OWIDplusLIVE - Tagesaktuelle N-Gramm-Analysen

Rüdiger, Jan Oliver
ruediger@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Wolfer, Sascha
wolfer@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Cotgrove, Louis
cotgrove@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

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 Ressourcen (OWIDplusLIVE und das zugrundeliegende Live-RSS-Korpus) vor, die einen explorativen Zugang zur Erforschung dieses Einflusses bieten. Zudem soll der sprachwissenschaftlichen Forschungsgemeinschaft ein Instrument an die Hand gegeben werden, auch andere sprachliche Entwicklungen in der Zukunft möglichst unmittelbar zu entdecken und anhand von Frequenzverläufen nachzuzeichnen. Das folgende Beispiel (Abb. 1) zeigt vier nacheinander gestellte Suchabfragen zu den Bi-Grammen: *zweite* (in blau), *dritte* (grün), *vierte* (gelb) und *fünfte Welle* (rot) [Stand: 26. September 2022].

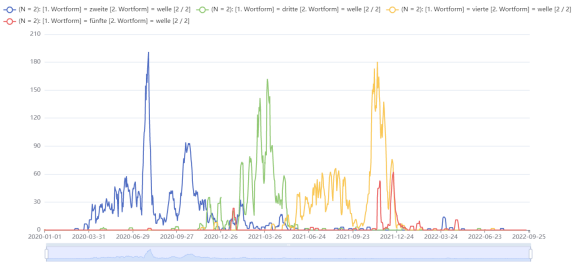

Abb. 1

Das zugrundeliegende Korpus besteht aus Titeln und kurzen Einführungstexten (sog. RSS-Feeds) zu Artikeln aus (derzeit) 13 deutschsprachigen Online-Quellen (Details zu den Quellen und zur Quellenauswahl siehe Vorprojekt: Wolfer u. a. 2020). Das Korpus wird seit dem 01.01.2020 täglich erhoben und umfasste am 26. September 2022 ca. 84,1 Millionen Token. Die Daten sind auch in Form von täglichen Unigramm- (inkl. Wortarten-Tagging) und Bigramm-Frequenzlisten frei auf OWIDplus (www.owid.de/plus/cowidplus2020) verfügbar.

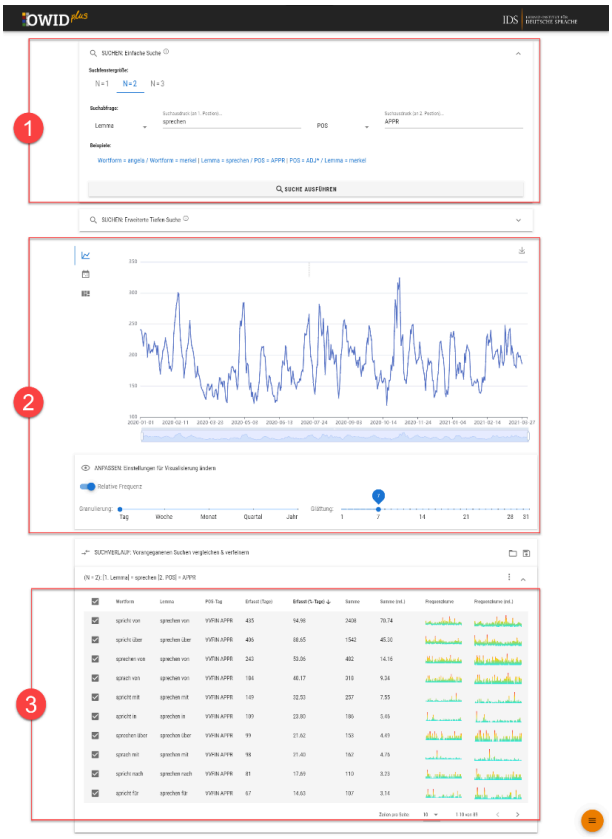

Abb. 2

OWIDplusLIVE wurde mit dem Ziel entwickelt, eine flexible und performante Lösung unabhängig vom Gegenstand (COVID-19) zu bieten. Damit löst dieses Tool den zuvor erstellten Prototypen ‚cOWIDplus Viewer' (Wolfer u. a. 2020) ab. OWIDplusLIVE professionalisiert den Prototypen in folgenden fünf Bereichen: (1) zusätzliche Annotations-Layer, konkret: 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 (4) wurden sowohl die webbasierte Oberfläche als auch das dahinterliegende Daten-Backend von Grund auf neu entwickelt. Die bestehende Feed-Verarbeitungspipeline konnte ohne größere Änderungen übernommen werden. Zentral für den Ansatz hinter OWIDplusLIVE ist die (5) gezielte Verzahnung von Technologien (Falk u. a. 2020; Banon u. a. 2022; You u. a. 2022), 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. Alle im Projekt entwickelten Komponenten (API und Web-Frontend) stehen kostenfrei als OpenSource (unter der AGPL-3.0 Lizenz) zur Verfügung - siehe: https://github.com/notesjor/IDS.OWID.Plus.Live

Die Abfrage durch die Nutzer*innen erfolgt 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. Die Oberfläche ist in drei Segmente eingeteilt, die im Folgenden benannt und weiter unten erklärt werden (siehe Abb. 2): (1) Der Abfragebereich. (2) Ein Bereich mit drei unterschiedlichen Visualisierungen. (3) Sowie die Detailansicht.

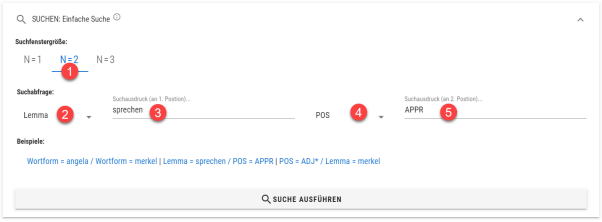

Abb. 3

Abb. 3 zeigt den Abfragebereich mit einer einfachen Suche nach Bi-Grammen auf unterschiedlichen Layern. Auf eine komplexe Such-Syntax wurde bewusst verzichtet. Platzhalter wie ‚?' und ‚*' sind jedoch möglich. 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.

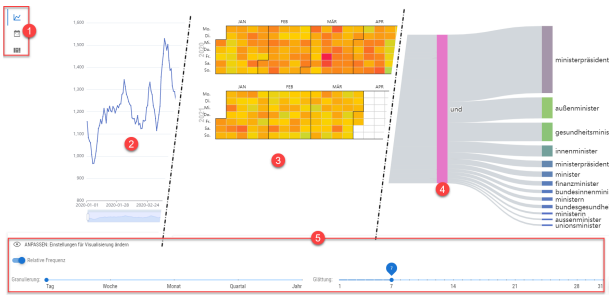

Abb. 4

Abb. 4 zeigt, kompakt zusammengeschnitten, die aktuell verfügbaren Visualisierungen. Diese können links (siehe Abb. 4 – Markierung 1) gewählt werden. Zur Verfügung steht ein tagesbasierter Frequenzverlauf (2 – siehe auch Abb. 1), 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 eine Glättung zu wählen.

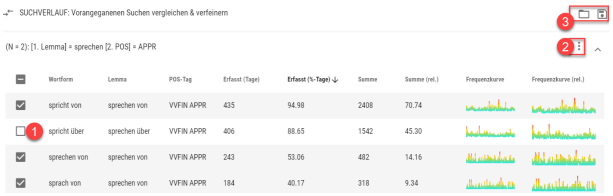

Abb. 5

Der Auszug der Detail-Ergebnisse im Suchverlauf (siehe Abb. 5) ermöglicht es, eine Teilmenge von Ergebnissen auszuwählen (1). 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.

Bibliographie

Banon, Shay und ‚Elastic NV contributors' . 2022. Elasticsearch. https://www.elastic.co/de/elasticsearch/ (zugegriffen: 28. Juli 2022).

Falk, Warren und ‚RocksDB contributors'. 2020. RocksDB. https://github.com/elastic/elasticsearch-net (zugegriffen: 28. Juli 2022).

Wolfer, Sascha; Koplenig, Alexander; Michaelis, Frank und 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.* https://doi.org/10.1075/ijcl.20078.wol (zugegriffen: 10. Oktober 2022).

You, Evan und ‚Vue.js contributors'. 2022. Vue.js. JavaScript. https://vuejs.org/ (zugegriffen: 28. Juli 2022).