Originally published in: International Journal of Corpus Linguistics vol. 25 (2020) no. 3, pp. 347-359.
DOI: *https://doi.org/10.1075/ijcl.20078.wol*

POSTPRINT

Sascha Wolfer, Alexander Koplenig, Frank Michaelis, and Carolin Müller-Spitzer
Leibniz Institute for the German Language

Tracking and analyzing recent developments in German-language online press in the face of the coronavirus crisis

cOWIDplus Analysis and *cOWIDplus Viewer*

Abstract: The coronavirus pandemic may be the largest crisis the world has had to face since World War II. It does not come as a surprise that it is also having an impact on language as our primary communication tool. In this short paper, we present three inter-connected resources that are designed to capture and illustrate these effects on a subset of the German language: An RSS corpus of German-language newsfeeds (with freely available untruncated frequency lists), a continuously updated HTML page tracking the diversity of the vocabulary in the RSS corpus and a *Shiny* web application that enables other researchers and the broader public to explore the corpus in terms of basic frequencies.

Keywords: RSS newsfeed corpus, data visualization, linguistic diversity, vocabulary, coronavirus

1. The coronavirus pandemic and its influence on language

Around the globe, the COVID-19 pandemic has affected almost every aspect of public life. Consequently, at the time of writing, the pandemic is *the* subject of discussion, not only in private face-to-face conversation (if you still have the opportunity to talk to someone face-to-face at all) but also in the news. As many activities of daily life like sports and cultural events came to a halt, respective newspaper desks might very well run out of events to report on or shift their focus to pandemic-related topics. Other, more general desks like politics are also mainly dealing with the effects of the pandemic on society, for example impend-

ing crises in the health-care systems, curfews or other measures taken by national governments.

This leads to the assumption that the vocabulary used in news articles, not only in printed but also in online media, is changing. To be precise, we assume a (temporary) frequency rise for specific word forms that are associated with the all-encompassing coronavirus pandemic. This does not exclusively mean that we can observe word forms that have never been observed in every-day press language. Although some terminological vocabulary (especially from epidemiology and virology) might find its way into every-day press language, it is also very likely that concepts from every-day language which usually play a secondary role are suddenly much more important. The pandemic can be seen as a prototypical event that affects lots of people on a wide scale and supersedes news coverage of other events. We assume that another event as comprehensive as the coronavirus pandemic would have similar effects on language.

Several linguistic datasets focusing on the COVID-19 pandemic have been released. One example is the Coronavirus Corpus that is part of the English-Corpora.org suite of corpora (https://www.english-corpora.org/corona). This corpus consists of English articles harvested from the web that deal with the coronavirus. Note, however, that we did not make any restrictions on which items from the RSS newsfeeds are incorporated in our corpus. On a more general level, the NOW corpus (Davies, 2016–), of which the Coronavirus Corpus is a subset, provides a case in point that resources that make it possible to track real time changes in a corpus are of great value both in linguistics and in the digital humanities.

Another corpus resource is currently being built at the center for digital lexicography of the German language (ZDL). The idea of the "coronakorpus" is to collect language data from relevant URLs from a range of publication sources in the German language (e.g. blogs, forums, tweets, and general information websites on the topic but also online newspaper articles; see https://github.com/adbar/coronakorpus for updates). As with the Coronavirus Corpus mentioned above, the texts have to address the coronavirus pandemic to be included into the corpus. A dataset of tweets related to COVID-19 as well as an overview of similar datasets is available from https://data.gesis.org/tweetscov19.

To investigate the effects of the coronavirus pandemic on German press language, we created three inter-connected resources which we introduce in this paper: a corpus of RSS feeds, a dynamic HTML site which presents analyses on vocabulary diversity throughout the coronavirus pandemic (*cOWIDplus Analysis*) and an online viewer application that allows exploration of frequency profiles through time (the *cOWIDplus Viewer*). The names of the two resources are derived from *OWIDplus* (www.owid.de/plus), which is an experimental platform

for multilingual lexical-lexicographic data, for quantitative lexical analyses and for interactive lexical applications within the *Online-Wortschatz-Informationssystem Deutsch* (*OWID*, *www.owid.de*) provided by the Leibniz Institute for the German Language in Mannheim.

2. The RSS corpus

Since the beginning of 2020, we have been collecting a corpus of RSS feeds for 13 sources in German: *Focus Online* (https://www.focus.de), *Frankfurter Allgemeine Zeitung* (https://www.faz.net), *Frankfurter Rundschau* (https://www.fr.de), *Süddeutsche Zeitung* (https://www.sueddeutsche.de), *Neue Zürcher Zeitung* (https://www.nzz.ch), *Spiegel Online* (https://www.spiegel.de), *Der Standard* (https://www.derstandard.at), *tageszeitung* (https://taz.de), *Die Welt* (https://www.welt.de) and *Die Zeit* (https://www.zeit.de). In addition to these more "traditional" news outlets, we included three outlets from editorial offices that have no counterpart in printed form with the same name: web.de, t-online.de, and heise.de. The sources were selected for popularity, variety in German-speaking countries (one source each from Austria and Switzerland and 11 sources from Germany), and political orientation (for example, the *Frankfurter Allgemeine Zeitung* is described as a conservative newspaper whereas the *tageszeitung* is known to have a left-wing orientation. Also, *heise.de*, a source with a strong focus on information technology, is included in the sample. We use a custom *R* (R Core Team, 2020) script and the *XML* package (Temple Lang, 2020) to collect the raw feed data. The RSS feeds are extracted every hour. This leads to duplicated items in the collected data because some items can and will still be in the RSS feeds after one hour. We exclude these items by only considering unique entries in the data structure. Note that this procedure could still leave duplicates in our data because sometimes the editorial staff decide to make changes to already published content (e.g. a small change in the title of an article). Technically, however, these count as new items in the source's RSS feeds and are therefore treated as separate entries in our corpus. The corpus preparation is explained in Section 2.1, followed by an overview of the corpus size and monthly corpus measures in 2.2.

2.1 Corpus preparation

All titles and descriptions (short pieces of text that introduce the articles) are extracted from the sources' RSS feeds along with the associated publication timestamp. Due to differing timestamp formats, we have to normalize the timestamps using the *R* package *lubridate* (Grolemund & Wickham, 2011). We discard the

exact timestamp and only keep the day of publication. We exclude all punctuation from the titles and description (henceforth "texts") except hyphens because we want to keep compounds. HTML markup (e.g. <*strong*> or </*strong*>) that could be used in RSS feeds is also excluded from the texts. All characters are converted to lower-case.

Some specific items are excluded from the texts, mainly because they are very specific to certain sources. The following list is the comprehensive set of word forms that are excluded following the above-mentioned procedure: *t-onlinede-redakteurin, t-onlinede-redakteur* (*redakteur* is German for "editor", *redakteurin* is the female form), *sport-live-blog, t-onlinede, focus-online-redakteurin, focus-online-redakteur, focus-online-reporter, spiegel-titelstory, faz-sprinter, heise, der-standardat, km/h*. Also, all *YouTube* links and digit-only wordforms are excluded. After these pre-processing steps, unigram and bigram frequency lists are created for each day.

2.2 Corpus size

Given the continuous nature of corpus collection, we can only provide a snapshot of the current (sub)corpus sizes. The following figures reflect the state of the corpus up to July 2nd, 2020. At the time of writing, the RSS corpus contains 13,649,334 tokens which are distributed over 329,463 types. The corpus sizes for each month are summarized in Table 1. Figure 1 breaks down token sizes to days. A linear model (the grey regression line included in Figure 1) does not show a significant effect of the date on the size of the corpus on that day ($\beta = 8.95$, $SE = 10.3$, $t = 0.87$, $p = 0.385$). This is confirmed by a permutation test with 50,000 replications resulting in $p_{\text{perm}} = 0.389$.

Table 1. Monthly corpus measures for the RSS corpus. Please note that the figures for July only include July 1st and 2nd

Month	Number of tokens	Share of tokens	Number of types
January 2020	2,291,576	16.8 %	113,978
February 2020	2,196,288	16.1 %	108,267
March 2020	2,311,588	16.9 %	102,570
April 2020	2,107,303	15.4 %	96,912
May 2020	2,268,369	16.6 %	105,173
June 2020	2,302,423	16.9 %	107,454
July 2020 (only 2 days)	171,787	1.3 %	21,517

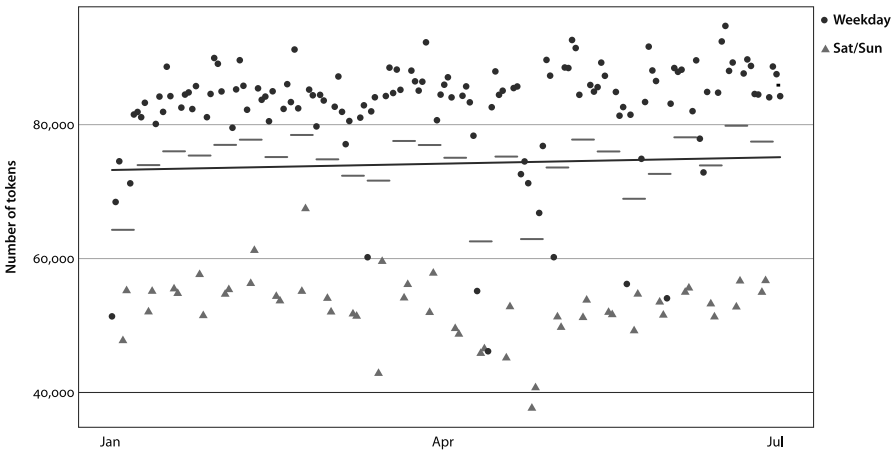

Figure 1. Daily corpus sizes of the RSS corpus with a linear model fit. Shapes indicate weekday or weekend (not taking into account holidays that did not fall on a weekend, e.g. May 1st). Horizontal line segments indicate the weekly mean (with day 1 of week 1 being January 1st, 2020)

Given the monthly corpus measures, we conclude that the size of the corpus can be thought of being more or less constant over time. To account for minor fluctuations (especially weekends vs. weekdays), we decided to report relative frequencies in the *cOWIDplus Viewer* (see Section 4).

3. *cOWIDplus Analysis*

With *cOWIDplus Analysis* (www.owid.de/plus/cowidplus2020), we provide a continuously updated resource in the form of an RMarkdown (Xie et al., 2018) HTML document that summarizes the findings (in German) with respect to the following research question: is there a quantitatively measurable narrowing of topics (and hence of the vocabulary) during the coronavirus pandemic in selected (online) news outlets published in the German language? If so, are topics (and hence the vocabulary) expanded after the crisis is "resolved"? Connected to this research question, we are testing two hypotheses:

i. Yes, there is indeed a narrowing of topics that can be measured by rather straightforward quantitative measures (a clear narrowing of vocabulary is indicated by all of the measures and by the investigation of frequency lists);
ii. When the pandemic and its consequences are under control, the situation will normalize and the vocabulary in online news outlets will diversify again in a continuous way (i.e. the variables used to quantify vocabulary diversity will return to the pre-pandemic level).

We operationalize the "control level" of the pandemic by seven indicators (from which five have to be met) that are directly related to daily life: (i) no contact bans, (ii) free travel, re-opening of (iii) schools and (iv) restaurants, (v) no restrictions regarding the number of clients in grocery stores, resumed (vi) sports competitions and (vii) cultural events with audiences.

The quantitative measures we use to operationalize the narrowing of the vocabulary are information-theoretic redundancy (Shannon, 1948: 14), mean segmental type-token ratio (MSTTR; Johnson, 1944: 3) and the accumulated token frequency share of the 100 most frequent word forms per day. We update the resource each week (possibly switching to updates every two weeks in the future) with the latest data and continuously evaluate the hypotheses. To enable the scientific public to replicate the analyses and to investigate further research questions, the following data is available for download on the *cOWIDplus Analysis* website: (i) daily frequency lists of unigrams (POS-tagged and lemmatized) and bigrams, (ii) weekly frequency lists (of unigrams) and (iii) the daily values for the central measures mentioned above (redundancy, MSTTR, and top 100 frequency share).

To quickly summarize the results so far: it turns out that the diversity of the vocabulary as indicated by all of the above-mentioned measures was only restricted for about one month, especially between mid-March and mid-April. All measures peaked at the beginning of April and then returned to a "pre-pandemic" level until mid-June. In retrospect, we have to concede that we overestimated the length of the effect. However, a more detailed conclusive analysis remains to be done, also taking specific areas of vocabulary into account (for example medical and epidemiological terminology). *R* code for the *cOWIDplus Analysis* is included in the website itself and can be accessed via the "Code" buttons above each results chunk.

4. *cOWIDplus Viewer*

Providing the daily frequency lists (see previous section) is a first step to enabling other researchers to use the RSS feed data. However, the amounts of knowledge and resources that are necessary to deal with the data could pose an obstacle for many people who might also be interested in the data. Inspired by the popular *Google Books Ngram Viewer* (Michel et al., 2011), we created the *cOWIDplus Viewer* (https://www.owid.de/plus/cowidplusviewer2020), which enables all users to search the RSS corpus for specific words forms, strings within word forms or bigrams. We explain the architecture and interface of the *Viewer* in Section 4.1, followed by examples of the interface in use (4.2).

4.1 Architecture and interface

The *Viewer* is a web application built using the *Shiny* framework (Chang et al., 2020). *Shiny* is an *R* package that allows to build interactive web apps from R. It comes with pre-defined HTML widgets that can be used to build an interactive browser-based user interface. On the server side, user input is processed within *R* and the results are returned to the user's browser. *R* code is accessible via *GitHub* (https://github.com/saschawo/cOWIDplusViewer).

The interface of the *cOWIDplus Viewer* is divided into two pages accessible via tabs just below the title: the main page and the bigram finder page. Figure 2 shows a screenshot of the current English version of the *cOWIDplus Viewer's* main page.

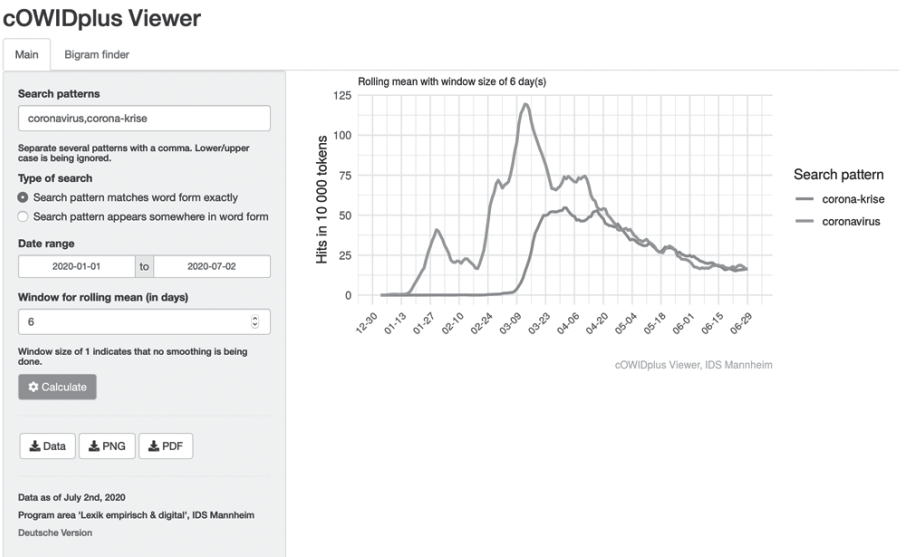

Figure 2. English version of the *cOWIDplus Viewer* main page with the default search patterns

All search parameters are located in the input panel on the left. Directly at the top is the search pattern field where different patterns have to be separated by a comma. After experimenting with the possibility of entering regular expressions in the search field, we decided against this option because researchers who want to use regular expressions could still use the downloadable dataset (see Section 3) and we wanted to keep the entry threshold as low as possible. (Not allowing regular expressions is also preferable due to security and, for some regular expressions, performance reasons.) All special characters used in regular expressions

are deleted from users' input and white space around entries is trimmed before further processing. Instead of full-blown regular expressions, we opted for a simple switch where users can choose the type of search for unigram searches (bigrams are always searched exactly as entered by the user). The two options are exact matching (the internal search algorithm uses the regular expression ^<pattern>$ for this) and matching within other word forms. Furthermore, the user can choose a specific date range which is propagated to the x-axis of the frequency graph on the right, the table of absolute frequencies on the bottom and the downloadable data file. The last parameter sets the window size of the rolling mean calculation which is realized by the "frollmean" function (with a center-aligned window) in the *R* package *data.table* (Dowle & Srinivason, 2019) and can be chosen between 1 (no smoothing) and 14 days. Due to the potentially lengthy calculations, we chose to use an action button ("Calculate") to trigger the calculation. Three download buttons give the user the opportunity to download the data which is currently displayed (as PNG or PDF images or in tabulated CSV format). Each week (and possibly, in the future, every two weeks), the underlying corpus is updated, so we have to make sure that the user is aware of the current date of the data: this is indicated at the bottom of the input panel.

On the right side of the main page, we show a graph of the relative frequencies in time which is smoothed if a window of greater than 1 day is selected for the rolling mean. If the user selects the search type "within-word-form", a table is displayed below the plot. This table shows all hits for all patterns, sorted by absolute frequencies within the selected date range. This table is quite important for this type of search because it might well be the case that the word forms that actually appear most often do not match the search patterns exactly. For example, when searching for *maske* ("mask") the most frequent hit is currently *maskenpflicht* ("obligation to wear a mask"), *masken* (inflected form of *maske*) is second, *schutzmasken* ("protective masks") takes rank 3, and *masked* (part of the name of the TV show "The Masked Singer") rank 4. The entry *maske* itself is currently on the fifth rank in terms of absolute frequencies with 740 hits. The table can be searched, sorted and the users can flip pages (not visible in Figure 2).

Figure 3 shows the second page of the *cOWIDplus Viewer* (the "Bigram finder") with the default search parameters. We implemented the bigram finder with a potential user in mind who might not have a very clear-cut research question regarding a specific bigram. Here, users can search the corpus to discover bigrams that might be worth checking out on the main page. There are three types of bigram searches: the first (and default) one searches for the pattern somewhere in a bigram, partial matches are allowed. The second and third types are more restrictive: here, one can search for bigrams where the search pattern is the first (second option) or second part (third option) of the bigram. Consequently, the

default search pattern *corona* returns bigrams like *das coronavirus* for the first option but only patterns like *corona und* ("corona and") for the second option or *wegen corona* ("because of corona") for the third option.

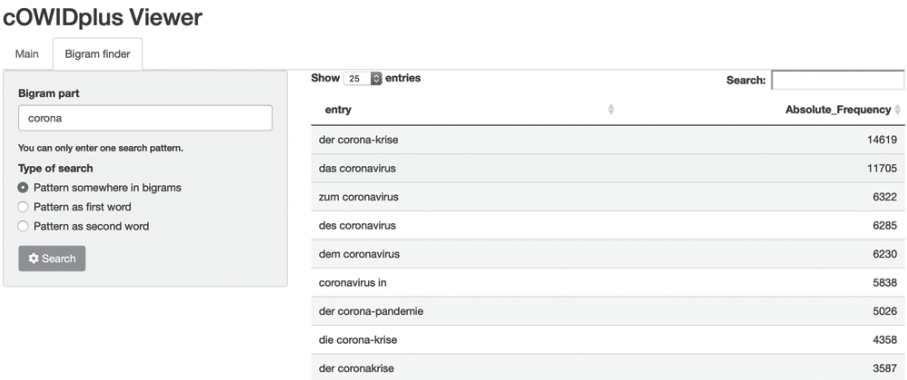

Figure 3. The bigram finder of the *cOWIDplus Viewer* (second page of the application)

4.2 Examples

At the moment, most of the effects people might be interested in are those related to the coronavirus pandemic, as already introduced in Section 1. Here, we show query examples where effects of the pandemic on social life are quite evident. One such pair of queries extracted with a method outlined in Koplenig (2017) is *fc* (abbreviation for *Fußballclub*, "football club") and *corona*. These two word forms show almost perfect opposite frequency time courses. As soon as all games in the Bundesliga (Germany's men's football competition) were cancelled (the last fixture in the first division was played on March 11th, 2020), the relative frequency of *fc* starts to drop dramatically with *corona* gaining substantially. With the Bundesliga resuming on May 16th, the two patterns converged in relative frequencies again.

Another example gives an impression of both a linguistically interesting case of naming conventions combined with a glimpse of political discussions. Figure 5 shows the developments in relative frequencies for the (partial) matches of *ausgangssperre* ("curfew"), *kontaktverbot* ("contact ban"), and *lockerung* ("relaxation"). The first mention of a potential curfew captured in the RSS corpus was on March 12th. Afterwards, its frequency peaked on March 15th. However, a curfew was never implemented nation-wide in Germany. Instead, on March 22nd, it was decided to implement a contact ban (which could be described as a much lighter version of a curfew).

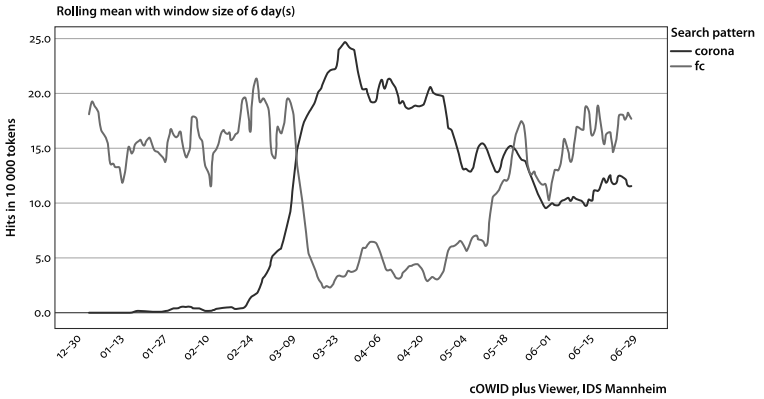

Figure 4. *cOWIDplus Viewer* result graph for exact matches of the word forms *fc* and *corona*

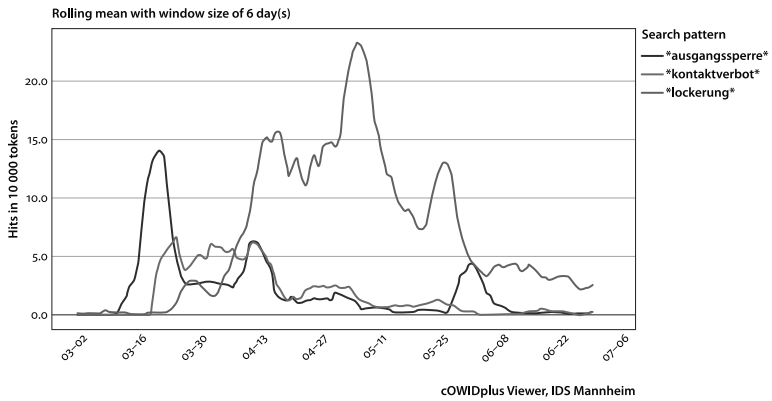

Figure 5. *cOWIDplus Viewer* result graph for matches of *ausgangssperre* ("curfew"), *kontaktverbot* ("contact ban"), and *lockerung* ("relaxation"), and longer forms found via the "within-word-form" option

Although *kontaktverbot* never reached a peak as high as *ausgangssperre*, it was consistently more frequent than *ausgangssperre* for a prolonged time. Then, around mid-April, discussions about potential relaxations began which is reflected by a sharp increase of frequencies of *lockerung*. With certain relaxations becoming reality at the beginning of May, the frequencies increased even more.

The last example shows how unigram and bigram searches can be mixed. In Figure 6, relative frequencies for *alltag* ("every-day life") and *neue normalität* ("new normality") are shown. Although *alltag* remains more frequent throughout the whole time period, the German-speaking online-press begins to include references to a certain *neue normalität* into their RSS feeds around the middle of April. It is quite reasonable to assume that it needed some time after the pandemic

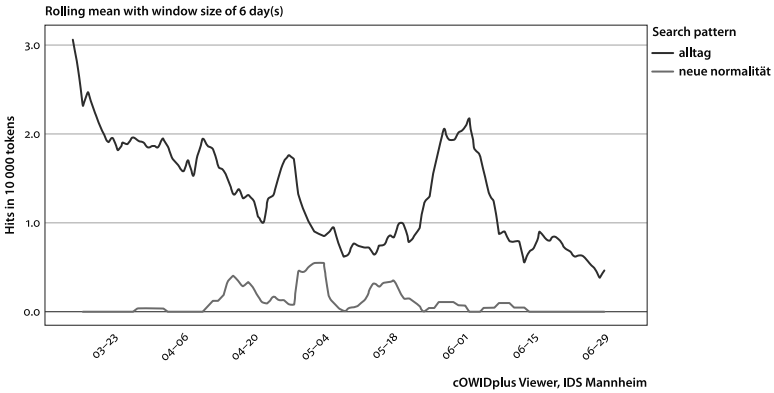

Figure 6. *cOWIDplus Viewer* result graph for exact matches of the unigram *alltag* ("everyday life") and the bigram *neue normalität* ("new normality") since March 15th

arrived in Germany that such a "new normality" is being established and written about.

At the moment, it is not possible for users to examine a word's usage in context. We are currently evaluating possibilities to share this data and implementing the functionality without violating copyright restrictions of the RSS data.

5. Conclusions

The *Viewer* is already being used in research on German neologisms related to the pandemic. On https://www1.ids-mannheim.de/sprache-in-der-coronakrise [accessed on July 2nd, 2020], several short papers by our colleagues, e.g. on terminology from medicine, conspiracy theories, and relaxations of measures, featuring frequency plots from the *cOWIDplus Viewer* are available. More generally, the RSS corpus and the *Viewer* application can serve as a starting point for corpus-based lexicography interested in (close to) real-time frequency developments in press language. Of course, not all new words (or meanings) observable in the RSS corpus will end up being described in a dictionary. But it can help to find candidates that have to be observed until lexicographers can decide whether or not to describe them.

As time passes and the impact of the coronavirus crisis on language presumably becomes weaker, we believe that the *Viewer* will still prove to be a valuable tool to explore a historical record of German press language during a global crisis. Also, it enables the research community to compare the effects of such a (hopefully) singular event like this global pandemic to other events with large-scale implications (e.g. the US presidential elections later in 2020).

The *cOWIDplus* resources can be expanded in several ways to make them more useful for linguistic analyses: one obvious expansion of the scope would be to include more RSS sources in the corpus, also from other languages. One restriction of the *Viewer* application is that word forms cannot be investigated in context at the moment (e.g. via KWIC views). Another possible expansion concerns the discovery of interesting word forms. At the moment, the users have to come up with their own ideas about which words might be interesting. Following the method proposed by Koplenig (2017), the tool could automatically provide word forms that changed the most in terms of their corpus frequencies. It remains to be seen whether it is feasible to include such a feature in the current *Viewer* application or whether it is more practicable to develop another application to provide such a functionality.

References

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). *shiny: Web application framework for R* (Version 1.4.0.2) [Computer software]. https://CRAN.R-project.org/package=shiny

Davies, M. (2016–). Corpus of news on the web (NOW): 10 billion words from 20 countries, updated every day. https://www.english-corpora.org/now/

Dowle, M., & Srinivason, A. (2019). *data.table: Extension of "data.frame"* (Version 1.12.8) [Computer software]. https://CRAN.R-project.org/package=data.table

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with {lubridate}. *Journal of Statistical Software*, *40*(3), 1–25. https://doi.org/10.18637/jss.v040.i03

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs: General and Applied*, *56*(2), 1–15. https://doi.org/10.1037/h0093508

Koplenig, A. (2017). A data-driven method to identify (correlated) changes in chronological corpora. *Journal of Quantitative Linguistics*, *24*(4), 289–318. https://doi.org/10.1080/09296174.2017.1311447

Michel, J.-B., Shen, Y. K., Aiden, A. P., Verses, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, L. E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(14), 176–182. https://doi.org/10.1126/science.1199644

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.2). R Foundation for Statistical Computing [Computer software]. https://www.R-project.org/

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Temple Lang, D. (2020). *XML: Tools for parsing and generating XML within R and S-Plus* (Version 3.99-0.3) [Computer software]. https://CRAN.R-project.org/package=XML

Xie, Y., Allaire, J., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC. https://bookdown.org/yihui/rmarkdown. https://doi.org/10.1201/9781138359444

Address for correspondence

Sascha Wolfer
Department for Lexical Studies
Leibniz Institute for the German Language
R5, 6-13
68161 Mannheim
Germany
wolfer@ids-mannheim.de

Co-author information

Alexander Koplenig
Department for Lexical Studies
Leibniz Institute for the German Language
koplenig@ids-mannheim.de

Carolin Müller-Spitzer
Department for Lexical Studies
Leibniz Institute for the German Language
mueller-spitzer@ids-mannheim.de

Frank Michaelis
Department for Lexical Studies
Leibniz Institute for the German Language
michaelis@ids-mannheim.de