel 2021: 287) auf zwei zentralen Ebenen beschreibe und visualisiere: der Überblickseite eines exemplarischen Reddit-Forums (Abschnitt 2.1) und den Kommentarverlauf zu einem Forum-Beitrag (Abschnitt 2.).

2.1. Subreddits als thematische Unterforen und vernetzte Online-Communities

Reddit ist eine Nachrichten-Website und ein Online-Forum, deren Inhalte von den Mitgliedern stammen und gemeinschaftlich kuratiert werden. Der Name ist eine Wortspielerei mit engl. *I read it*, ‚ich hab’s gelesen‘ (vgl. Reddit 2014). Reddit kann man sich als eine Art vernetztes Über-Forum vorstellen: Die Website besteht aus einer Vielzahl von thematischen Unterforen, den so genannten *Subreddits*. Jedes Subreddit hat ein bestimmtes Thema (z.B. im Bereich Technik, Politik oder Musik) und kann gezielt aufgerufen werden, indem man „reddit.com/r/subreddit_name“ eingibt. Laut eigenen Angaben zählen zu den beliebtesten deutschsprachigen Subreddits wie r/Arbeitsleben, r/de, r/de_1AmA, r/finanzen, r/fragreddit, r/ich_iel und andere (vgl. Reddit 2022).

Subreddits sind selbstverwaltet und mehr oder weniger stark als Online-Communities und Diskursgemeinschaften ausgestaltet (zum Verhältnis dieser beiden Begriffe vgl. Naef 2021). Neben einem eigenen visuellen Design legen die Nutzer:innen dabei auch die Regeln für ihre Foren fest und übernehmen die Moderation. Nutzer:innen können mit einem Account mehreren Communities folgen und passiv oder aktiv partizipieren, indem sie neue Beiträge erstellen sowie veröffentlichte Beiträge lesen, bewerten und kommentieren. Eine Besonderheit der Plattform, die von anderen Forensystemen übernommen wurde, ist das Bewertungssystem: Nutzer:innen können die Sichtbarkeit einzelner Beiträge und Kommentare bestimmen, indem sie diese jeweils hoch- oder herunterstufen (sog. *upvoting* und *downvoting*). Für den oder die Nutzer:in, die den Post oder Kommentare beigetragen hat, ergeben sich daraus sogenannte Karma-Punkte. Auf diese Weise können Beiträge auch aus den verschiedenen Unterforen auf die Hauptseite von Reddit (die sog. *Frontpage*) gelangen. Dieses Bewertungssystem bestimmt somit übergreifend die Sichtbarkeit und Relevanz von Beiträgen, auch über die einzelnen Subreddits hinaus (vgl. Reddit 2014).

Abbildung 1 zeigt diese multimodalen Affordanzen am Beispiel von *r/FragReddit*. Zu sehen ist die Übersichtsseite des Subreddits, auf der die eingereichten Forenbeiträge (Posts) angezeigt und nach verschiedenen Kriterien sortiert werden können. Erkennbar sind die Möglichkeiten der Individualisierung: Nutzer:innen wählen ein eigenes Farbdesign und eine Header-Grafik für das Forum aus und bestimmen so das visuelle Design ihrer Community mit. Zentrale Elemente sind in der rechten Randspalte zu finden, wo mehrere Module über das Forum selbst, seine Regeln und die Moderator:innen informieren. Diese Grundelemente gehören zu den Standards des Plattformdesign (vgl. Pfurtscheller i.E.). Im Fall von *r/FragReddit* sind hier einige Punkte bemerkenswert: So wird in der Infobox explizit auf ein anderes Forum verwiesen und das Forum als deutschsprachiges Pendant einer anderen großen Community markiert („vgl. */r/AskReddit* – hier in der deutschsprachigen Version“). Die acht Regeln betreffen Ausrichtung und Ziel des Forums (Regel 1: „Dieses Subreddit ist für interessante, hintergründige oder provozierende Fragen“), aber auch die Komplexität der Fragen und den Sprachgebrauch (Regel 4: „Der Titel muss eine grammatikalisch korrekte Frage sein, die mit einem Fragezeichen endet“).

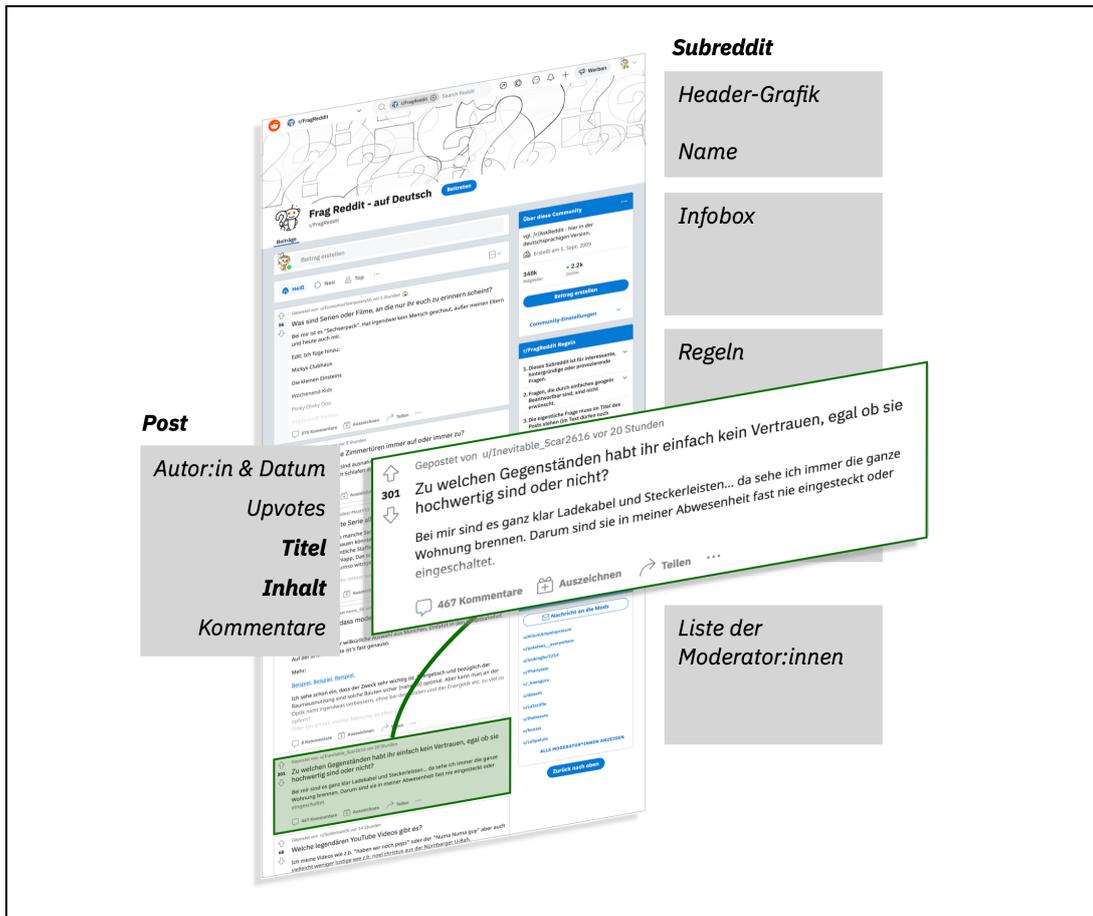


Abbildung 1

Screenshot der Startseite von r/FragReddit, hervorgehoben werden die Bestandteile und der Aufbau der Foren-Website

Zentrale Ressourcen sind dafür die einzelnen Posts des Forums. Der in Abbildung 1 hervorgehobene Post hat zum Zeitpunkt der Aufzeichnung 301 positive Bewertungen (*Upvotes*) und 467 Kommentare erhalten. Bevor wir näher auf diesen Post und seine Anschlusskommunikation eingehen, können wir zusammenfassend drei zentrale Ebenen von Reddit festhalten, die Reddit als Website und digitale Mediumgebung ausmachen:

1. *Subreddits* als thematische Unterforen, die von den Nutzer:innen selbstständig angelegt und verwaltet werden können. Reddit bietet dafür Moderationswerkzeuge, Gestaltungsmöglichkeiten für das visuelle Erscheinungsbild sowie weitere Community-Tools an (wie z.B. Wikis).
2. *Posts* als Beiträge in Subreddits. Alle Reddit-Posts haben einen Titel und einen Inhalt. Neben reinen Textbeiträgen sind auch Links, Bilder und Videos als Inhalte möglich. Bei genügend positiven Bewertungen (*Upvotes*) können Posts auch außerhalb der Unterforen auf der Reddit-Startseite landen und damit viel Reichweite gewinnen.
3. *Kommentare* als schriftliche Anschlusskommunikation, die durch sequenzielle Kommentar-auf-Kommentar-Einrückungen beliebig tiefe Diskussionsstränge ermöglichen. Auch Kommentare können von den Nutzer:innen positiv oder negativ bewertet werden.

2.2 Kommentare, Kontexte und Interaktionsbedingungen auf Reddit

Als digitale Medienumgebung ermöglicht Reddit Formen des sprachlichen Austauschs, die man als interpersonal-öffentliche Kommunikation (vgl. Haas / Brosius 2011) auffassen kann. Die Kommunikation auf Reddit ist asynchron und hauptsächlich schriftlich, mit gelegentlichen Bild- und Videoinhalten. Anonymität wird durch die Verwendung von Pseudonymen ermöglicht und gefördert. Daraus ergeben sich Interaktionsbedingungen, die mit Kommunikation in traditionellen Online-Foren vergleichbar ist (vgl. Klemm 2012; Pappert / Roth 2016; Kaltwasser 2019; Pappert / Roth 2019).

Ich möchte diese Merkmale und Interaktionsbedingungen im Folgenden am Beispiel des oben bereits eingeführten r/FragReddit veranschaulichen und vertiefen. Dort veröffentlicht ein:e Nutzer:in mit dem Pseudonym *Inevitable_Scar2616* am späten Nachmittag des 1. April 2023 folgende Frage: „Zu welchen Gegenständen habt ihr einfach kein Vertrauen, egal ob hochwertig oder nicht?“. Der Beitrag wurde von mehr als 301 Personen hochgestuft und erhielt 467 Kommentare. In Abbildung 2 ist eine Visualisierung dieses Posts und aller Antwortkommentare zu sehen. Die Kommentierungen sind in diesem *distant viewing* chronologisch sortiert und eingefärbt: Rot markiert sind die obersten Top-Level-Kommentare, schwarz die Folgekommentare. Deutlich zu erkennen ist, dass im Zeitverlauf später veröffentlichte Kommentare weniger Folgekommentare auslösen. Grundsätzlich erlaubt Reddit Asynchronität: Nutzer:innen können Beiträge und Kommentare verfassen, wann immer sie möchten, und andere können ohne zeitliche Einschränkungen darauf reagieren. Das Beispiel zeigt aber auch, dass ein Großteil der Kommentare in einem Zeitfenster von ca. zwei Stunden nach der Veröffentlichung geschrieben wurde. Antworten auf die gestellte Frage, die einen Tag später kommen, werden weniger beachtet und diskursiv bearbeitet als solche, die zeitlich sehr nahe am Veröffentlichungszeitpunkt liegen. Die Visualisierung verdeutlicht weiter die hierarchische Struktur, die im Beispiel bis zu 11 Antwortebenen umfasst. Diese hierarchische Struktur in Verbindung mit der Asynchronität ermöglicht es, dass Interaktionen zeitversetzt und gleichzeitig an unterschiedlichen Stellen stattfinden. Jeder Kommentar ist seinerseits kommentierbar und mit einem permanenten Link versehen. Dass die Nutzer:innen über Antworten benachrichtigt werden, gehört zu den weiteren Anforderungen der Plattform.

Die Kommunikation auf Reddit ist in mehrfacher Hinsicht kontextgebunden, insofern Posts innerhalb themenspezifischer Subreddits veröffentlicht werden (Kontextebene 1: Subreddit als thematisches Forum), Kommentare sich auf Posts beziehen (Kontextebene 2: Posts als Bezugsgröße) und in einer potenziell endlosen sequenziell-hierarchischen Anschlusskommunikation eingebunden sind (Kontextebene 3: Kommentarsequenz).

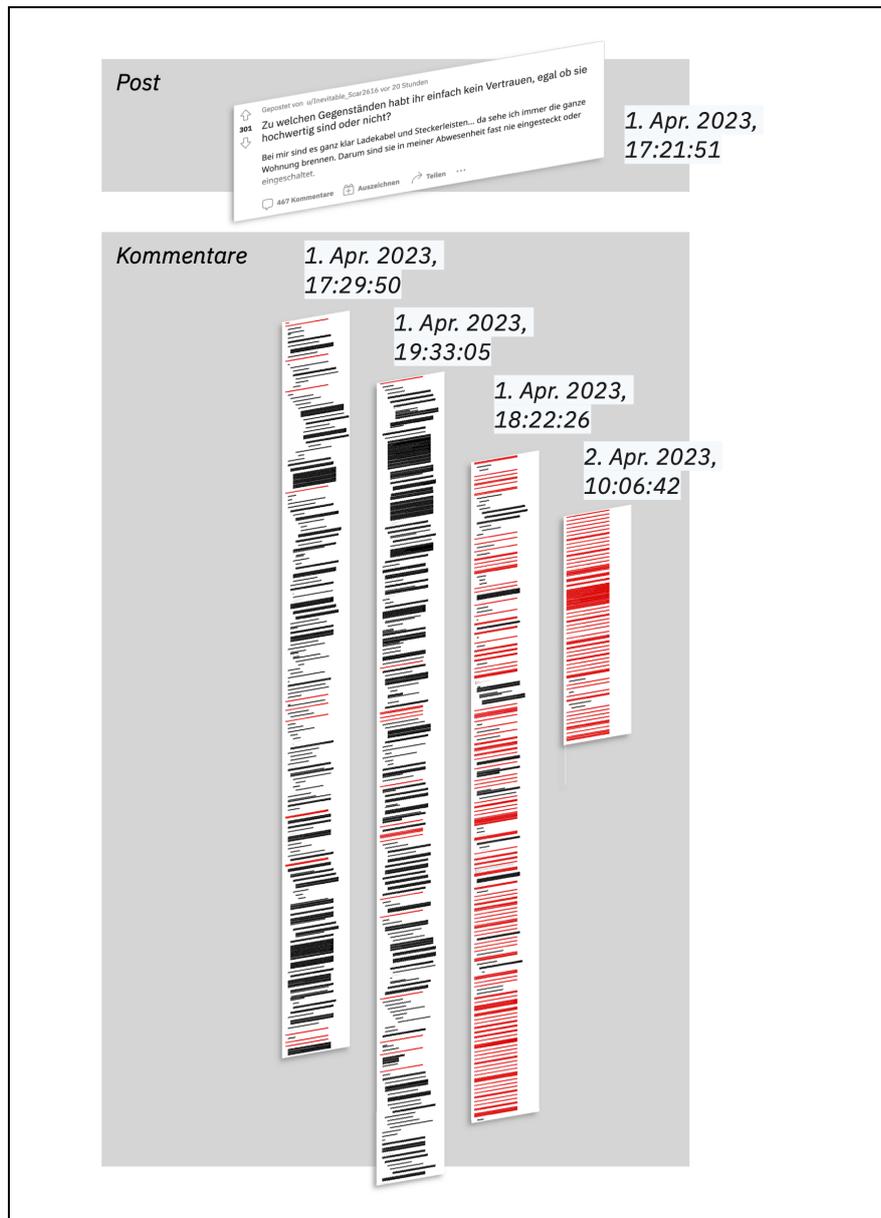


Abbildung 2

Visualisierung einer Kommentarsequenzen zu einem Reddit-Post auf r/FragReddit (rot: initialer Kommentar, schwarz Folgekommentar)

Im Kontext von r/FragReddit gibt es eine recht klare dyadische Struktur: In der Regel beziehen sich die initialen Top-Level-Kommentare direkt auf die gestellte Frage und haben damit den Status von Antworten. Im Beispiel wird das nicht zuletzt durch die sprachliche Kürze deutlich (die Angaben in Klammern beziehen sich auf: Account, Timestamp in CEST, Punkte, Folgekommentare):

- (1) Klettergeschirr. Das Leben hängt an einem dünnen Seil. Neeee... Nicht mein Ding. (restwasserschale, 2023-04-01 17:29:50, 46 P, 7 FK)
- (2) Kuchen Transportboxen. Ich nehm die immer unten drunter, weil ich den henkeln nicht vertraue (Celeana23, 2023-04-01 17:41:59, 1280 P, 48 FK)
- (3) Leitern (This_Pumpkin_4331, 2023-04-01 17:44:08, 103 P, 1 FK)
- (4) Diese Klimmzugstangen, die man in den Türrahmen spannt. (astarch, 2023-04-01 17:51:13, 847 P, 27 FK)

- (5) Trägerlose BHs - egal wie hochwertig ,ich hätte immer Angst plötzlich oben ohne dazustehen (dann lieber transparente Träger) (distelwaldweg, 2023-04-01T18:05:28, 201 P, 12 FK)
- (6) Drucker... Das Brauch ich denke ich nicht weiter auszuführen. (ActuaryCute3771, 2023-04-01 18:22:26, 448 P, 49 FK)

Wie schon an diesen Belegen deutlich wird, ist der Sprachgebrauch auf Reddit in der Regel eher informell und interaktionsorientiert (vgl. Storrer 2018): Nutzer:innen beteiligen sich an Diskussionen, stellen Fragen, geben Antworten und kommentieren die Beiträge anderer Nutzer:innen. Sprachlich zeigt sich diese Interaktionsorientierung an der Kürze und Kontextgebundenheit der Kommentare: Am Kommentaranfang wird auf bestimmte Gegenstände Bezug genommen (*Klettergeschirr, Kuchen Transportboxen, Leitern, Diese Klimmzugstangen, Trägerlose BHs, Drucker*). Jede dieser Nennungen ist auf die ursprüngliche Frage nach Gegenständen, denen man nicht vertraut, bezogen und nur in diesem Kontext verständlich. Das sprachliche Handeln ist im Sinne diese Art der Interaktion gestaltet, bei der es weniger darum geht, vollständig ausformulierte und unabhängig verständliche Texte zu schreiben, sondern vielmehr darum, einen Beitrag zum Forentheema zu verfassen und die Interaktion im Forum am Laufen zu halten. Dabei finden sich viele Ausdrucksweisen, die man aus Alltagsgesprächen kennt. Im Unterschied zu mündlichen Gesprächen ist jedoch sowohl die initiale Frage, als auch die Antworten nicht an einzelne Gesprächspartner:innen gerichtet, sondern an alle Nutzer:innen, die an dieser Frage interessiert sind und sich im Forum beteiligen wollen.

Ähnlich wie Wikipedia ist Reddit eine Plattform, auf der beide Schreibweisen vertreten sind. Auf Reddit finden sich neben interaktionsorientierten Kommentaren und Diskussionen, die auf schnelle Reaktionen und das Gelingen der laufenden Interaktion abzielen, auch textorientierte Beiträge, wie z.B. längere Posts oder Wiki-Einträge. Diese Mischung aus Text- und Interaktionsorientierung ermöglicht es, Reddit als Lehr- und Lernmedium flexibel einzusetzen. Im DaF-/DaZ-Unterricht kann das Konzept des interaktionsorientierten Schreibens dazu genutzt werden, um Lernende mit den Besonderheiten der internetbasierten Kommunikation vertraut zu machen und ihre Fähigkeiten in verschiedenen Kommunikationskontexten und -stilen zu fördern. Reddit bietet aus linguistischer Sicht damit eine Fülle von Möglichkeiten für den DaF-/DaZ-Unterricht und seiner Nutzung als Sprachressource im Kontext eines datengeleiteten Lernens (vgl. Johns 1991; Boulton 2017), das auf der Idee basiert, dass Lernende eine Sprache besser verstehen und erlernen, wenn sie selbständig Muster und Regeln in Texten entdecken und analysieren. Die Schriftlichkeit, Kontextgebundenheit und Informalität der Interaktion auf Reddit ermöglichen es den Lernenden, ein breites Spektrum an Sprachkompetenzen und kommunikativen Fähigkeiten zu entwickeln und zu erweitern. Durch die eigenständige oder angeleitete Auseinandersetzung mit Beiträgen und Kommentaren auf Reddit kann ein fundiertes Verständnis für die Vielseitigkeit der deutschen Sprache und ihrer Verwendung in Online-Kontexten und Diskursgemeinschaften entwickelt und Fertigkeiten im Umgang mit unterschiedlichen digitalen Kommunikationssituationen und Schreibregistern auf- und ausgebaut werden.

3. Reddit als Korpus und digitale Sprachressource

In diesem Abschnitt wird untersucht, wie Reddit als Korpus für sprachliche Untersuchungen betrachtet und im Kontext DaF/DaZ genutzt werden kann. Reddit als Korpus betrachten heißt, die auf der Plattform verfügbaren Textdaten systematisch zu sammeln, aufzubereiten und für linguistische Analysen oder den Sprachunterricht zu verwenden. Grundsätzlich sind Reddit-Korpora immer auch Web-Korpora (vgl. Gatto 2014; Kehoe 2020). Korpus ist dabei zunächst im weiten Sinn zu verstehen als Sammlung von Texten, die aus Reddit-Posts und -Kommentaren besteht. In den folgenden Abschnitten erörtere ich, welche unterschiedlichen Methoden und Werkzeuge zum Einsatz kommen können, um Reddit als Korpus und digitale Sprachressource zu nutzen. Ausgehend von den Überlegungen von

Bernardini / Baroni / Evert (2006) unterscheidet drei Zugänge: Reddit als Korpus-Surrogat (Kap. 3.1), Reddit als Korpus-Fundus (Kap. 3.2) und Reddit als Korpus im engeren Sinne (Kap. 3.3).

3.1 Reddit als Korpus-Surrogat

Die erste Möglichkeit, Reddit als digitale Ressource für sprachbezogene Fragen zu nutzen, besteht darin, die Suchfunktion von Reddit oder externe Suchmaschinen zu verwenden, um auf der Plattform nach bestimmten Wörtern, Wortgruppen oder Themen zu suchen. Diese Art der Nutzung ist sehr zugänglich und erfordert wenig technisches Know-how. Auch Personen ohne spezifische korpuslinguistische Kenntnisse, Lehrende und Lernende, können auf diese Weise Texte und Diskussionsstränge zu einem bestimmten Thema oder Sprachgebrauch finden. In diesem Fall dient Reddit mit seinen angebotenen Suchmöglichkeiten in gewisser Weise als Ersatz für ein traditionelles Korpus und entsprechende Korpus-Software. Bernardini / Baroni / Evert (2006) sprechen daher von Korpus-Surrogat.

Die grundlegenden Suchmöglichkeiten und Filteroptionen auf Reddit umfassen die Eingabe von Suchbegriffen sowie das Anwenden von Filtern, um die Ergebnisse basierend auf Relevanz, Zeitraum oder Typ weiter einzugrenzen. Um die integrierte Suchfunktion zu nutzen, kann man die Reddit-Startseite (<https://www.reddit.com>) oder ein bestimmtes Subreddit besuchen und verwenden und das Suchfeld im oberen Navigationsbereich nutzen. Nutzer:innen können dabei gezielt nach bestimmten Subreddits, Schlagworten oder Themen suchen, um eine passende Datengrundlage für ihre Verwendung im DaF-/DaZ-Kontext zu schaffen. Zusätzlich bietet Reddit erweiterte Suchmöglichkeiten. Beispiele für solche Such-Operatoren sind: Anführungszeichen für exakte Wortgruppen; *AND*, *OR* und *NOT* für die Kombination von Begriffen; der *site*-Operator für das Auffinden von Beiträgen mit Links zu einer bestimmten Website und der *subreddit*-Operator für das Beschränken der Suche auf ein bestimmtes Subreddit (vgl. Reddit 2014).

Die explorative Nutzung von Reddit ermöglicht den Zugang zu vielfältigen Diskussionen und sprachlichen Variationen, stellt jedoch bei wissenschaftlicher oder didaktischer Verwendung auch erhebliche Herausforderungen dar. Zwar kann sie bei kleineren, fokussierten Untersuchungen hilfreich sein, doch stoßen systematische Analysen und die Auswertung größerer Korpora schnell an ihre Grenzen. Der Suchprozess kann sich als mühsam und zeitaufwendig erweisen, insbesondere wenn relevante Informationen in umfangreichen Diskussionssträngen versteckt sind. Im DaF/DaZ-Unterricht kann die Nutzung der Reddit-Suche das Sprachverständnis durch Konfrontation mit diversen Ausdrücken, Interaktionsmustern und Diskursen erweitern. Doch birgt die Community-generierte und nicht redaktionell geprüfte Natur von Reddit-Inhalten Risiken hinsichtlich Qualität und Relevanz. Lehrkräfte müssen deshalb gezielte Suchanfragen vorbereiten und dabei auch das Risiko unangemessener oder pädagogisch irrelevanter Inhalte berücksichtigen. Schließlich erfordert die Verwendung von Reddit im Unterricht ein gewisses Maß an digitaler Kompetenz sowohl von den Lehrkräften als auch von den Schüler:innen, was eine weitere Hürde darstellen kann. Insgesamt erfordert die Nutzung von Reddit als Korpus-Surrogat daher eine umsichtige und reflektierte Herangehensweise, die die genannten Herausforderungen und problematischen Aspekte berücksichtigt.

3.2 Reddit als Korpus-Fundus

Eine andere Verwendungsweise liegt vor, wenn Reddit nicht nur punktuell durchsucht, sondern verschiedene Diskussionsbeiträge und Kommentare aus Reddit ausgewählt und als Fundstellen kopiert oder heruntergeladen werden. Dabei entstehen kleinere oder größere Datensammlungen, die für

sprachliche Analysen und Übungen im DaF/DaZ-Unterricht verwendet werden können. Reddit dient dabei als Korpus-Fundus (*corpus shop* nach Bernardini / Baroni / Evert 2006).

Um ein solches Korpus auf Reddit zu erstellen, können Lehrkräfte beispielsweise thematisch oder stilistisch ähnliche Texte aus verschiedenen Subreddits zusammentragen. Die dabei entstehenden Textsammlungen haben in der Regel den Status von „wegwerfbaren“ Korpora (*disposable corpora*, vgl. Varantola 2003). Dieser Begriff bezieht sich auf temporäre und situationsbezogene Textsammlungen, die für eine spezifische Forschungsfrage oder einen didaktischen Zweck ad-hoc erstellt und anschließend auch wieder verworfen werden können. Diese Textsammlung ermöglicht es, unterschiedliche Aspekte der deutschen Sprache in authentischen Online-Kommunikationssituationen zu untersuchen, wie zum Beispiel den Gebrauch von Fachwortschatz, informellen Ausdrücken oder bestimmten grammatikalischen Strukturen. Im Unterricht kann ein solches Korpus dazu verwendet werden, um gezielte Übungen und Analysen durchzuführen, bei denen die Lernenden beispielsweise Gemeinsamkeiten und Unterschiede zwischen den Texten herausarbeiten, den Wortschatz erweitern oder bestimmte sprachliche Muster erkennen und anwenden können.

Dennoch birgt auch diese Methode Herausforderungen und problematische Aspekte. Die Auswahl und Zusammenstellung der Texte erfordert nicht nur Zeit und Aufwand, sondern auch ein hohes Maß an Urteilsvermögen, um geeignete und didaktisch wertvolle Beiträge zu identifizieren. Beispielsweise könnten Beiträge, die vulgäre oder beleidigende Sprache enthalten, für den Unterricht unpassend sein, obwohl sie reale Beispiele für die Verwendung der deutschen Sprache darstellen. Darüber hinaus können urheberrechtliche Fragen aufkommen, wenn Inhalte von Reddit kopiert und in einem anderen Kontext verwendet werden. Schließlich kann die Qualität und Relevanz der ausgewählten Beiträge stark variieren, da sie von der Community generiert und nicht redaktionell geprüft werden. So könnte ein Beitrag, der eine reiche Quelle für den Fachwortschatz zum Thema Umweltbewusstsein sein könnte, tatsächlich irreführende oder falsche Informationen enthalten. Dies kann insbesondere für weniger erfahrene Nutzer:innen oder Lernende problematisch sein, die Schwierigkeiten haben könnten, verlässliche und hochwertige Informationen von weniger relevanten oder gar irreführenden Beiträgen zu unterscheiden. Auch hier ist also ein reflektierter und kritischer Umgang mit den Ressourcen von Reddit von entscheidender Bedeutung.

Neben einer händischen Datensammlung via den oben dargestellten Suchmöglichkeiten ist auch eine maschinengestützte Zusammenstellung möglich. Eine effizientere Methode zur Datenerhebung ist der Zugriff auf Reddit-Daten über das *Application Programming Interface* (API). Verschiedene Programmiersprachen wie Python und R bieten Bibliotheken und Pakete, z. B. PRAW (vgl. Boe 2023) für Python und redditExtractoR (vgl. Rivera 2023), die den Zugriff auf die Reddit API erleichtern und das automatisierte Sammeln von Daten ermöglichen. Mit diesen Tools können Forscher:innen und Lehrende gezielt nach Subreddits, Diskussionssträngen und Kommentaren suchen und große Mengen an Daten für die linguistische Analyse sammeln. Allerdings sind für die Nutzung dieser Tools grundlegende Programmierkenntnisse erforderlich, was für einige Anwender:innen eine Hürde darstellen kann. Es gibt jedoch zahlreiche Ressourcen und Tutorials, die den Einstieg in die Programmierung und die Nutzung der Reddit API erleichtern.

3.3 Reddit-Korpora im engeren Sinn

Korpora im engeren Sinn beziehen sich auf strukturierte und systematisch angelegte Sammlungen von Sprachdaten, die für linguistische Analysen und Forschungszwecke verwendet werden. Aus einer angewandten Perspektive wäre es in der Regel wünschenswert, auf solche aufbereiteten und bereinigten Korpora zugreifen zu können, im besten Fall mit einem benutzer:innenfreundlichen Interface, etwa via CQPweb (vgl. Hardie 2012). Im Fall von Reddit ist die Auswahl solcher aufbereiteter und

bereinigter Korpora begrenzt, jedoch gibt es einige Optionen, die Forscher:innen und Lehrenden im DaF/DaZ-Kontext zur Verfügung stehen.

Das größte deutschsprachige Korpus ist das GeRedE-Korpus (*German Reddit Exchanges*, vgl. Blombach et al. 2020). Das GeRedE-Korpus umfasst etwa 270 Millionen Tokens, rund 380.000 Beiträge und 6,8 Millionen Kommentare zwischen 2010 und 2018. Um deutsche Inhalte zu erhalten, wurden die Korpusdaten mit dem zuvor schon von Barbaresi (2015) beschriebenen Verfahren aus einem großen, frei verfügbaren Datensatz gefiltert und anschließend weiter aufbereitet. Dabei wurden verschiedene Reddit-Metadaten übernommen und weitere Annotationsschichten (u.a. ein POS-Tagging) hinzugefügt, um die Analyse zu erleichtern.

Eine weitere Möglichkeit ist ConvoKit (vgl. Chang et al. 2020): ConvoKit ist eine Python-Bibliothek, die entwickelt wurde, um die Analyse von gesprächsorientierten Korpora zu erleichtern. Sie enthält vorverarbeitete Reddit-Korpora. ConvoKit bietet Nutzer:innen mit Python-Kenntnissen eine benutzerfreundliche Schnittstelle und eine Reihe von Funktionen, die es ermöglichen, die Struktur und den Inhalt von Interaktion auf Reddit zu analysieren. Jedes einzelne Posting oder Kommentar wird als Äußerung (*utterance*) betrachtet, während jedes Foren-Gespräch (*conversation*) aus einem Beitrag und den zugehörigen Kommentaren besteht. Für jede Äußerung werden Informationen wie Sprecher:in, Konversations-ID, Antwort auf, Zeitstempel und Text zur Verfügung gestellt, sowie weitere Reddit-spezifische Metadaten. Insgesamt sind 948.169 Subreddits enthalten, wobei jedes Korpus Beiträge und Kommentare aus einem individuellen Subreddit von dessen Gründung bis Oktober 2018 umfasst. ConvoKit ermöglicht es auch, verschiedene Subreddits miteinander zu kombinieren, um ein größeres Korpus für die Analyse zu erstellen. Dies kann insbesondere im DaF/DaZ-Kontext nützlich sein, wenn man beispielsweise mehrere deutschsprachige Subreddits miteinander kombinieren möchte.

Die dritte Möglichkeit, die ich hier vorstellen möchte, ist kein Korpus per se, sondern entspricht vielmehr einer Korpus- und Analysesoftware: Das 4CAT Capture and Analysis Toolkit (vgl. Peeters / Hagen 2022) ist ein Software-Tool, das entwickelt wurde, um die Analyse von Social-Media-Daten (einschließlich Reddit-Daten) zu erleichtern. 4CAT ermöglicht es, Datensätze basierend auf verschiedenen Suchkriterien zu erstellen und sie anschließend auf verschiedene Weise zu analysieren. Obwohl 4CAT nicht speziell auf Reddit ausgerichtet ist, bietet es dennoch einige Funktionen, die auch für die Analyse von Reddit-Daten nützlich sind. Die Software-Umgebung ermöglicht es, Reddit-Posts und -Kommentare nach bestimmten Kriterien wie Keywords, Zeitraum oder Subreddits zu durchsuchen und daraus ein eigenes Subkorpus zu erstellen. Dabei werden Funktionen wie die Anonymisierung von persönlichen Informationen (via Hashing der Nutzer:innen-Kennungen) und zahlreiche Filter- und Analysemöglichkeiten angeboten.

Allen drei Ansätzen ist gemeinsam, dass sie auf demselben Datensatz basieren. Sowohl das GeRedE-Korpus, als auch ConvoKit und 4CAT beziehen ihre Daten aus dem größten frei zugänglichen Reddit-Korpus: dem sog. Pushshift-Reddit-Dataset (vgl. Baumgartner et al. 2020), das eine nahezu vollständige Sammlung aller Reddit-Posts und -Kommentare enthält. Dieser Datensatz wurde für verschiedene Forschungszwecke intensiv genutzt (vgl. Gaffney / Matias 2018) und ist die Datenbasis für alle anderen hier erwähnten Korpora und Datensammlungen (so sie nicht direkt via Reddit API erstellt wurden). Es ist auch möglich, direkt auf die Rohdaten des Baumgartner-Korpus zuzugreifen. Neben der Möglichkeit, über ein Webinterface im Datensatz zu suchen (z.B. über <https://redditsearch.io/>), ist eine Option die Nutzung der Pushshift API. Eine weitere Möglichkeit besteht darin, direkt einzelne Datendumps von archivierten Reddit-Posts und -Kommentaren herunterzuladen (vgl. Baumgartner et al. 2020).

Insgesamt eröffnen bestehende Reddit-Korpora interessante Möglichkeiten für die Analyse von Sprache und Kommunikation im DaF/DaZ-Kontext. Bei der Verwendung dieser Ressourcen sollten jedoch stets die Repräsentativität, Aktualität und Vollständigkeit der Daten kritisch hinterfragt

und die Analyseergebnisse entsprechend interpretiert werden. Der Pushshift-Datensatz von Baumgartner et al. (2020) – und damit auch die davon abgeleiteten Korpora in ConvoKit (vgl. Chang et al. 2020) und GeRedE (vgl. Blombach et al. 2020) – ist kein vollständiges Abbild von Reddit und enthält einige Lücken (vgl. Gaffney / Matias 2018). Auch wenn dieser Befund für sprachbezogene Fragestellungen ohne Anspruch auf Repräsentativität weniger stark ins Gewicht fällt, ist bei der Verwendung aller Korpora die Qualität und Vollständigkeit der Daten kritisch zu betrachten. Es ist grundsätzlich davon auszugehen, dass es in Reddit-Korpora zu fehlenden oder unvollständigen Daten kommen kann, was bei der Interpretation der Analyseergebnisse eine vertiefende Reflexion hinsichtlich Validität und Reliabilität erfordert und auch an die Lernenden vermittelt werden muss.

4. Fallstudie: Sprachvergleich von r/Austria und r/de

4.1 Ziel und Fragestellung

Im folgenden Abschnitt möchte ich exemplarisch zeigen, wie sich die beschriebenen Ressourcen in ConvoKit (vgl. Chang et al. 2020) für eine korpusgeleitete Untersuchung des Sprachgebrauchs und der Sprachvariation auf Reddit nutzen lassen. Der Fokus liegt auf den beiden Subreddits r/de und r/Austria, die für den DaF-/DaZ-Kontext besonders relevant sind: Beide Subreddits repräsentieren große deutschsprachige Communities und bieten somit eine breite Auswahl an Diskussionsthemen. Darüber hinaus bieten sie die Möglichkeit, Unterschiede und Gemeinsamkeiten zwischen der deutschen und österreichischen Sprach- und Diskurspraxis zu untersuchen. r/de ist ein deutschsprachiger Subreddit, der sich auf Diskussionen, Nachrichten und Inhalte über Deutschland konzentriert. Mit über einer Million Mitgliedern ist es einer der größten deutschsprachigen Subreddits und dient als eine wichtige Quelle für Informationen und Diskussionen über Deutschland und die deutsche Kultur. r/Austria hingegen fokussiert Österreich und bietet eine Plattform für Diskussionen, Nachrichten und Inhalte rund um das Land und die österreichische Kultur. Mit über 200.000 Mitgliedern ist dieser Subreddit ebenfalls bedeutend für den deutschsprachigen Raum, insbesondere für die österreichische Community.

Im Kontext von DaF-DaZ ermöglicht ein Vergleich der deutschsprachigen Subreddits somit einen korpusbasierten und korpusgeleiteten Zugang zum Thema Sprachvariation im Internet. Für den Unterrichtskontext können die beiden Subreddits darüber hinaus in vielfältiger Weise als Lehr- und Lernressource genutzt werden (vgl. Kap. 2). So wäre es beispielsweise möglich, die Reddit-Foren als interaktives Lehr- und Lernmedium zu nutzen, in dem sich die Lernenden selbstständig an Diskussionen beteiligen und „kreativ und kommunikativ im virtuellen Raum am Sprachenlernen [partizipieren]“ (Vergeiner 2021: 161). Die Lehrenden sind dabei in mehrfacher Hinsicht gefordert, die Lernenden vorzubereiten und zu unterstützen. Eine didaktische Aufgabe besteht darin, die vorherrschenden Sprach- und Diskurspraktiken als Gegenstand sprachlicher Variation und soziokulturell geprägter Mediennutzung zu kontextualisieren. Dazu benötigen die Lehrenden vertiefende Informationen zu den Subreddits, die über Korpora erschlossen werden können.

Vor diesem Hintergrund möchte die exemplarische Fallstudie dazu beitragen, Lehrende über die Möglichkeiten und Herausforderungen des Einsatzes der beiden Subreddits r/de und r/Austria als Lehr- und Lernmedien zu informieren und ihnen eine Grundlage für die Planung und Umsetzung von entsprechenden Unterrichtseinheiten und -aktivitäten zu bieten. Die Fragestellung lautet: Inwiefern unterscheiden sich die korpusbasiert greifbaren Sprach- und Diskurspraktiken in den Subreddits r/de und r/Austria und welche Implikationen ergeben sich daraus für den DaF/DaZ-Unterricht?

4.2 Daten und Vorgehen

Für die Beantwortung der Fragestellung wurde eine datengeleitete Analyse mithilfe von ConvoKit (vgl. Chang et al. 2020) durchgeführt. ConvoKit ist ein Python-Toolkit, das zur Analyse und Verarbeitung von (schriftbasierten) Interaktionen entwickelt wurde und eine einfache Handhabung und flexible Analyse von interaktionsorientierten Gesprächsdaten ermöglicht. Die Reddit-Korpora in ConvoKit basieren auf dem Pushshift-Datensatz (vgl. Baumgartner et al. 2020) und enthalten Informationen über Beiträge, Kommentare, Nutzer:innen und andere Metadaten, die in den jeweiligen Subreddits zu finden sind. Das Korpus ist als Stichprobe zu verstehen (vgl. Gaffney / Matias 2018).

In ConvoKit wird ein Gespräch (*conversation*) als eine hierarchisch strukturierte Sammlung von Äußerungen (*utterances*) betrachtet, die einen gemeinsamen thematischen oder diskursiven Zusammenhang aufweisen. Terminologisch beziehen sich Gespräche daher auf den gesamten dialogischen Verlauf eines Diskussionsthreads, der aus einem initialen Beitrag (Post) und allen dazugehörigen Kommentaren besteht. Sowohl der ursprüngliche Beitrag (Post) als auch die darauf folgenden Kommentare der Sprecher:innen (= Nutzer:innen) werden dabei als Äußerungen bezeichnet.

Das Analyseverfahren war korpusgeleitet und explorativ. Im Rahmen der Analyse wurden zunächst grundlegende statistische Daten zu den Korpora von r/de und r/Austria erhoben, um einen Überblick über die Größe und Struktur der beiden Subreddits zu erhalten. Dazu zählten die Anzahl der Sprecher:innen, die Anzahl der Äußerungen und Gespräche sowie die durchschnittliche Länge der Äußerungen in beiden Subreddits. Insgesamt ist r/de ungefähr 8,5-mal größer als r/Austria in Bezug auf die Anzahl der Äußerungen und etwa 4-mal größer hinsichtlich der Anzahl der Unterhaltungen und der Sprecher:innen. Die Vergleichsdaten zu den Korpora sind in Tabelle 1 dargestellt.

Merkmal	r/Austria	r/de
Anzahl der Sprecher:innen	18.764	80.605
Anzahl der Äußerungen	562.325	4.792.155
Anzahl der Unterhaltungen	37.353	228.822
Durchschnittliche Anzahl der Wörter pro Äußerung	35,15	32,32
Durchschnittliche Anzahl der Äußerungen pro Unterhaltung	15,05	20,94
Durchschnittliche Anzahl der Äußerungen pro Sprecher:innen	29,97	59,45
Durchschnittliche Anzahl der Wörter pro Sprecher:inner	1.053,48	1921,66
Anteil der einmaligen Sprecher:innen	35.13%	37.45%

Tabelle 1

Übersicht von r/Austria und r/de: Vergleich der Anzahl der Sprecher:innen, Äußerungen, Unterhaltungen, der durchschnittlichen Anzahl von Wörtern pro Äußerung und Sprecher:in, sowie der Anteil der einmaligen Sprecher:innen

Für einen quantitativen Vergleich wurden dann die relativen Worthäufigkeiten in beiden Subreddits ermittelt und verglichen. Dazu wurden via eines Python-Skripts zunächst alle Äußerungen in Listen extrahiert und mittels regulärer Ausdrücke vorverarbeitet, um Groß- und Kleinschreibung zu vereinheitlichen und Zeichen, die keine Wörter sind, zu entfernen. Anschließend wurden alle Tokens gezählt und ihre relative Häufigkeit berechnet, indem die absolute Häufigkeit jedes Wortes durch die Gesamtzahl der Wörter im jeweiligen Subreddit dividiert wurde. Die Ergebnisse wurden in einer Tabelle sowie Visualisierungen dargestellt und verglichen. Die einhundert Ausdrücke mit der größten Divergenz wurden anschließend in verschiedene induktiv-entwickelte Kategorien eingeteilt und manuell geclustert, um einen vertieften Einblick in die unterschiedlichen Themen und Diskurspraktiken der beiden Subreddits zu erhalten. Alle genutzten Python-Skripts, Analysedaten und Visualisierungen sind online verfügbar.

4.3 Ergebnisse

Vergleicht man die Teilkorpora der beiden Subreddits r/Austria und r/de in ihrer Gesamtheit, so ergeben sich zunächst einige allgemeine Größen, die einen genaueren Eindruck von der Ausprägung der beiden Diskursgemeinschaften vermitteln.

Die Ergebnisse zeigen, dass r/de sowohl in Bezug auf die Anzahl der Sprecher:innen als auch die Anzahl der Äußerungen und Unterhaltungen deutlich größer ist als r/Austria. Es ist interessant festzustellen, dass die durchschnittliche Anzahl der Wörter pro Äußerung in r/Austria etwas höher ist als in r/de, während die durchschnittliche Anzahl der Äußerungen pro Unterhaltung in r/de höher ist als in r/Austria. Dies könnte darauf hindeuten, dass in r/de möglicherweise mehr Diskussionen stattfinden, während in r/Austria die Beiträge möglicherweise ausführlicher sind.

Darüber hinaus zeigt sich, dass die durchschnittliche Anzahl der Äußerungen und Wörter pro Sprecher:innen in r/de höher ist als in r/Austria, was darauf hindeuten könnte, dass die Nutzer:innen in r/de aktiver sind oder mehr zum Diskurs beitragen. Der Anteil der einmaligen Sprecher:innen ist in beiden Subreddits ähnlich, wobei r/de einen etwas höheren Anteil aufweist. Dies könnte bedeuten, dass in beiden Subreddits eine ähnliche Anzahl von Nutzer:innen nur einmalig teilnehmen und dann nicht weiter aktiv sind.

In der quantitativen Analyse wurden insgesamt 1.811.203 Wortformen hinsichtlich ihrer relativen Häufigkeit gezählt und die Top 100 pro Subreddit manuell kategorisiert. Insgesamt konnten folgende übergreifende Cluster gebildet werden: Sprache, Land, Politik, Medien/Technologie. In r/Austria sind zudem englischsprachige Ausdrücke deutlich häufiger als in r/de. Abbildung 3 zeigt eine Treemap-Visualisierung, die den Vergleich der beiden Subreddits hinsichtlich der relativen Häufigkeit der Top-100-Wörter veranschaulicht. In der Treemap werden die Wörter als Rechtecke dargestellt, wobei die Größe der Rechtecke der relativen Häufigkeit des jeweiligen Wortes entspricht. Die Rechtecke sind außerdem nach den oben genannten Clustern gruppiert, um die Verteilung und Gewichtung der verschiedenen Kategorien innerhalb der Subreddits besser erkennen zu können.

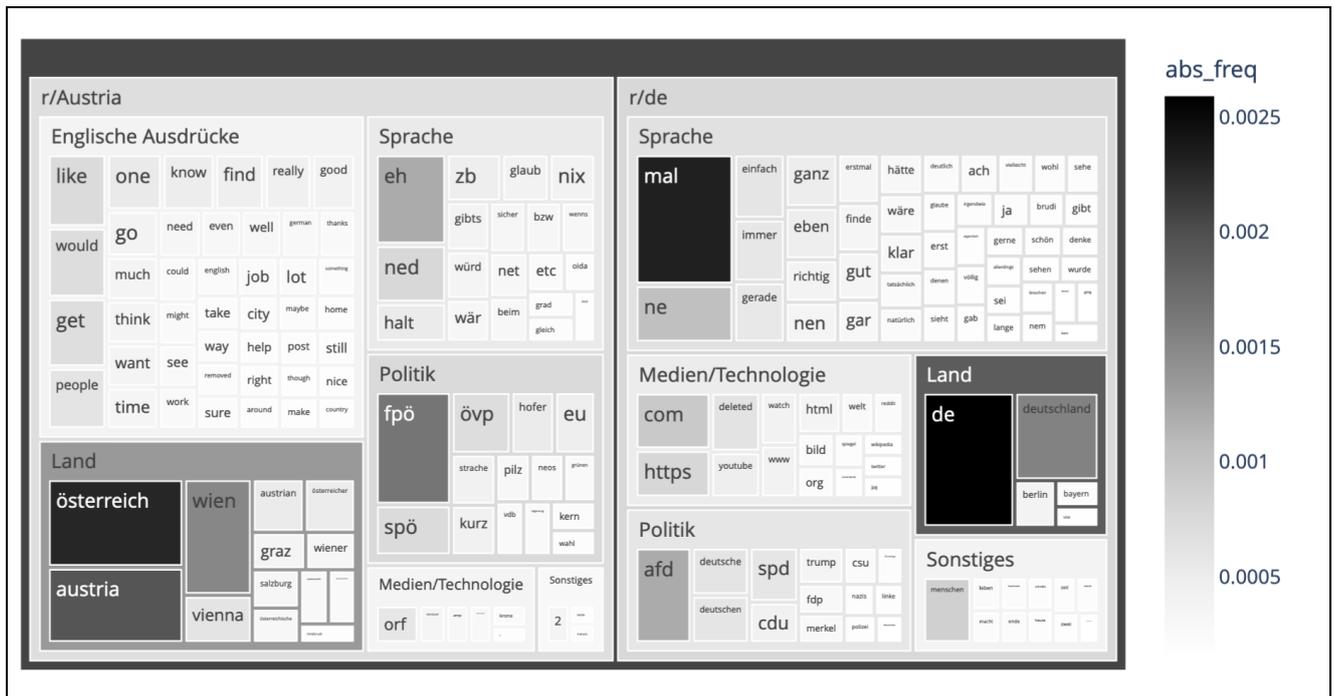


Abbildung 3

Vergleich der Top-100-Wörter mit größter relativer Worthäufigkeit in r/Austria und r/de (manuelles Clustering, untersuchte Tokens: 1.811.203)

Die Ergebnisse des quantitativen Vergleichs zeigen, inwiefern sich die beiden Subreddits r/de und r/Austria in ihren Sprach- und Diskurspraktiken unterscheiden. Die Häufigkeit von politischen, geographischen, umgangssprachlichen sowie medien- und technikbezogenen Begriffen variiert zwischen den beiden Communities. Wenig überraschend sind politische und geographische Ausdrücke, die sich auf österreichische Kontexte beziehen, in r/Austria stärker vertreten, während politische und geographische Ausdrücke, die sich auf Deutschland und deutsche Kontexte beziehen, in r/de häufiger sind. Darüber hinaus sind in beiden Subreddits nationale Medien und internetspezifische Begriffe (URLs, Plattformnamen) präsent. Unterschiede gibt es auch im Bereich der funktionalen Wortformen, die auf regionale Sprachvarietäten hinweisen. Insgesamt sind in r/Austria österreichische Sprachvarietäten stärker vertreten, während in r/de eher bundesdeutsche Sprachvarietäten auftreten.

Trotz dieser Unterschiede gibt es auch Gemeinsamkeiten in den Diskurspraktiken beider Subreddits. Zum einen verwenden beide Communities eine informelle und interaktionsorientierte Sprache. Dies zeigt sich in der Verwendung von umgangssprachlichen Ausdrücken und Diskurspartikeln, die dazu dienen, den schriftlichen Dialog zu strukturieren und die Beziehung zwischen den Sprecher:innen zu gestalten. Zweitens zeigen beide Subreddits ein Interesse an politischen und geographischen Themen, wobei sie sich auf ihre jeweilige nationale oder regionale Perspektive konzentrieren. Dies spiegelt das gemeinsame Bedürfnis der Teilnehmer:innen wider, sich über aktuelle Ereignisse und Entwicklungen auszutauschen und ihre Meinungen und Erfahrungen zu teilen. Schließlich zeigen die Korpusdaten auch, dass allgemeine Diskurspraktiken wie das Stellen von Fragen, das Anbieten von Informationen und das Äußern von Meinungen in beiden Subreddits zu beobachten sind.

Ein zentrales Ergebnis der Analyse ist, dass die sprachlichen Unterschiede zwischen den beiden Subreddits nicht nur auf verschiedenen Schwerpunkten in Bezug auf Themen wie Politik und Geographie beruhen, sondern auch auf funktionalen Ausdrücken, die spezifisch für die schriftliche Interaktion in den jeweiligen Communities sind. In Abbildung 4 ist der Cluster ‚Sprache‘ im Detail visualisiert.

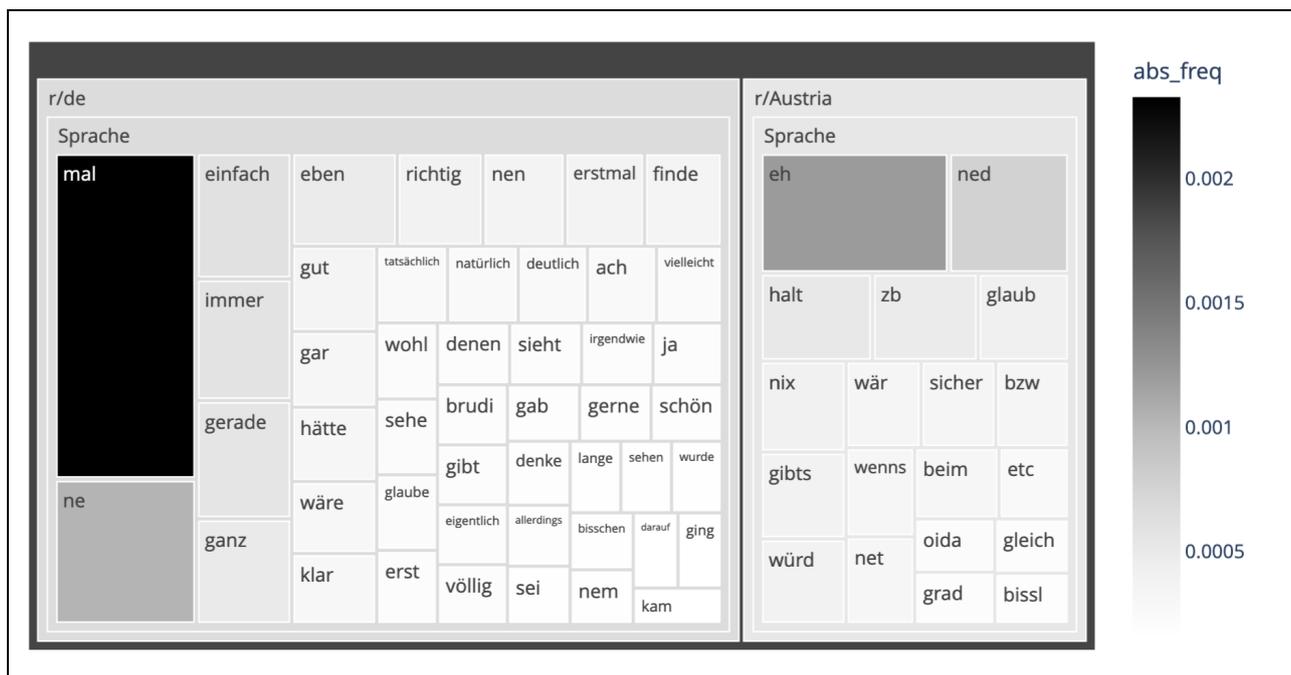


Abbildung 4
Detailansicht des Clusters ‘Sprache’ in r/Austria und r/de.

Die Analyse bringt einige sprachliche Unterschiede zwischen den beiden Subreddits hervor. Im österreichischen gesprochenen Deutsch sind Diskursmarker wie *eh*, *halt* und *net* weit verbreitet und erfüllen unterschiedliche Funktionen. *Eh* dient beispielsweise der Abschwächung oder Betonung einer Aussage, *halt* kann Zustimmung, Resignation oder Unvermeidbarkeit signalisieren und *net* stellt eine regionale Variante von *nicht* dar. Diese Ausdrücke sind in der Tendenz für den österreichischen Sprachgebrauch charakteristisch und weisen auf regionale sprachliche Besonderheiten hin. Im Gegensatz dazu sind tendenziell für den deutschen Sprachgebrauch typische Diskursmarker und Ausdrücke wie *mal*, *ne* und *nen* in r/de stärker vertreten. *Mal* dient der Abschwächung von Aufforderungen oder Bitten und der Verdeutlichung zeitlicher Nähe, während *ne* und *nen* Artikelreduktionen sind, die in informellen Gesprächen verwendet werden. Diese Diskursmarker und Ausdrücke sind für den deutschen Sprachgebrauch charakteristisch und verweisen auf sprachliche Besonderheiten dieser Community.

Ein weiteres bemerkenswertes Ergebnis betrifft die Variation bei den Konjunktiv-II-Wortformen *wär* (häufig in r/Austria, selten in r/de) und *wäre* (häufig in r/de, selten in r/Austria) sowie die Variation von *bissl* vs. *bisschen*. Diese Beobachtungen zeigen, wie regionale Dialekte und Sprachgewohnheiten in den beiden Subreddits zum Ausdruck kommen. Die größere Verwendung von *wär* und *bissl* in r/Austria zeigt, dass Nutzer:innen im österreichischen Subreddit eher zu informelleren Formen neigen, während Sprecher:innen in r/de eher die Standard-Formen *wäre* und *bisschen* verwenden.

Abschließend sei noch ein Befund aus dem Bereich der Nomina hervorgehoben. *Oida* und *Brudi* sind beides informelle Ausdrücke, die in den jeweiligen Subreddits r/Austria und r/de häufig verwendet werden. Beide Ausdrücke haben eine gewisse Ähnlichkeit in ihrer Funktion, obwohl sie unterschiedliche sprachliche Formen haben und in verschiedenen regionalen Varianten des Deutschen vorkommen. *Oida* hat seinen Ursprung im Wienerischen Dialekt. Es leitet sich von *Alter* ab und wird oft als interjektivischer Ausdruck der Überraschung, Verärgerung, Zustimmung oder auch als Anrede verwendet. *Brudi*, abgeleitet von *Bruder*, ist in Deutschland als informelle Anrede oder Bezeichnung für einen Freund oder Bekannten gebräuchlich. Es hat eine ähnliche Funktion wie *Oida*

in Bezug auf die informelle Anrede und drückt eine gewisse Verbundenheit zwischen Sprecher:innen aus. Beide Ausdrücke zeigen eine informelle und vertraute Kommunikation unter den Mitgliedern der jeweiligen Subreddits. Sie dienen dazu, eine Gemeinschaft zu schaffen und den Austausch in einer entspannten Atmosphäre zu fördern. Obwohl *Oida* und *Brudi* in unterschiedlichen regionalen Varianten des Deutschen vorkommen, haben sie gemeinsam, dass sie die Identität und kulturellen Eigenheiten ihrer jeweiligen Sprechergruppen repräsentieren und zur sozialen Kohäsion innerhalb der Subreddits beitragen.

4.4 Diskussion

Die angestellte Analyse der beiden Subreddits *r/Austria* und *r/de* liefert einige aufschlussreiche Ergebnisse bezüglich sprachlicher Variationen und Diskurspraktiken innerhalb der beiden deutschsprachigen Communities. Durch den Vergleich der relativen Worthäufigkeiten wurden sowohl thematische als auch sprachliche Unterschiede aufgedeckt, die für den DaF/DaZ-Unterricht von Bedeutung sein können.

Die aufgeführten Ergebnisse liefern wertvolle Informationen, die für einen fortgeschrittenen DaF/DaZ-Unterricht genutzt werden können. Der Unterschied im Gebrauch von Diskursmarkern und regionalen Ausdrücken kann in Lektionen über regionale Variationen im deutschsprachigen Raum einfließen. In diesem Kontext können Lernende ermutigt werden, die Bedeutung und Verwendung dieser Wörter zu untersuchen (z.B. anhand exemplarischer Belegstellen) und ihren Wortschatz durch Integration in eigene Textproduktionen zu erweitern. Allerdings müssen Lehrkräfte auch die Herausforderungen berücksichtigen, die mit der Verwendung von authentischen Online-Texten in der Klasse verbunden sind. Es ist wichtig, dass die Lernenden lernen, kritisch mit diesen Texten umzugehen und ihre Eignung als Lernressourcen einzuschätzen. Des Weiteren müssen die Lehrkräfte die notwendige Kontextualisierung und Begleitung bereitstellen, um sicherzustellen, dass die Lernenden von diesen Materialien profitieren können. Insgesamt ermöglichen die hier dargestellten Analysen die Entwicklung eines komplexen und realitätsnahen Bildes der deutschen Sprache, das über das traditionelle Klassenzimmer hinausgeht. Sie erfordern jedoch eine sorgfältige didaktische Vorbereitung und Anpassung an die Bedürfnisse und Fähigkeiten der Lernenden.

Die sprachlichen Unterschiede, insbesondere bei Partikeln, Adverbien und Diskursmarkern, weisen auf regionale Variationen und Kommunikationsstile im Kontext des interaktionsorientierten Schreibens im Internet hin, die für den DaF/DaZ-Unterricht relevant sind. Ein Verständnis dieser Unterschiede und Schreibformen kann dazu beitragen, das Bewusstsein für sprachliche Variationen im deutschsprachigen Raum und der Interaktion im Internet zu fördern und den Lernenden helfen, sich besser auf die verschiedenen Kommunikationskontexte einzustellen. Ein großer Vorteil der Verwendung von Reddit als Korpus besteht darin, dass die Lernenden die Möglichkeit haben, Wörter und Ausdrücke im Kontext zu sehen. Anders als bei isolierten Wörterbuchdefinitionen können Lernende durch die Kontextualisierung von Wörtern und Ausdrücken in echten Diskussionen und Interaktionen ein tieferes und praktischeres Verständnis der Sprache erlangen.

Die Unterschiede in den Diskurspraktiken, etwa in Bezug auf politische Themen, Mediennutzung und geographische Schwerpunkte, bieten wichtige Einblicke in die kulturellen Unterschiede zwischen den beiden Communities. Diese Erkenntnisse können im DaF/DaZ-Unterricht genutzt werden, um den Lernenden ein besseres Verständnis für die kulturellen und politischen Hintergründe der deutschsprachigen Länder und Regionen zu vermitteln.

Aus den Ergebnissen der Fallstudie lassen sich mehrere Implikationen für den DaF/DaZ-Unterricht ableiten:

1. Bewusstsein für regionale sprachliche Variationen schaffen: Lehrer:innen können die Unterschiede zwischen den Subreddits nutzen, um den Lernenden die Vielfalt der deutschen Sprache und die regionalen Variationen näherzubringen. Diskussionen und Vergleiche der unterschiedlichen Ausdrücke und Sprachvarianten können das Interesse der Lernenden wecken und ihre Motivation erhöhen.
2. Kulturelles Verständnis fördern: Lehrer:innen können die Diskussionen aus den Subreddits nutzen, um den Lernenden die kulturellen Unterschiede und Gemeinsamkeiten zwischen den deutschsprachigen Ländern und Regionen näherzubringen. Dies kann dazu beitragen, Stereotypen abzubauen und das Verständnis für die kulturelle Vielfalt im deutschsprachigen Raum zu fördern.
3. Aktuelle Themen und Diskurse integrieren: Lehrer:innen können die in den Subreddits diskutierten Themen und Diskurse nutzen, um den Unterricht aktuell und relevant zu gestalten. Dies kann dazu beitragen, den Lernenden die Relevanz des Deutschlernens für ihr tägliches Leben und ihre persönlichen Interessen zu verdeutlichen.

Insgesamt zeigen die Ergebnisse der Korpusanalyse, dass die Analyse von Sprach- und Diskurspraktiken in den Subreddits r/de und r/Austria wertvolle Einblicke in die Sprach- und Diskurspraktiken in beiden Online-Communities für den DaF/DaZ-Unterricht bieten kann. Dennoch gibt es einige Limitationen, die bei der Interpretation der Ergebnisse zu berücksichtigen sind, von denen ich hier drei herausheben möchte: Erstens ist die Datenquelle als nicht-vollständig zu werten: Die Daten wurden aus dem Pushshift-Datensatz extrahiert, der nicht alle Beiträge und Kommentare aus den Subreddits enthält, sondern als lückenhaftes Sample verstanden werden muss (vgl. Gaffney / Matias 2018). Dadurch könnte die Analyse möglicherweise nicht alle relevanten Aspekte der Sprach- und Diskurspraktiken in den beiden Communities erfassen. Die Größe der via Convokit erzielbaren Stichprobe bietet insgesamt aber wohl doch eine solide Grundlage für die Identifizierung von Mustern und Tendenzen in der Verwendung von Wörtern und Ausdrücken. Insgesamt konnten mehr als 5 Millionen Äußerungen (Reddit-Kommentare) in 266.175 Reddit-Posts untersucht werden, die von 95.708 Nutzer:innen stammen. Die zweite Limitation betrifft die Repräsentativität. Die untersuchten Subreddits sind nicht repräsentativ für Sprecher:innen aus Deutschland oder Österreich, sondern repräsentieren lediglich die Sprach- und Diskurspraktiken der Nutzer:innen, die sich aktiv an diesen Online-Communities beteiligen. Daher sollten die Ergebnisse nicht verallgemeinert werden, um allgemeine Aussagen über die deutsche oder österreichische Standardsprache zu treffen (vgl. dazu auch Kaltenböck / Mehlmauer-Larcher 2005). Limitationen ergeben sich drittens auch aus der Selbstselektion der Nutzer:innen: Die Nutzer:innen von Reddit sind möglicherweise nicht repräsentativ für die breitere Bevölkerung, da sie sich aufgrund ihrer Interessen, Bildung oder persönlichen Präferenzen für die Teilnahme an den Subreddits entscheiden. Diese Selbstselektion kann zu einer Verzerrung der Sprach- und Diskurspraktiken führen, die in der Analyse beobachtet werden.

Trotz dieser Limitationen hat die Korpusanalyse wertvolle Informationen über die sprachlichen Unterschiede und Gemeinsamkeiten zwischen den Subreddits r/de und r/Austria hervorgebracht. Die Analyse hat gezeigt, dass es Unterschiede in der Verwendung von Diskursmarkern, politischen Begriffen, Ortsnamen und Mediennamen gibt, die für den DaF/DaZ-Unterricht genutzt werden können. Die Ergebnisse bieten auch einen authentischen Einblick in die Kommunikation von deutschsprachigen Nutzer:innen in Online-Communities, was für die Entwicklung von Lehrmaterialien und die Gestaltung des Unterrichts von Bedeutung ist. Obwohl die Analyse keine umfassenden Aussagen über die deutsche oder österreichische Standardsprache treffen kann, trägt sie dennoch zur Förderung eines tieferen Verständnisses für die Vielfalt und Dynamik der deutschen Sprache im Internet bei.

Ein solches Wissen über unterschiedliche Kommunikationsstile im Unterricht zu vermitteln, ist zweifellos anspruchsvoll. Zu erkennen, auf welcher vielfältigen Weise die deutsche Sprache im Internet, etwa auf Plattformen wie Reddit, verwendet wird, erscheint jedoch als ein entscheidender Aspekt, wenn es um die Vermittlung eines umfassenden und aktuellen Sprachverständnisses geht. Ein fortgeschrittener DaF/DaZ-Unterricht, der sich dieser Aufgabe stellt, kann auf die vorgestellten Korpusressourcen zurückgreifen. Die Umsetzung im Unterricht erfordert jedoch eine kontinuierliche

Kontextualisierung und Sensibilisierung der Lernenden für die diaphasische Dimension, d.h. dass die Art und Weise, wie wir sprechen und schreiben, von verschiedenen Faktoren abhängt, u.a. vom Kontext, von der Situation und von der Beziehung zwischen den Partner:innen. Die Vermittlung dieser Aspekte kann eine besondere Herausforderung darstellen, da es nicht nur darum geht, den Lernenden die Standardsprache beizubringen, sondern ihnen auch die Fähigkeit zu vermitteln, verschiedene sprachliche Register zu erkennen und zu verwenden.

Die vorgestellten Korpusressourcen ermöglichen es, solche nicht-standardisierten Schreibweisen und Formulierungen, die sich an der gesprochenen Sprache orientieren, zum Gegenstand des Unterrichts zu machen. Social-Media-Kommunikation ist jedoch auch für den Sprachunterricht ein „moving target“ (vgl. Beißwenger 2020: 292) und zwingt dazu, Unterrichtsansätze und -materialien anzupassen und zu aktualisieren, um den sich stetig wandelnden Kommunikationsstilen und der Dynamik des Sprachgebrauchs im Internet Rechnung zu tragen. Lehrkräfte müssen daher nicht nur in der Lage sein, relevante und lehrreiche Inhalte auf Reddit zu identifizieren, sondern auch in der Lage sein, diese Inhalte in einen für den Unterricht geeigneten Kontext zu stellen. Gleichzeitig bietet diese Herausforderung auch eine wertvolle Gelegenheit: Sie ermöglicht es den Lernenden, ein tieferes und praktischeres Verständnis der deutschen Sprache zu entwickeln und sie besser auf die reale Sprachverwendung in verschiedenen Kontexten, insbesondere in der Online-Kommunikation, vorzubereiten.

In der exemplarischen Fallstudie wurde nur ein Aspekt der sprachlichen Unterschiede zwischen r/de und r/Austria untersucht, nämlich die relative Worthäufigkeit. Es gibt jedoch weitere korpuslinguistische Methoden und Ansätze, die für eine umfassendere Analyse der Sprach- und Diskurspraktiken in den beiden Subreddits von Interesse sein könnten. Zum Beispiel könnte das Part-of-Speech (POS)-Tagging stärker berücksichtigt werden, um Informationen über die syntaktischen Eigenschaften der verwendeten Wörter zu erhalten. Dies könnte helfen, weitere Unterschiede in der Verwendung von Wortarten und grammatischen Strukturen zwischen den beiden Communities aufzudecken. Darüber hinaus bietet ConvoKit auch Möglichkeiten zur Vektorisierung von Texten und zur Anwendung von Machine-Learning-Methoden. Durch die Anwendung von Textklassifikations- oder Clustering-Algorithmen könnten möglicherweise weitere Muster und Zusammenhänge in den Daten erkannt werden, die bei einer reinen Frequenzanalyse verborgen bleiben. Eine Kombination dieser zusätzlichen Methoden könnte zu einer umfassenderen und differenzierteren Analyse der sprachlichen Unterschiede und Gemeinsamkeiten zwischen r/de und r/Austria führen. Dies wiederum würde den Lehrkräften im DaF/DaZ-Unterricht noch mehr Informationen und Anhaltspunkte liefern, um die Lernenden auf die verschiedenen sprachlichen Besonderheiten und Diskurspraktiken in den beiden Communities vorzubereiten und zu unterstützen.

5. Schluss

Ziel des Beitrags war es, die Social-Media-Plattform Reddit als Lehr- und Lernmedium vorzustellen und die Möglichkeiten ihrer Nutzung als digitale Sprachressource zu untersuchen. Insgesamt wurde aufgezeigt, dass Reddit als Medium und digitale Sprachressource für den DaF/DaZ-Kontext ein vielversprechendes Potenzial bietet. Durch die systematische Analyse und didaktische Aufbereitung der auf der Plattform verfügbaren sprachlichen Ressourcen können Lehrende und Lernende ein besseres Verständnis für Sprachvariation, kulturelle Unterschiede und die Dynamik des Sprachwandels im digitalen Zeitalter gewinnen und so die Qualität des DaF/DaZ-Unterrichts weiter erhöhen.

Trotz der vielversprechenden Möglichkeiten, die Reddit als Korpus für den DaF/DaZ-Unterricht bietet, sollte aber an dieser Stelle auch darauf hingewiesen werden, dass eine kritische Reflexion über die Verwendung von Social-Media-Daten und deren Herausforderungen unerlässlich ist (vgl.

Kaltenböck / Mehlmauer-Larcher 2005). Eine sorgfältige Auswahl und Aufbereitung von Texten und Diskussionen ist notwendig, um den Lernenden eine angemessene und vielfältige Auswahl an authentischen Sprachressourcen bereitzustellen. Zukünftige Forschungsarbeiten sollten daher weiterhin eine ausgewogene Betrachtung der Potenziale und Schwierigkeiten bei der Verwendung von Reddit-Daten im DaF/DaZ-Kontext verfolgen, um die bestmögliche Integration in den Unterricht zu erreichen.

Durch die Implementierung von korpusbasierten Unterrichtsstrategien könnten Lernende direkt mit den sprachlichen Ressourcen aus Reddit interagieren, um Muster und Regeln in authentischen Texten zu entdecken und zu analysieren. Als mögliche Vertiefung und Weiterentwicklung der präsentierten Ergebnisse und Ansätze sind zunächst konkrete fachdidaktische Umsetzungen und Anwendungen in spezifischen Lehr-Lern-Szenarien denkbar. Die Verwendung von Reddit als Korpus und digitale Sprachressource könnte nicht nur die Entwicklung von Sprachbewusstsein und kritischer Analyse fördern, sondern auch dazu beitragen, Lernenden die spezifischen Kommunikationsnormen und -stile von internetbasierter Kommunikation im Deutschen näherzubringen. In diesem Zusammenhang würden weitere empirische Studien korpusbasierter und korpusgeleiteter Unterrichtsstrategien mit Reddit-Daten einen wertvollen Beitrag zum DaF/DaZ-Feld leisten. So wäre insbesondere die thematisierte Nutzung von Reddit-Daten im Kontext von Data-Driven Learning (vgl. Johns 1991) auch praktisch zu untersuchen. Empirische Studien und Reflexionen könnten zeigen, wie Lehrkräfte und Lernende Reddit-Daten und -Ressourcen in konkreten didaktischen Szenarios nutzen, um Sprachvariation, kulturelle Unterschiede und die Dynamik des Sprachwandels im digitalen Zeitalter zu erkunden.

Literatur und Ressourcen

Barbaresi, Adrien (2015): Collection, Description, and Visualization of the German Reddit Corpus. In: *2nd Workshop on Natural Language Processing for Computer-Mediated Communication*, 7-11. <https://hal.science/hal-01207311> (04.09.2023).

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, 830-839. <https://doi.org/10.1609/icwsm.v14i1.7347>.

Beißwenger, Michael (2020): Internetbasierte Kommunikation als Textformen-basierte Interaktion: ein neuer Vorschlag zu einem alten Problem. In: Marx, Konstanze / Lobin, Henning / Schmidt, Axel (Hrsg.): *Deutsch in Sozialen Medien*. Berlin / Boston: de Gruyter, 291-318.

Bernardini, Silvia / Baroni, Marco / Evert, Stefan (2006): A WaCky Introduction. In: Baroni, Marco / Bernardini, Silvia (Hrsg.): *Wacky! Working papers on the web as Corpus*. Bologna: GEDIT, 9-40.

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*. Marseille: European Language Resources Association, 6310-6316. <https://aclanthology.org/2020.lrec-1.774> (04.09.2023).

Boe, Bryce (2023): PRAW: The Python Reddit API Wrapper. Python Python Reddit API Wrapper Development. <https://github.com/praw-dev/praw> (04.09.2023).

Boulton, Alex (2017): Corpora in language teaching and learning. In: *Language Teaching* 50: 4, 483-506. <https://doi.org/10.1017/S0261444817000167>.

Chang, Jonathan P. / Chiam, Caleb / Fu, Liye / Wang, Andrew / Zhang, Justine / Danescu-Niculescu-Mizil, Cristian (2020): ConvoKit: A Toolkit for the Analysis of Conversations. In: *Proceedings of the 21th Annual*

- Meeting of the Special Interest Group on Discourse and Dialogue*. Stroudsburg, PA: Association for Computational Linguistics, 57-60. <https://aclanthology.org/2020.sigdial-1.8> (04.09.2023).
- Ebersbach, Anja / Glaser, Markus / Heigl, Richard (2016): *Social Web*. 3., überarbeitete Auflage. Konstanz: UVK.
- Gaffney, Devin / Matias, J. Nathan (2018): Caveat Emptor, Computational Social Science: Large-Scale Missing Data in a Widely-Published Reddit Corpus. In: *PLOS ONE* 13: 7, 1-13. <https://doi.org/10.1371/journal.pone.0200162>.
- Gatto, Maristella (2014): *The web as corpus: theory and practice*. London / New York: Bloomsbury.
- Gilmore, Alex (2007): Authentic materials and authenticity in foreign language learning. In: *Language Teaching* 40: 2, 97-118. <https://doi.org/10.1017/S0261444807004144>.
- Haas, Alexander / Brosius, Hans-Bernd (2011): Interpersonal-öffentliche Kommunikation in Diskussionsforen – Strukturelle Äquivalenz mit der Alltagskommunikation? In: Wolling, Jens / Will, Andreas / Schumann, Christina (Hrsg.): *Medieninnovationen. Wie Medienentwicklungen die Kommunikation in der Gesellschaft verändern*. Konstanz: UVK, 103-119.
- Hardie, Andrew (2012): CQPweb — combining power, flexibility and usability in a corpus analysis tool. In: *International Journal of Corpus Linguistics* 17: 3, 380-409.
- Johns, Tim (1991): Should you be persuaded: Two samples of data-driven learning materials. In: *ELR Journal* 4, 1-16.
- Kaltenböck, Gunther / Mehlmauer-Larcher, Barbara (2005): Computer corpora and the language classroom: on the potential and limitations of computer corpora in language teaching. In: *ReCALL* 17: 1, 65-84. <https://doi.org/10.1017/S0958344005000613>.
- Kaltwasser, Dennis (2019): *Forenkommunikation in Onlinezeitungen. Pressekommunikation im medialen Wandel*. Linguistische Untersuchungen, Bd. 13. Gießen: Giessen University Library Publications.
- Kehoe, Andrew (2020): Web Corpora. In: Paquot, Magali / Gries, Stefan Th. (Hrsg.): *A Practical Handbook of Corpus Linguistics*. Berlin: Springer, 329-351.
- Klemm, Michael (2012): Doing being a fan im Web 2.0. Selbstdarstellung, soziale Stile und Aneignungspraktiken in Fanforen. In: *Zeitschrift für germanistische Linguistik* 2012: 56, 3-32. <https://doi.org/10.1515/zfal-2012-0002>.
- Koch, Wolfgang (2022): Reichweiten von Social-Media-Plattformen und Messengern. Ergebnisse der ARD/ZDF-Onlinestudie 2022. In: *Media Perspektiven* 10, 471-478. https://www.ard-zdf-onlinestudie.de/files/2022/2210_Koch.pdf (04.09.2023).
- Linthe, Maja (2020): Texte zwischen Links und Likes: Die Textqualitätskategorien des Mannheimer Analyseleitfadens für Social-Media-Texte. In: *Deutsche Sprache* 48: 2, 126-145.
- Naef, Marcel (2021): „Und ich hab damals für dich ge-Voted...“ Online-Kommentarforen zwischen Text, Diskurs und Gemeinschaft. In: *Lublin Studies in Modern Languages and Literature* 45: 2, 27-38. <https://www.ceeol.com/search/article-detail?id=989769> (04.09.2023).
- Nodari, Claudio / Steinmann, Cornelia (2010): Lernerautonomie. In: Krumm, Hans-Jürgen / Fandrych, Christian / Hufeisen, Britta / Riemer, Claudia (Hrsg.): *Deutsch als Fremd- und Zweitsprache*. 2. Halbband. Berlin / New York: de Gruyter Mouton, 1157-1162.
- Pappert, Steffen / Roth, Kersten Sven (2019): Diskurspragmatische Perspektiven auf neue Öffentlichkeiten in Webforen. In: Hauser, Stefan / Opilowski, Roman / Wyss, Eva L. (Hrsg.): *Alternative Öffentlichkeiten. Soziale Medien zwischen Partizipation, Sharing und Vergemeinschaftung*. Bielefeld: transcript, 19-52.

- Pappert, Steffen / Roth, Kersten Sven (2016): Diskursrealisationen in Online-Foren. In: *Zeitschrift für Angewandte Linguistik* 65:1, 37-66.
- Peeters, Stijn / Hagen, Sal (2022): The 4CAT Capture and Analysis Toolkit. In: *Computational Communication Research* 4: 2, 571-589. <https://computationalcommunication.org/ccr/article/view/120> (04.09.2023).
- Pfurtscheller, Daniel (im Druck): Media as Cookie Cutters. Exploring the Digital Mediality of News on Instagram. In: Luginbühl, Martin / Schneider, Jan: *Media as Procedures*. Amsterdam: Benjamins.
- Reddit (2014): Häufig Gestellte Fragen. Reddit. https://www.reddit.com/wiki/de/faq/#wiki_h.E4ufig_gestellte_fragen (04.09.2023).
- Reddit (2021): Hallo and Guten Tag! Reddit Is Coming to Germany - Upvoted. Blog. <https://www.redditinc.com/blog/hallo-and-guten-tag-reddit-is-coming-to-germany> (04.09.2023).
- Reddit (2022): Herzlichen Glückwunsch! Reddit Celebrates One Year in Germany - Upvoted. <https://www.redditinc.com/blog/reddit-celebrates-one-year-in-germany> (04.09.2023).
- Riedner, Renate (2018): Authentizität in der Fremdsprachendidaktik – kritische Anmerkungen zu einem problematischen Konzept. In: *Deutsch als Fremdsprache* 55: 1, 34-43.
- Rivera, Ivan (2023): RedditExtractoR: Reddit Data Extraction Toolkit. <https://CRAN.R-project.org/package=RedditExtractoR> (04.09.2023).
- Schneider, Jan Georg (2018): Medialität. In: Liedtke, Frank / Tuchen, Astrid (Hrsg.): *Handbuch Pragmatik*. Stuttgart: J.B. Metzler, 272-281.
- Semrush (2023): reddit.com Website Traffic, Ranking, Analytics [February 2023]. *Semrush* <https://www.semrush.com/website/reddit.com/overview/> (01.04.2023).
- Spitzmüller, Jürgen / Warnke, Ingo Hans Oskar (2011): *Diskurslinguistik: Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse*. Berlin / New York: de Gruyter.
- Storrer, Angelika (2014): Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze - empirische Befunde. In: Plewina, Albrecht / Witt, Andreas (Hrsg.): *Sprachverfall? Dynamik - Wandel - Variation*. Berlin / Boston: de Gruyter, 171-196.
- Storrer, Angelika (2018): Interaktionsorientiertes Schreiben im Internet. In: Deppermann, Arnulf / Reineke, Silke (Hrsg.): *Sprache im kommunikativen, interaktiven und kulturellen Kontext*. Berlin / Boston: de Gruyter, 219-244.
- Storrer, Angelika (2020): Textqualität digital: Ein Modell zur Qualitätsbewertung digitaler Texte. In: *Deutsche Sprache* 48: 2, 101-125.
- Tonhauser, Ingo (2020): Textarbeit im Fremdsprachenunterricht als Frage fachdidaktischer Kompetenz. In: *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 25: 1, 1451-1470.
- Thonhauser, Ingo / Hufeisen, Britta (2016): Authentische, didaktisierte und didaktische Texte – Überlegungen zur Textarbeit aus drei verschiedenen Perspektiven. In: Freudenberg-Findeisen, Renate (Hrsg.): *Auf dem Weg zu einer Textsortendidaktik. Linguistische Analysen und text(sorten)didaktische Bausteine nicht nur für den fremdsprachlichen Deutschunterricht*. Hildesheim et al.: Olms, 149-164.
- Varantola, Krista (2003): Translators and disposable corpora. In: Zanettin, Federico / Bernardini, Silvia / Stewart, Dominic (Hrsg.): *Corpora in Translator Education*. Manchester: St. Jerome, 55-70.
- Vergeiner, Elisabeth (2021): Interaktive Webtools für den Grammatikunterricht. In: *ÖDaF-Mitteilungen* 37: 2, 160-169. <https://doi.org/10.14220/odaf.2021.37.2.160>.

Würffel, Nicola (2021): Lehr- und Lernmedien. In: Altmayer, Claus / Biebighäuser, Katrin / Haberzettl, Stefanie / Heine, Antje (Hrsg.): *Handbuch Deutsch als Fremd- und Zweitsprache: Kontexte – Themen – Methoden*. Stuttgart: J.B. Metzler, 282-300.

Biographische Notiz: Daniel Pfurtscheller ist seit 2021 Assistenzprofessor am Institut für Germanistik der Universität Innsbruck. Zu seinen Forschungsschwerpunkten zählt Medienlinguistik, Internetlinguistik, Multimodalitätsforschung und Interaktion in digitalen Medien.

Kontaktanschrift:

Ass.-Prof. Mag. Dr. Daniel Pfurtscheller
Institut für Germanistik
Universität Innsbruck
Innrain 52, 6020 Innsbruck
Österreich
daniel.pfurtscheller@uibk.ac.at

