57. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 9. bis 11. März 2021
(als Online-Konferenz)

SPRACHE IN POLITIK UND GESELLSCHAFT

Perspektiven und Zugänge

METHODENMESSE

Dienstag, 09. März 2021, 09:50 Uhr

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

Sascha Wolfer, Alexander Koplenig, Frank Michaelis, Carolin Müller-Spitzer, Jan Oliver Rüdiger (Leibniz-Institut für Deutsche Sprache)

Schon bald nach den ersten bestätigten Corona-Fällen in Deutschland 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 Wortschatzebene.

Wir möchten drei Ressourcen vorstellen, anhand derer wir für einen Teilbereich der deutschen Sprache diese Einflüsse mess- und sichtbar machen: (i) Ein RSS-Korpus deutschsprachiger Online-Presse, (ii) eine kontinuierlich aktualisierte Internetseite (cOWIDplus Analyse), auf der wir die thematische Konzentration auf das Pandemiegeschehen kontinuierlich anhand des o.g. Korpus überwachten und (iii) eine Web-Applikation (cOWIDplus Viewer), mit der das RSS-Korpus anhand von Uni- und Bigramm-Frequenzen exploriert werden kann.

Das RSS-Korpus besteht aus Titeln und kurzen Einführungstexten zu Artikeln aus 13 deutschsprachigen Online-Quellen, wird momentan 2-wöchentlich aktualisiert und umfasst derzeit ca. 31,6 Millionen Token und ca. 565 000 Types (Stand 20.1.2021). Es ist in Form von täglichen Unigramm- (inkl. Wortarten-Tagging) und Bigramm-Frequenzlisten frei herunterladbar auf der Webseite zur cOWIDplus Analyse. Auf dieser Seite zeigen wir außerdem, dass die Diversität des Vokabulars im Untersuchungskorpus zwischen Mitte März und Mitte April drastisch eingeschränkt ist – zeitgleich zur Hochzeit der ersten Pandemiewelle in Deutschland. Diese quantitative Analyse nutzt drei Maße (informationstheoretische Redundanz, mean segmental type-token ratio und den Frequenzanteil der häufigsten 100 Types pro Tag), die durchweg die Konzentration des Vokabulars anzeigen. Wöchentliche Frequenzlisten von Inhaltswörtern zeigen, dass diese Effekte hauptsächlich von mit der Corona-Pandemie verbundenen Worttypen stammen.

Der cOWIDplus Viewer bietet die Möglichkeit, mit einem Web-Interface die Uni- und Bigrammfrequenzen des RSS-Korpus zu explorieren. Neben einer tagesgenauen Selektion der zu untersuchenden Zeitperiode, verschiedenen Such- und Glättungsmodi erlaubt diese Web-Applikation auch, die Ergebnisse in Tabellen- und Abbildungsform herunterzuladen. Eine Bigramm-Suche erlaubt es zudem, potentiell interessante Bigramme über deren Einzelkomponenten aufzufinden, um im zweiten Schritt deren Frequenzverlauf zu visualisieren.

Da sich während der Corona-Pandemie herausgestellt hat, dass die cOWIDplus-Ressourcen ein wertvolles Instrument darstellen, um kurzfristige Veränderungen des Sprachgebrauchs zu untersuchen, arbeiten wir an einer neuen, erweiterten Lösung, die das Korpus zeitnaher zugreifbar macht und zusätzliche Viewer-Funktionen beinhaltet (z. B. Trigramme, Lemma- und Wortarten-Layer sowie zusätzliche Visualisierungen). Im Rahmen des Vortrags werden wir eine erste lauffähige Vorschauversion dieser neuen Ressource zeigen.