Jan Oliver Rüdiger/Sascha Wolfer/Alexander Koplenig/
Frank Michaelis/Carolin Müller-Spitzer/Samira Ochs/
Louis Cotgrove

OWIDplusLIVE

Day-to-day collection, exploration, analysis, and visualization of N-Gram frequencies in German (online press) language

Keywords LIVE-Data; N-Gram; German; data exploration; visualization

With OWIDplusLIVE, we would like to introduce the EURALEX community to two resources that provide analytical access to daily updated data (data: frequency data and N-grams – reference point: previous day).

The project started following the first confirmed COVID-19 cases in Germany. It was already clear at the time that the social impact of the pandemic would be immense. And yet, in retrospect, it is very surprising how broad and wide-reaching the influence of the pandemic has been, especially at the level of German-language vocabulary (see Wolfer et al. 2022). It remains to be seen how persistent and lasting these influences are, i. e., how many of the words that have found their way into everyday usage will continue to be consistently used in the future.

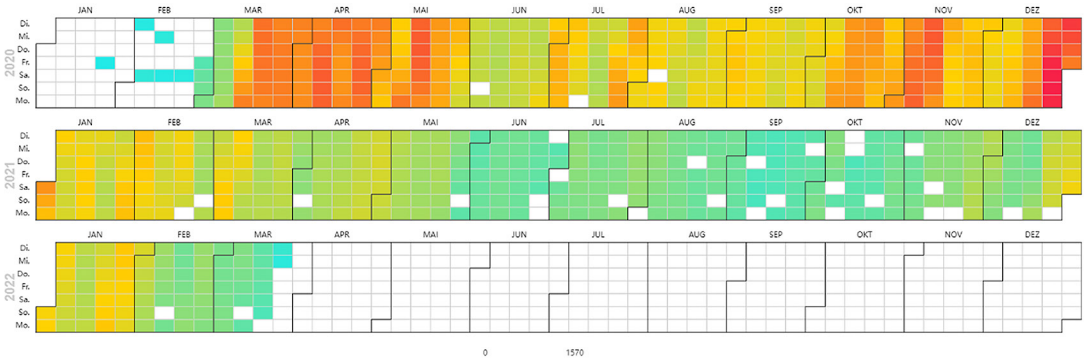

Fig. 1: OWIDplusLIVE – calendar view – query "corona VVFIN"

Against this background, it is necessary to develop instruments for monitoring language as close to real time as possible. This will allow the analysis of new events, conflicts and socially relevant topics.

Most "traditional" corpora don't allow to analyze corpus material in real-time (first notable projects in this field are e. g.: Davies 2013; Vogel et al. 2021; Barbaresi 2022). Therefore, we present two resources to track language use in a limited subset of the German language. The first resource is an RSS corpus containing titles and so-called "descriptions" (leads or teasers) to online newspaper articles from 13 German-language online sources (one each from Switzerland and Austria, 11 from Germany). Currently (2022-Mar-23), the corpus contains

This page is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
DOI: *https://doi.org/10.14618/t86d-e088*

about 67.9 million tokens in about 1.5 million feed items since January 1, 2020. The second resource is called OWIDplusLIVE (https://www.owid.de/plus/live-2021/), a web application that allows to investigate the frequency of uni-, bi- and trigrams in the mentioned RSS corpus. Users can perform searches at three different layers: word-form, lemma and part-of-speech (within our NLP pipeline we use (Schmid 1995) for automatic annotation). Wildcard searches are also possible. In order to make the application accessible to a wide range of user groups, we have deliberately avoided implementing a complex search syntax. We also provide a (German) video tutorial for a quick start.

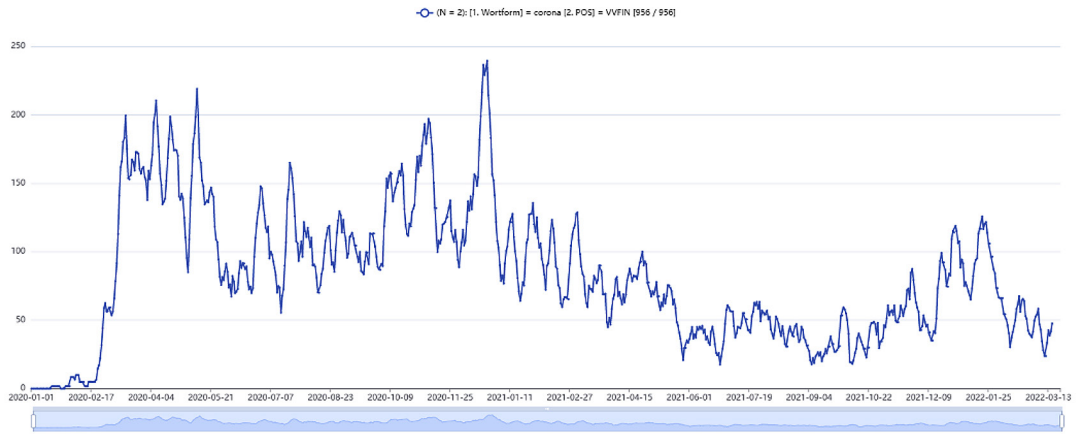

Fig. 2: OWIDplusLIVE – line chart – query "corona VVFIN"

Currently, three visualizations are available: 1) a calendar view (see Fig. 1) where each day is colored according to the (relative or absolute) frequency of the current day; 2) a line chart (see Fig. 2) showing the (relative or absolute) frequencies. Here, frequencies can also be compared between multiple queries; 3) a Sankey diagram (see Fig. 3) that reveals language usage patterns, especially for more complex queries. All visualizations have interactive elements such as mouseover information and/or zoom and selection options. All queries can be further explored in a list of all results. This feature is a key feature of this tool (e. g., compared to other similar N-Gram viewers) enabling the transparent exploration of results. In other words, you not only get an aggregated total frequency for the query "corona VVFIN" (word form "corona" followed by a finite verb) but also separate time series (these can be manually de/selected). This enables results to be fine-tuned according to the needs of the researcher. Furthermore, we want to avoid to have a 'black box'-software. Researchers can understand which time series are included in the visualization and what impact a time series has on the overall result.

Results, queries, visualizations, and result data can be shared and exported via a URL, JSON, TSV (tab-delimited text) format.

The source code and a documentation are freely available. In the context of lexicography, OWIDplusLIVE can be used for all issues related to vocabulary change and lexical innovation.

This page is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
DOI: *https://doi.org/10.14618/t86d-e088*

References

Barbaresi, A. (2022): Webmonitor. https://www.dwds.de/d/korpora/webmonitor (last access: 23-03-2022).

Davies, M. (2013): Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day. https://corpus.byu.edu/now/ (last access: 23-03-2022).

Schmid, H. (1995): TreeTagger. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (last access: 23-03-2022).

Vogel, F./Bäumer, B./Deus, F./Knobloch, C./Rüdiger, J. O./Schmallenbach, J./Schölzel, H./Tripps, F./Weber, S./Wilton, A. (2021): www.Diskursmonitor.de – gemeinschaftlich erarbeitete Online-Plattform zur Aufklärung und Dokumentation strategischer Kommunikation. http://doi.org/10.5281/ZENODO.5780230 (last access: 23-03-2022).

Wolfer, S./Koplenig, A./Michaelis, F./Müller-Spitzer, C./Rüdiger, J. O. (2022): Wie können wir den Einfluss der Corona-Pandemie auf die Verteilungen im deutschen Online-Pressewortschatz messen und explorieren? In: Kämper, H./Plewnia, A. (eds.): Sprache in Politik und Gesellschaft. Perspektiven und Zugänge. (= Jahrbuch des Instituts für Deutsche Sprache 2021). Berlin/Boston, pp. 331–338. http://doi.org/10.1515/9783110774306-022 (last access: 23-03-2022).

Contact information

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

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

Alexander Koplenig
Leibniz-Institut für Deutsche Sprache
koplenig@ids-mannheim.de

Frank Michaelis
Leibniz-Institut für Deutsche Sprache
michaelis@ids-mannheim.de

Carolin Müller-Spitzer
Leibniz-Institut für Deutsche Sprache
mueller-spitzer@ids-mannheim.de

Samira Ochs
Leibniz-Institut für Deutsche Sprache
ochs@ids-mannheim.de

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

This page is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
DOI: *https://doi.org/10.14618/t86d-e088*