Erschienen in: OWIDplus. - Mannheim: Leibniz-Institut für Deutsche Sprache (IDS), 2020. *https://www.owid.de/plus/cowidplus2020/*
und in: Aktuelle Stellungnahmen zur Sprache in der Coronakrise, erschienen am 06.05.2020. *https://www1.ids-mannheim.de/sprache-in-der-coronakrise/*

Sascha Wolfer / Alexander Koplenig / Frank Michaelis / Carolin Müller-Spitzer

cOWIDplus ANALYSE: WIE SEHR SCHRÄNKT DIE CORONA-KRISE DAS VOKABULAR DEUTSCHSPRACHIGER ONLINE-PRESSE EIN?

Die Autor*innen sind Mitarbeiter*innen des Programmbereichs Lexik empirisch und digital in der Abteilung Lexik am Leibniz-Institut für Deutsche Sprache, Mannheim.

cOWIDplus Analyse ist eine kontinuierlich aktualisierte Ressource zu der Frage, ob und wie stark sich der Wortschatz ausgewählter deutscher Online-Pressemeldungen während der Corona-Pandemie systematisch einschränkt und ob bzw. wann sich das Vokabular nach der Krise wieder ausweitet. In diesem Artikel erläutern wir die hinter der Ressource stehende Forschungsfrage, die zugrunde gelegten Daten, die Methode sowie die bisherigen Ergebnisse.

Corona und Sprache

Die Corona-Pandemie beeinflusst fast jede Facette des öffentlichen Lebens, und das praktisch auf der ganzen Welt. Es ist dabei nur allzu verständlich, dass die Pandemie nicht nur in persönlicher Face-to-Face-Kommunikation (direkt oder digital) einen großen Teil einnimmt, sondern auch die Nachrichten großflächig beherrscht. Große Teile des öffentlichen Lebens sind eingestellt oder starken Restriktionen unterworfen, was selbstverständlich einen Einfluss auf die Berichterstattung hat. Kulturredaktionen gehen die öffentlichen Veranstaltungen aus, über die sie berichten können. Dasselbe gilt für Sportredaktionen. Doch auch Politik- und Wirtschaftsressorts müssen auf die Krise reagieren und berichten über die Effekte, die die Pandemie auf die Gesellschaft hat. Dazu gehören Kontaktsperren, Hilfspakete oder andere Maßnahmen, die von Regierungsseite aus unternommen werden.

Dies alles legt die Vermutung nahe, dass sich das Vokabular, das in Artikeln verwendet wird – und zwar nicht nur in Print-, sondern auch in Onlinemedien – wandelt. Um es präziser auszudrücken: Es ist eine Einschränkung des Vokabulars auf Gegenstandsbereiche um die Corona-Pandemie zu erwarten. Das muss nicht unbedingt bedeuten, dass weniger verschiedene Wörter (hier i.S.v. Wortformen) verwendet werden. Vielmehr legt diese Annahme nahe, dass es eine Verschiebung der Häufigkeitsverteilung über Wörter hinweg gibt, und zwar in der Form, dass sich die Verteilung zugunsten (temporär) wichtiger Wörter verschiebt. Anders ausgedrückt: Weniger Wörter, d. h. Types, vereinigen mehr Tokens, d. h. mehr Wortvorkommen, auf sich. Solch eine Veränderung ist durch quantitative Maße detektier- und messbar. Wir werden unten darstellen, wie wir die angenommene Veränderung in der Häufigkeitsverteilung operationalisieren.

Fragestellung

Aus diesen Überlegungen leitet sich unsere zweiteilige Fragestellung ab, die folgendermaßen formuliert werden kann:
Können wir eine quantitativ messbare Einengung der behandelten Themen in der deutschen Online-Presse während der Corona-Pandemie feststellen? Wenn das so ist: Lässt sich in diesen quantitativen Messungen auch feststellen, wann die Krise nicht mehr das beherrschende Thema ist? Themen sollen in unserem Fall abgelesen werden an dem verwendeten Vokabular in den Online-Pressemeldungen.

Hypothesen

Unsere daraus abgeleiteten Hypothesen sind zweigeteilt, denn wir können eine Hypothese zu Sprachdaten formulieren, die uns bereits vorliegen und eine weitere für Sprachdaten, die wir erst noch sammeln werden.

- *Hypothese 1* (zu bereits vorhandenen Daten): Ja, es gibt eine Verengung der behandelten Themen in der deutschen Online-Presse, die wir mit ziemlich einfach zu berechnenden quantitativen Maßzahlen erfassen können. Mit einem eher qualitativen Ansatz können wir außerdem durch das Betrachten von Frequenzlisten feststellen, dass es in der Tat Pandemie- oder Corona-spezifisches Vokabular ist, das von dieser Einengung „profitiert“ und sowohl obere Frequenzränge einnimmt als auch viele Treffer auf sich vereint.

Quelle	Titel	Beschreibung
FAZ 1.4.2020	Französische Kontroverse: Alain Finkielkraut widerlegt Giorgio Agamben und Peter Sloterdijk	Wir bleiben eine Zivilisation: Der französische Philosoph Alain Finkielkraut geht mit Äußerungen von Giorgio Agamben und Peter Sloterdijk zur Corona-Krise hart ins Gericht.
NZZ 9.4.2020	Deutschland zählt seit Januar erstmals weniger aktive Fälle – das Coronavirus in 16 Grafiken	Wie stark ist Deutschland vom Virus betroffen? Was unternehmen andere Länder im Kampf gegen Sars-CoV-2? Die wichtigsten Daten und Fakten zum Coronavirus.
heise 8.4.2020	Buzzfeed Deutschland sucht Käufer – Kurzarbeit in Medienhäusern	Die Werbeausfälle haben Folgen für Medienhäuser. Viele setzen auf Sparmaßnahmen, daneben sollen neue Einnahmequellen erschlossen werden.“
FAZ 1.3.2020	Coronavirus: Erste Todesfälle in Amerika und Australien	Hamsterkäufe und Kursverfall auf der einen, zur Besonnenheit mahnende Worte von Politikern und Experten auf der anderen Seite: Das Virus breitet sich weiter aus. In Hessen wurden vier weitere Infektionen gemeldet. Zwei Länder melden die ersten Todesfälle.

Tab. 1: Zufälliger Auszug aus dem RSS-Feed-Datensatz

- *Hypothese 2* (zu noch zu erhebenden Daten): Wenn die Pandemie und ihre Konsequenzen mehr unter Kontrolle sind, wird sich die Situation normalisieren. Dementsprechend wird das Vokabular in den verwendeten Quellen wieder expandieren, bis ein Stand erreicht ist, der dem Prä-Pandemie-Niveau entspricht.

Daten

Unsere Datengrundlage sind öffentlich zugängliche RSS-Feeds von 13 deutschsprachigen Quellen. RSS-Feeds (übersetzt etwa: Nachrichteneinspeisungen) werden verwendet, um Artikel einer Webseite oder deren Kurzbeschreibungen in maschinenlesbarer Form bereitzustellen. Insbesondere für Nachrichtenmeldungen, die z. B. in Newstickern oder Liveblogs laufen, wird diese Technik benutzt. Für unsere Analysen haben wir RSS-Feeds von *Focus online*, *Frankfurter Allgemeine Zeitung*, *Frankfurter Rundschau*, *Süddeutsche Zeitung*, *Neue Zürcher Zeitung*, *Spiegel Online*, *Der Standard*, *tageszeitung*, *Die Welt*, *Die Zeit*, sowie als nicht-pressesprachliche Online-Quellen *web.de*, *t-online.de* und *heise.de* gesammelt. Diese Quellen sind nach Beliebtheit und zumindest minimaler Variabilität über deutschsprachige Länder ausgesucht (für Textbeispiele vgl. Tab. 1). Aus diesen RSS-Feeds werden alle Titel und Beschreibungen extrahiert, jegliche Interpunktion entfernt sowie alle Tokens in Kleinschreibung überführt. Für jeden Tag in unserem Datensatz werden daraufhin Uni- und Bigramm-Frequenzlisten erstellt, allerdings werden für *cOWDIplus Analyse* momentan nur die Unigramm-Frequenzlisten weiter ausgewertet. Die Auswertung der Bigrammlisten ist für weitere Update-Versionen geplant. Seit Januar 2020 sammeln wir diese Daten und setzen dies kontinuierlich fort.

Maße

Um die Diversität des verwendeten Vokabulars quantitativ zu erfassen, verwenden wir die folgenden Maße:[1]

Redundanz: Redundanz beschreibt, wie viel Informationen in einer Informationsquelle mehrfach vorhanden sind. Angewendet auf die vorliegenden Daten könnte man zum Beispiel vermuten, dass je stärker sich das Vokabular auf einige wenige sehr häufige Types verdichtet, desto höher wird die Redundanz in den Daten. Redundanz ist ein Entropie-basiertes Maß.

Mean segmental type-token ratio (MSTTR): Das Verhältnis (ratio) zwischen der Anzahl unterschiedlicher Wörter (Types) und der Anzahl laufender Wörter (Tokens), wird oft als ein Maß für lexikalische Diversität oder Vielfalt benutzt. Die Type-Token-Ratio (TTR) ist ebenfalls zwischen 0 und 1 skaliert. Je niedriger die Type-Token-Ratio für ein Korpus ist, desto weniger divers, d.h. desto eingeschränkter ist das Vokabular. Da die TTR von der Größe des Korpus beeinflusst wird, wird häufig die mean segmental type-token ratio berechnet.

Dabei wird ein Korpus in gleich große Teile aufgeteilt (hier: 500 Tokens) und die TTR wird für jeden einzelnen Teil berechnet (meist wird dabei der letzte Teil ignoriert, da dieser weniger als 500 Token umfasst). Der Durchschnittswert der einzelnen Teil-TTRs ist die mean segmental type-token ratio.

Frequenzanteil der TOP100: Wir benutzen dieses Maß als einen weiteren Indikator für die Diversität des Vokabulars. Zunächst werden alle Types in absteigener Tokenfrequenz sortiert. Dann wird der Anteil der Tokenfrequenz der 100 frequentesten Typen an der Gesamttoken-Anzahl berechnet. Damit ist auch dieses Maß zwischen 0 und 1 skaliert: je höher der Wert, desto mehr Tokens entfallen in der Verteilung auf die 100 häufigsten Types.

Ergebnisse

Einschränkung des Vokabulars

Für die Redundanz und den Frequenzanteil der TOP100-Types scheint sich die erste Hypothese zu bestätigen (vgl. Abb. 1): erstens steigt die Redundanz über die Zeit, zweitens sinkt die MSTTR und drittens steigt auch der Frequenzanteil der 100 häufigsten Typen. Interessant sind aktuell (Ende April 2020) die Werte am Ende der jeweiligen Kurven, in denen alle Maße eine entgegengesetzte Entwicklung zu nehmen scheinen. Die Redundanz sinkt erneut, die MSTTR steigt und auch der Frequenzanteil der TOP100-Types geht wieder zurück. Die entsprechenden Anpassungslinien (in den Abbildungen die blauen Linien) befinden sich ungefähr auf einem Niveau von Anfang März 2020. Dies deutet darauf hin, dass sich die Diversität der analysierten Pressemeldungen bereits jetzt normalisieren könnte. Ob diese Werte aber bereits eine Trendwende einleiten, werden erst die folgenden Wochen zeigen. Auch ist an diesen Werten allein noch nicht abzulesen, ob die Corona-Krise insgesamt im Vokabular nicht mehr so stark im Fokus steht, oder ob das Corona-Vokabular selbst weiter ausdifferenziert wird.

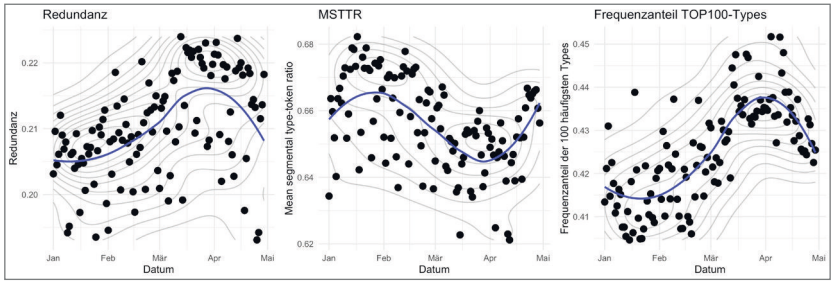

Abb. 1: Verlauf der verschiedenen Diversitätsmaße seit Januar 2020

Ein überraschendes Ergebnis zeigt Abbildung 2: Auch wenn in diesem Kontext nicht von einem kausalen Zusammenhang auszugehen ist, ähnelt der zeitliche Verlauf der gemessenen Redundanz visuell stark dem Verlauf der täglichen Corona-Neuinfektionen in Deutschland.

Rolle des Corona-bezogenen Vokabulars

Um einen Eindruck davon zu bekommen, worauf die quantitativ gemessene Einschränkung des Vokabulars zurückzuführen ist, zeigen wir auf *cOWIDplus Analyse* wöchentliche Frequenzlisten, die aus den tagesbasierten Listen aggregiert werden (für einen Auszug vgl. Abb. 3). Wenn man dabei die hochfrequenten Funktionswörter aus den 20 häufigsten Worttypen ausschließt, kann man sehen, wie ab der vierten Woche beginnend am 22. Januar 2020 die Wortform *coronavirus* auf Platz 1 der Frequenzliste steht, und das obwohl erst am 27. Januar 2020 von der Johns Hopkins Universität der erste Fall in Deutschland verzeichnet wird. Allerdings gibt es zwei Ausnahmen: In Woche 6 fand die Wahl des thüringischen Ministerpräsidenten statt und in der achten Woche verübte in Hanau ein Mann einen Anschlag. In diesen Wochen verdrängt u. a. *thüringen* bzw. *hanau coronavirus* vom ersten Platz. In Woche 10 tritt erstmals ein weiterer Eintrag hinzu, der klar

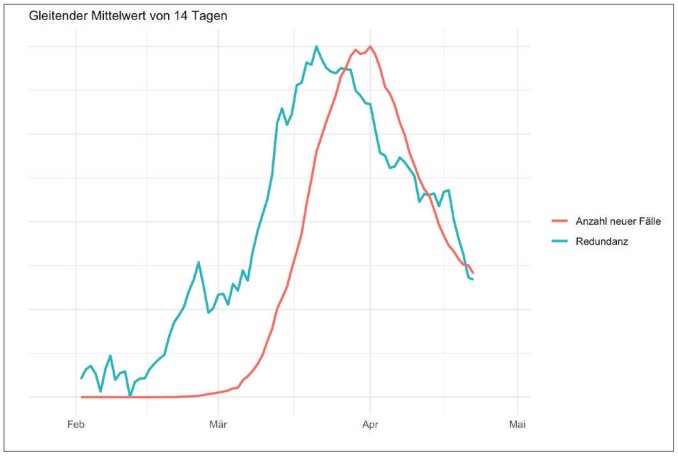

Abb. 2: Korrelation zwischen neu erfassten Corona-Infektionen und der gemessenen Redundanz in den RSS-Feeds

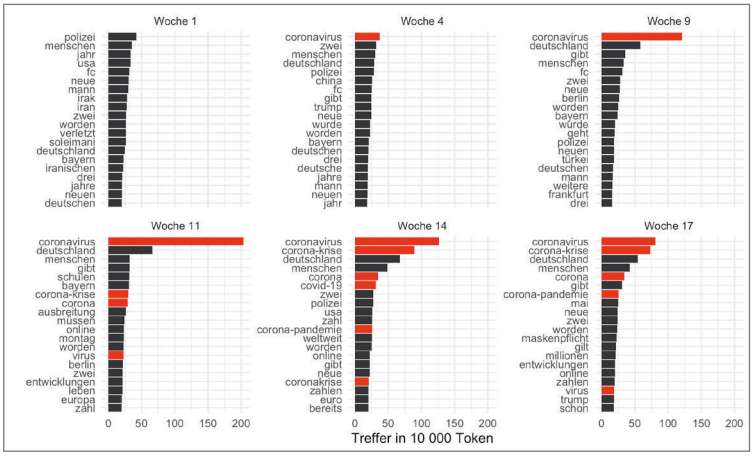

Abb. 3: Ausgewählte Wochenfrequenzlisten mit den 20 häufigsten Wortformen (ohne Funktionswörter)

mit der Pandemie verknüpft ist: „covid-19“. „corona-pandemie“ selbst befindet sich erst in Woche 13 zum ersten Mal unter den 20 frequentesten Typen. In der 17. Woche tritt zum ersten Mal die Wortform *maskenpflicht* unter die 20 frequentesten Formen. Diese wochenbasierte Analyse zeigt damit, dass es insbesondere Corona-spezifisches Vokabular ist, das für die Einschränkung des Vokabulars verantwortlich ist.

Die erste oben genannte Hypothese kann also bereits (vorläufig) bestätigt werden: Es gibt eine deutlich messbare Verengung der behandelten Themen in der deutschen (Online-)Presse. Wir können zudem in den wochenbasierten Analysen zeigen, dass dafür v.a. das Vokabular rund um die Corona-Krise verantwortlich ist. Ob wir jetzt schon wieder in eine Phase eintreten, in der diese Einengung sich abschwächt, worauf die letzten zwei Wochen hindeuten, sodass also nur Mitte März diese stark messbare Einschränkung festzustellen ist, werden die folgenden Wochen zeigen. Dementsprechend können wir auch erst in einiger Zeit überprüfen, ob sich die zweite Hypothese bestätigen lässt.
Da die RSS-Feeds sich auch als eine interessante Datengrundlage erwiesen haben, um einzelne Frequenzverläufe von Wörtern zu untersuchen (vgl. z. B. den Beitrag von Annette Klosa-Kückelhaus zu Maske oder Mundschutz), haben wir ein zweites Tool, den cOWIDplus Viewer, entwickelt, in dem man die Daten selber explorieren kann. Mehr Informationen dazu finden Sie auch in dem Artikel „cOWIDplus Viewer: Sprachliche Spuren der Corona-Krise in deutschen Online-Nachrichtenmeldungen. Explorieren Sie selbst!“.

Anmerkung

[1] Eine genauere Erläuterung der verwendeten Maße mit den zugehörigen Formeln findet sich unter <www.owid.de/plus/cowidplus2020/>.

Quellen (in der Reihenfolge ihrer Nennung)

cOWIDplus Analyse: <www.owid.de/plus/cowidplus2020/>.

Redundanz (Informationstheorie). Wikipedia. <https://de.wikipedia.org/w/index.php?title=Redundanz_(Informationstheorie)&oldid=191969545>.

Entropie (Informationstheorie). Wikipedia. <https://de.wikipedia.org/w/index.php?title=Entropie_(Informationstheorie)&oldid=194939230>.

Type-Token-Relation. Wikipedia. <https://de.wikipedia.org/w/index.php?title=Type-Token-Relation&oldid=197204797>.

Klosa-Kückelhaus, Annette. Maske oder Mundschutz?, <www1.ids-mannheim.de/fileadmin/aktuell/Coronakrise/klosa_mundschutz.pdf>.

Alle Webseiten zuletzt eingesehen am 4.5.2020.