Sascha Wolfer/Alexander Koplenig/Frank Michaelis/
Carolyn Müller-Spitzer/Jan Oliver Rüdiger (Mannheim)

Wie können wir den Einfluss der Corona-Pandemie auf die Verteilungen im deutschen Online-Pressewortschatz messen und explorieren?

Schon bald nachdem die ersten Coronavirus-Infektionsfälle auch in Deutschland bestätigt wurden, deutete sich an, dass die gesellschaftlichen Auswirkungen der Pandemie immens sein würden. Es war daher teilweise vorauszusehen, dass die Pandemie auch ihren Niederschlag in der Sprache finden würde. Und doch ist erstaunlich, wie weitreichend und tiefgreifend das Pandemiegeschehen und die gesellschaftlich-politischen Reaktionen Einfluss auf unseren Sprachgebrauch übten und üben, insbesondere auf der Ebene des Wortschatzes. Wir stellen zwei interaktive Ressourcen (*cOWIDplus Viewer* und *OWIDplusLIVE*) vor, die einen explorativen Zugang zur Erforschung dieses Einflusses bieten. Damit sollen der sprachwissenschaftlichen Forschungsgemeinschaft zwei Instrumente an die Hand gegeben werden, mit denen sich sprachliche Entwicklungen in der Zukunft möglichst unmittelbar entdecken und anhand von Frequenzverläufen nachzeichnen lassen.

Das zugrundeliegende Korpus besteht aus Titeln und kurzen Einführungstexten (sog. RSS-Feeds) zu Artikeln aus (derzeit) 13 deutschsprachigen Online-Quellen und umfasst am 4. Oktober 2021 ca. 53 Millionen Token und ca. 791.000 Types. Die Daten sind auch in Form von täglichen Unigramm- und Bi-Gramm-Frequenzlisten (inkl. Wortarten-Tagging) frei auf *OWIDplus* (www.owid.de/plus/cowid-plus2020) verfügbar.

cOWIDplus Viewer

Abb. 1: Startseite des cOWIDplus Viewer (mit hervorgehobenen Bereichen, auf die wir im Text Bezug nehmen)

Der *cOWIDplus Viewer* (<http://www.owid.de/plus/cowidplusviewer2020/>) bietet die Möglichkeit, mit einem Web-Interface (siehe Abb. 1) Uni- und Bi-Gramm-Frequenzen im RSS-Korpus zu explorieren. Neben einer tagesgenauen Selektion der zu untersuchenden Zeitperiode (in Abb. 1 durch 2 hervorgehoben), verschiedenen Such- (1) und Glättungsmodi (3) können die Ergebnisse auch in Tabellen- und Abbildungsform heruntergeladen (4) werden. Eine Bi-Gramm-Suche (5) erlaubt es zudem, potentiell interessante Bi-Gramme über deren Einzelkomponenten aufzufinden, um im zweiten Schritt deren Frequenzverlauf zu visualisieren. Für weitere Informationen zum *cOWIDplus Viewer* sowie beispielhafte Auswertungen siehe Müller-Spitzer et al. (2021) und Wolfer et al. (2020). Der *cOWIDplus Viewer* basiert auf *Shiny* (Chang et al. 2020), einer Technologie, die relativ unkompliziert die Erstellung interaktiver dynamischer Webapplikationen erlaubt. Allerdings ist *Shiny* nicht auf die Anforderungen des immer weiterwachsenden Datenbestands ausgelegt. Zudem konnten wir aufgrund dieser Beschränkungen einige Funktionalitäten, die uns sinnvoll erschienen, nicht umsetzen. Deshalb haben wir uns dazu entschieden, die Ressource in technologisch veränderter und deutlich erweiterter Form fortzuführen.

OWIDplusLIVE wurde mit dem Ziel entwickelt, eine flexible und performante Lösung unabhängig vom Gegenstand COVID-19 zu bieten. Dabei erstreckt sich die Flexibilität auf die Bereiche (1) zusätzliche Annotations-Layer, nämlich Lemma und Wortart (Part-of-Speech, POS), (2) größere N-Gramme (aktuell Tri-Gramme, aber auch die Möglichkeit N-Gramme größer 3 zu erfassen) sowie (3) die Möglichkeit zusätzlicher Visualisierungen. Dafür wurden sowohl die webbasierte Oberfläche als auch das dahinter liegende Daten-Backend von Grund auf neu entwickelt. Die bestehende Feed-Verarbeitungspipeline konnte ohne größere Änderungen übernommen werden.

Abb. 2: OWIDplusLIVE verteiltes Konzept

Zentral für den Ansatz hinter *OWIDplusLIVE* ist die gezielte Verzahnung von Technologien, die es ermöglichen, die Anwendung einfach mit neuen Daten (und ggf. Analysemöglichkeiten) zu erweitern, die Berechnungen über mehrere Server zu verteilen sowie Anfragen so effizient wie möglich zu verarbeiten. Abbildung 2 visualisiert diese Verzahnung. Weitere Informationen und technische Details sind unter <https://www.owid.de/plus> verfügbar.

Die Abfrage durch die Nutzer/-innen erfolgt auch weiterhin über eine webbasierte Oberfläche. Ein Großteil der Berechnungen und Visualisierungen findet im Browser der Nutzer/-innen statt. *OWIDplusLIVE* ist verfügbar unter <https://www.owid.de/plus/live-2021>. Sowohl die API als auch die webbasierte Oberfläche stehen als OpenSource (unter der AGPL-3.0 Lizenz) auf <https://github.com/notesjor/OWIDplusLIVE> zur Verfügung.

Abb. 3: Übersicht OWIDplusLIVE (inkl. manueller Hervorhebungen)

Abbildung 3 zeigt eine Übersicht der gesamten *OWIDplusLIVE* Web-Oberfläche. Ganz oben auf der Seite (Bereich 1) ist die Suche verortet. Wie Abbildung 4 zeigt, können Abfragen auf unterschiedlichen Layern erfolgen. Auf eine komplexe Such-Syntax wurde bewusst verzichtet, Platzhalter wie ‚?‘ und ‚*‘ sind jedoch möglich. Neben einer einfachen Suche auf unterschiedlichen Layern ist zudem eine erweiterte Tiefen-Suche integriert. Diese erlaubt es, pro N-Gramm-Position auf mehreren Layern zu suchen. Kombiniert mit Platzhaltern lassen sich so z. B. alle Lemmata finden, die die Zeichenfolge **bayer** enthalten und auf dem POS-Layer als Nomen annotiert wurden (Bsp.: *nordbayer*, *bayern*). Der Bereich 2 (Detail siehe Abb. 5) zeigt einen Zusammenschnitt der drei aktuell verfügbaren Visualisierungen¹. Der Bereich 3 (Detail siehe Abb. 6) zeigt den Suchverlauf inkl. einer Detail-Tabelle. Hier können einzelne Datenpunkte aus der Visualisierung ausgeschlossen werden.

Q SUCHEN: Einfache Suche ⓘ

Suchfenstergröße:

N=1   N=2   N=3

Suchabfrage:

Lemma 2   Suchausdruck (an 1. Position)... sprechen 3   POS 4   Suchausdruck (an 2. Position)... APPR 5

Beispiele:

Wortform = angela / Wortform = merkel | Lemma = sprechen / POS = APPR | POS = ADJ* / Lemma = merkel

Q SUCHE AUSFÜHREN

Abb. 4: Suche nach Bi-Gramm: 1. Lemma = sprechen / 2. POS = APPR (inkl. manueller Hervorhebungen)

Abbildung 4 illustriert eine einfache Suche nach Bi-Grammen auf unterschiedlichen Layern. Zuerst (1) wurde die Suchfenstergröße $N=2$ (Bi-Gramm) gewählt. An der ersten Position des Bi-Gramms wird auf dem Layer ‚Lemma‘ (2) nach ‚sprechen‘ (3) gesucht. An der zweiten Position wird auf dem POS-Layer (4) nach APPR (5) gesucht (APPR steht für die Wortart ‚Präposition; Zirkumposition links‘). Diese Abfrage ergibt somit Bi-Gramme wie „sprechen mit“, „spricht über“, „sprachen aufgrund“ usw.

¹ Für die Visualisierungen wurde die Bibliothek ECharts (Li et al. 2018) verwendet.

Abb. 5: Manueller Zusammenschnitt unterschiedlicher Visualisierungen (inkl. manueller Hervorhebungen)

Abbildung 5 zeigt, kompakt zusammengeschnitten, die aktuell verfügbaren Visualisierungen. Diese können unter Bereich (1) gewählt werden. Zur Verfügung steht ein tagesbasierter Frequenzverlauf (2), eine Kalenderansicht (3) und ein Sankey-Diagramm (4). Die Visualisierungen können über den unteren Bereich (5) angepasst werden. Es ist z. B. möglich, absolute und relative Frequenzen auszuwerten, eine Granulierung (Auswertung pro Tag, Woche, Monat, Quartal und Jahr) und davon abhängig einen gleitenden Mittelwert (zur Glättung) zu wählen.

Abb. 6: Auszug des Suchverlaufs (inkl. manueller Hervorhebungen)

Der Auszug der Detail-Ergebnisse im Suchverlauf (siehe Abb. 6) ermöglicht eine selektive Auswahl (1) der Ergebnisse. Die gesamten Daten einer einzelnen Suchabfrage können über das Dreipunkt-Menü (siehe Bereich 2) als JSON, TSV und URL exportiert werden, um die Daten weiterzugeben bzw. auch um die Daten mit anderen Programmen auszuwerten und zu visualisieren. Außerdem ist es möglich, den gesamten Suchverlauf (siehe Bereich 3), also alle Suchabfragen, als JSON zu exportieren und einen gespeicherten Suchverlauf wiederherzustellen.

OWIDplusLIVE stellt bereits jetzt eine Ressource für die tagesaktuelle Analyse sprachlicher Daten in RSS-Newsfeeds deutscher Online-Presse dar. Trotzdem gibt es an einigen Stellen Potential zur Weiterentwicklung. So könnten die analysierten Zeitabschnitte noch flexibler gestaltet werden, um auch Entwicklungen zu erfassen, die kleinteiliger als ein Tag (z. B. für die Analyse von Social-Media-Sprachdaten) oder grobkörniger als ein Jahr (z. B. für diachrone Analysen) sind. Außerdem sind zusätzliche Visualisierungen denkbar, die unterschiedliche Blickwinkel auf die Daten ermöglichen würden. Funktionale Ergänzungen wären bspw. die Integration von vordefinierten Visualisierungen in Homepages oder in Lernmanagementsysteme (z. B. Moodle).

Literatur

- Chang, Winston/Cheng, Joe/Allaire, Joseph J./Sievert, Carson et al. (2020): Shiny: Web application framework for R (Version 1.4.0.2). Internet: www.CRAN.R-project.org/package=shiny (Stand: 10.8.2021).
- Li, Deqing/Mei, Honghui/Shen, Yi/Su, Shuang/Zhang, Wenli/Wang, Junting/Zu, Ming/Chen, Wei (2018): ECharts: A declarative framework for rapid construction of web-based visualization. In: *Visual Informatics* 2, 2, S. 136–146.
DOI: <https://doi.org/10.1016/j.visinf.2018.04.011> (Stand: 10.8.2021).
- Müller-Spitzer, Carolin/Koplenig, Alexander/Michaelis, Frank/Wolfer, Sascha (2021): Wochenaktuelle lexikalische Spuren der Corona-Krise in deutschen Online-Nachrichtmeldungen. In: *Deutsche Sprache* 49, S. 1-23.
DOI: <https://doi.org/10.37307/j.1868-775X.2021.01.02> (Stand: 10.8.2021).
- Wolfer, Sascha/Koplenig, Alexander/Michaelis, Frank/Müller-Spitzer, Carolin (2020): Tracking and analyzing recent developments in German-language online press in the face of the coronavirus crisis: cOWIDplus Analysis and cOWIDplus Viewer.
In: *International Journal of Corpus Linguistics* 25, 3, S. 347–359. DOI: <https://doi.org/10.1075/ijcl.20078.wol> (Stand: 10.8.2021).