Contagious "Corona" Compounding by Journalists in a CLARIN Newspaper Monitor Corpus

Koenraad De Smedt
University of Bergen, Norway
desmedt@uib.no

Abstract

Newspaper monitor corpora, which incorporate new materials on a regular basis, are particularly useful for tracking linguistic changes spurred by current developments. The COVID-19 pandemic prompted a case study in the Norwegian Newspaper Corpus. The corpus was mined for productive compounds with the stems "corona" and its alternative spelling "korona", tracing their frequencies and dates of first occurrence during the first wave of the pandemic. The quantitative analysis not only monitored the daily volume and variation of such compounds, but also the dynamics of vocabulary growth, and a change in their preferred spelling. The paper concludes with reflections on methodology and data sources.

1 Introduction

The COVID-19 pandemic, which started to spread around the world in the spring of 2020, has quickly become the subject of much research, not just in medicine, but also in the social sciences and humanities. For various research purposes, corpora containing specific types of discourse have been compiled, from scientific articles (Lu Wang et al., 2020) to tweets (Dimitrov et al., 2020). Furthermore, large *monitor* corpora, which are regularly updated from a wide range of sources, are allowing lexicographers and others to detect linguistic changes in almost real time (OED Editorial, 2020; Paton, 2020).

Newspaper corpora are essentially time-stamped journalistic descriptions of daily events. Newspaper *monitor* corpora are moreover regularly updated; thereby they are not only a window into the course of current events, but they also provide up-to-date data samples of journalistic language. Such corpora are unfortunately scarce. In fact, the only monitor corpus that I could identify in the CLARIN resource family overview of newspaper corpora[1] is the Norwegian Newspaper Corpus (Andersen and Hofland, 2012) at the CLARINO Bergen Centre (De Smedt et al., 2016). This large resource, containing two billion words and growing, has been useful in earlier studies of neologisms, loan words and other vocabulary expansion (Andersen, 2012).

The pandemic provided an exceptional opportunity to further demonstrate the use of this monitor newspaper corpus. It is a rare experience to observe a sudden dramatic increase in the vocabulary in a very short period of time. Events related to the outbreak and pandemic were extensively discussed in the media all over the world. This stimulated the coining of new words in many languages. Particularly striking in Norwegian was the productivity of compounds, such as *testkø* ("testing queue"), *hjemmeisolering* ("home isolation"), *kommunekarantene* ("municipal quarantaine"), *smittesporingsapp* ("contagion tracing app") and *flokkimmunitetsstrategien* ("the herd immunity strategy"). Like other Germanic languages, Norwegian has indeed very productive compounding, and compounds are normally written as one word.

In this context, *corona/korona* stands out. From January 2020, the word by itself quickly became a common term for both the virus, the disease and the epidemic. It also became by far the most frequent initial part of compounds. What is unique about *corona/korona*, moreover, is that practically all its compounds

[1] https://www.clarin.eu/resource-families/newspaper-corpora

Koenraad De Smedt 2021. Contagious "Corona" Compounding by Journalists in a CLARIN Newspaper Monitor Corpus. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 83–92.

were completely new. Before 2020, *corona* occurred in only a handful of relevant compounds, such as *coronavirus*, (”corona virus”), *coronafamilien* (“the corona family”), and *corona-vaksiner* (“corona vaccines”).[2] During the first wave of the pandemic, there was an explosion of new compounds, such as *koronatelefon* (“corona telephone”), *koronadødsfall* (“corona death”), *koronafrykten* (“the fear of corona”), *corona-cruiset* (“the corona cruise”) and *coronatider* (“corona times”). In contrast to *virus*, the term *corona/korona* is more specific and eye-catching, something that appeals to newspaper editors. This may explain why its use seemed contagious in the journalistic sphere.

The current work is a case study showing the possibilities of mining a newspaper monitor corpus accessible through CLARIN. Its primary objective is to trace the productivity of compounds with *corona* and its alternative spelling *korona* during the first wave of the pandemic. The hypothesis was not only that an evolution in the tempo of vocabulary expansion had occurred, but also an evolution in the ratio between types and tokens, so the goal was to measure the extent and speed of these trends. Another objective was to trace spelling change in terms of changing proportions of the two variant spellings. There were indications that the normalization by the Language Council near the end of January 2020 had influenced journalists’ spelling, but the extent of the change had not been quantified previously.

2 Data and Method

The Norwegian Newspaper Corpus (Andersen and Hofland, 2012) was the data source for the present study. It is updated every night by harvesting publicly accessible articles from ten major Norwegian online newspapers.[3] At every automatic update, boilerplate is removed so that nearly clean text is left, and every article is tagged with the date and the source. This corpus is accessible in two ways.

One way to use the corpus is through an instance of the IMS Corpus Workbench (CWB; Evert and Hardie, 2011).[4] In this system, the corpus is split up in different sections, most or them covering one year. Searches can be specified by regular expressions and can be limited to a year, a month or a date. Some disadvantages of the CWB version are that search can only be performed in one section at a time, and that it is not possible to specify arbitrary start and end dates. Another disadvantage is that the system does not have a download function, so that relevant items must be extracted from the HTML encoding of the search result pages.

The corpus is also accessible through the Corpuscle corpus management and search system[5] (Meurer, 2012) at the CLARINO Bergen Center. This system has a better interface and a more powerful and efficient query system (Meurer, 2020). It allows the specification of arbitrary start and end dates in queries. It also offers download of matching strings, with optional annotation features, to a file with tab-separated values. Unfortunately, this version of the corpus is updated less regularly than the CWB version. Both versions were consulted, but the data from Corpuscle, which was up to date until March 8, 2021, are the basis for the present study.

The query `"[ck]orona.*" %c :: year = "202[0|1]"` was used in Corpuscle to retrieve all occurrences of words starting with *corona* or *korona*, in uppercase or lowercase, from the Bokmål[6] section of the corpus, tagged with the year 2020 or 2021. All matches were downloaded as a tab-separated file with keywords, newspaper codes and dates. The first observation was on January 9, 2020, and the last one on March 8, 2021. The period with observations thus spans a year and two months, or 425 days to be precise.

The base forms *corona/korona* and their inflected forms were removed, as well as obvious spelling errors and unrelated words such as *koronar* og *coronal*. The cleaned word list has 167957 tokens, which are all compounds, with or without hyphens. Preprocessing, analysis and plotting was performed with a shell script that called programs in Awk, Python and R.

[2] Before 2020 these referred to viruses other than SARS-CoV-2, primarily SARS-CoV and MERS-CoV.

[3] The following newspapers, with their codes, are represented in the corpus: Adresseavisen (AA, Trondheim), Aftenposten (AP, Oslo), Bergens Tidende (BT, Bergen), Dagsavisen (DA, Oslo), Dagbladet (DB, Oslo), Dagens Næringsliv (DN, Oslo), Fædrelandsvennen (FV, Kristiansand), Nordlys (NL, Tromsø), Stavanger Aftenblad (SA, Stavanger) and Verdens Gang (VG, Oslo).

[4] http://korpus.uib.no/avis/bokm.html

[5] http://clarino.uib.no/korpuskel

[6] The corpus also has a separate Nynorsk section, which is much smaller and was not used in this study.

3 Analysis

3.1 Spelling

From October 1998 until the end of 2019, the few existing compounds with *corona*, in senses related to the virus, occur in the Norwegian Newspaper Corpus only with initial *c*-, whereas *koronautbrudd* ("corona outbreak") with *k*- was only used in the sense of "solar flare." An n-gram search of both spellings in the digital newspaper collection of the National Library of Norway, which goes further back in time, confirms this practice, as shown in Figure 1. In January 2020, the spelling with *c*- was still very dominant in the Norwegian Newspaper Corpus.

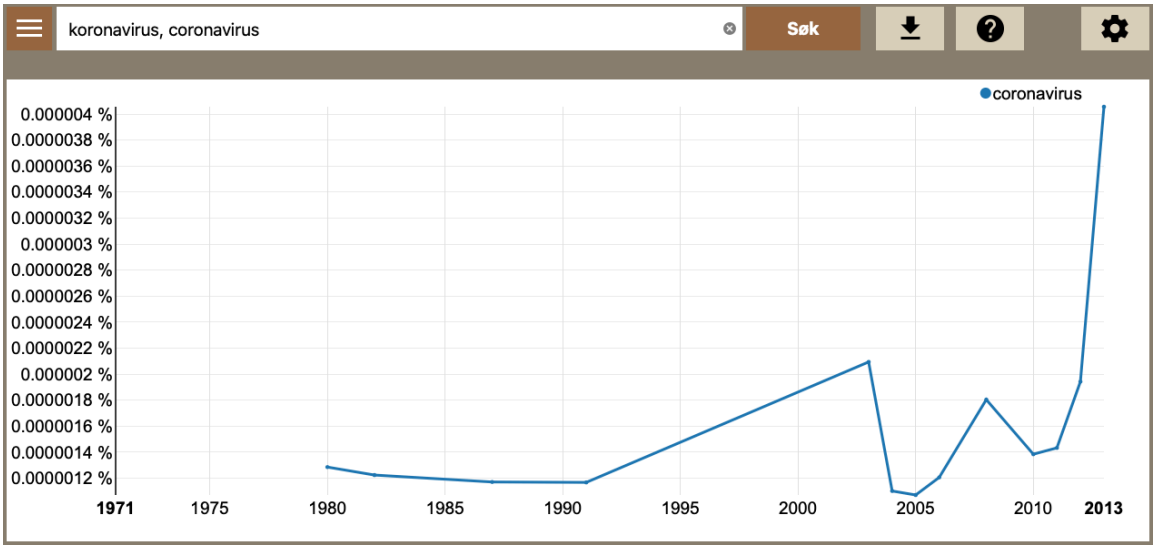

Figure 1. Relative frequencies in the newspaper collection of the National Library of Norway: *coronavirus* occurs from 1980 until (at least) 2013, *koronavirus* does not occur in this period.

In an online article on January 28,[7] however, the Language Council of Norway stated that the word is to be spelled with *k*-, thereby effectively normalizing the spelling for the first time and doing so in a way that went against the commonly practiced spelling. The present study is probably the first quantitative assessment of the effect of that normalization. After a brief period of fluctuation between spellings, the use of *k*- in a majority of cases was observed after the middle of February, as shown in Figure 2. However, after more than a year since the spelling change, there is no further convergence towards the new spelling. This seems due to the fact that not all newspapers adopted the newly normalized spelling. Figure 3 shows the variation per newspaper, revealing some clear discrepancies between newspapers as regards the choice between *c*- and *k*-. A final note on spelling is that among the 167957 tokens there were 18168 written with a hyphen, which is normally unnecessary, except to avoid the collision of two *a*s, such as in *korona-app*, or in combinations with a number, such as *korona-17.mai* ("corona 17th of May", Constitution Day in Norway).

3.2 Frequency, Variation and Productivity

The number of tokens per day is shown in Figure 4. The earliest occurrences of relevant compounds in the Norwegian Newspaper Corpus in 2020 were *coronavirus* (indef. sg.) and *coronaviruset* (def. sg.), on January 9, 2020. The use of these and other compounds remained modest for over a month, but on February 26, 2020, when the virus was detected in Norway, a marked increase can be seen. The maximum token count was 1765 on a single day.

The token counts are generally somewhat lower in the weekends when the volume of articles is lower. In that respect it might have been useful to count normalized frequencies on the basis of the volume of

[7] https://www.sprakradet.no/Vi-og-vart/hva-skjer/Aktuelt-ord/koronavirus/

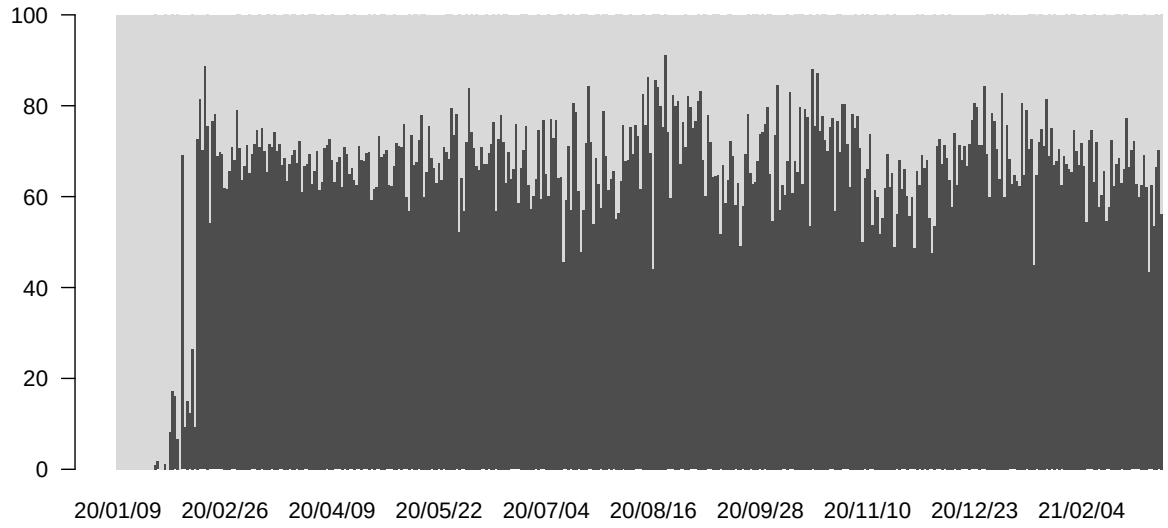

Figure 2. Distribution of *c*- (light) and *k*- (dark) over time. No bars are shown for days without any occurrences of either.

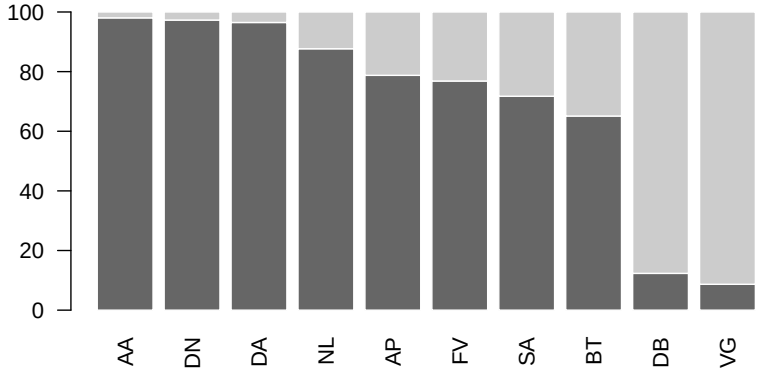

Figure 3. Distribution of *c*- (light) and *k*- (dark) per source; AA=Adresseavisen, AP=Aftenposten, BT=Bergens Tidende, DA=Dagsavisen, DB=Dagbladet, DN=Dagens Næringsliv, FV=Fædrelandsvennen, NL=Nordlys, SA=Stavanger Aftenblad and VG=Verdens Gang.

harvested words per day, but unfortunately the daily volumes are not provided by the corpus interface. Token counts in themselves are however not the primary focus of the present investigation.

Whereas the volume of tokens indicates how much is written about a topic in general, the breadth of the discussion in terms of subtopics may rather be revealed by looking at the number of distinct types. For this purpose, normalization was applied by deleting the first part of the compound so that the above-mentioned spelling variation and the possible use of a hyphen are disregarded. The remaining wordforms were lemmatized.[8] Lemmatization resulted in a few errors, most of which were corrected with a manually constructed script. Furthermore, lemmatization was not entirely consistent, e.g., some deverbal adjectives were reduced to the verb lemma, whereas others were not. Also, some ambiguities may not have been correctly resolved, because the lemmatizer was run on a simple list of wordforms, which does not provide any helpful context, as compared to applying the lemmatizer to running text. A better solution would obviously be to lemmatize the whole corpus, but that was not a realistic option at the time of this study. Nevertheless, the lemmatizer output was in general useful and its minor imperfections do not seem to have distorted the general picture.

[8]Lemmatization was done by means of the model *nb_ core_ news_ md-2.3.0* in *spaCy*, https://github.com/explosion/spacy-models/releases//tag/nb_core_news_md-2.3.0, details at https://spacy.io/models/nb#nb_core_news_md.

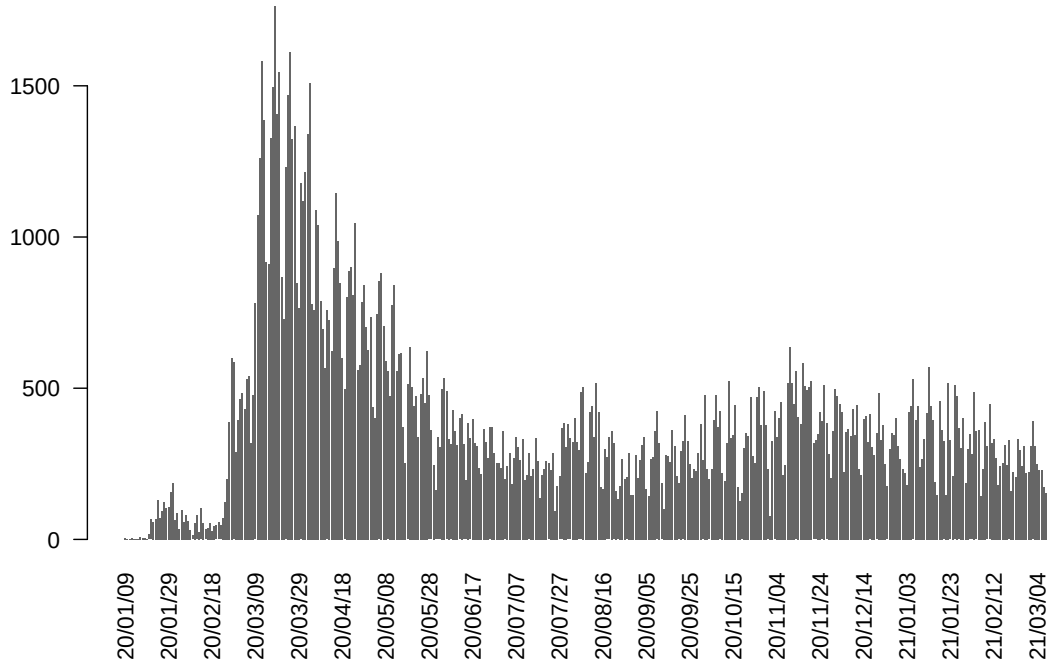

Figure 4. Number of occurrences observed per day.

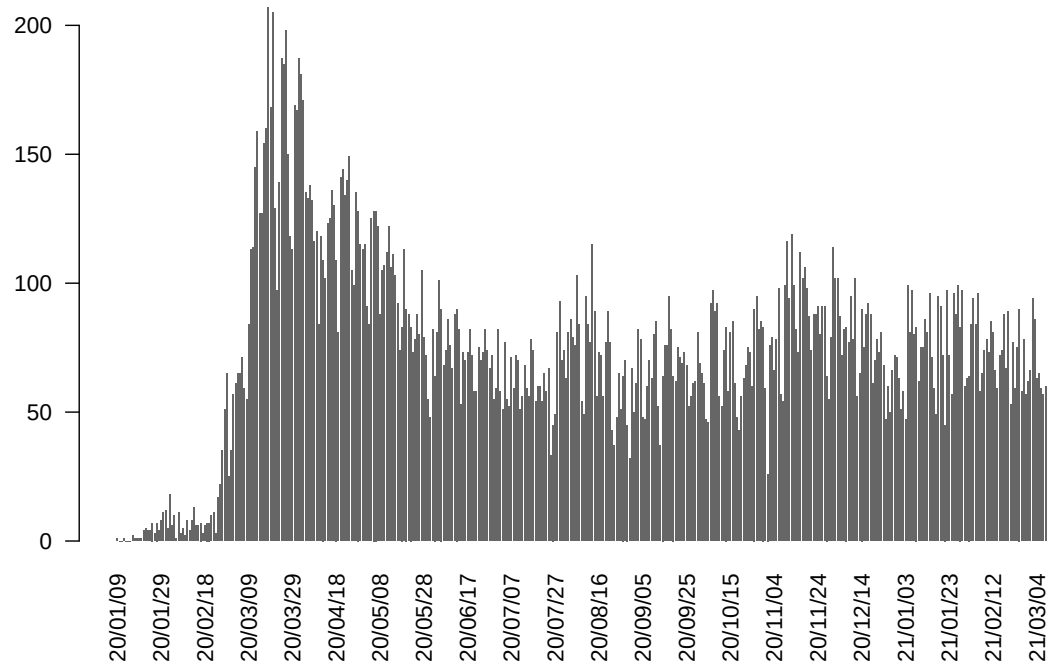

Figure 5. Number of types (lemmatized) observed per day.

Normalization and lemmatization meant, for instance, that the word forms *koronatiltakene* ("the corona measures") and *Corona-tiltak* ("corona measure(s)") were both reduced to the same lemma *tiltak*. Alternative spellings of lemmas such as *oppmyking* and *oppmykning* ("softening") remain however separate items. In the end, the original 167957 tokens, consisting of 3012 distinct word forms, were reduced to 2133 lemma types. A frequency list was made of all the types, showing a typical Zipf distribution.[9] Compounds containing *virus* make up close to half of the total number of tokens.

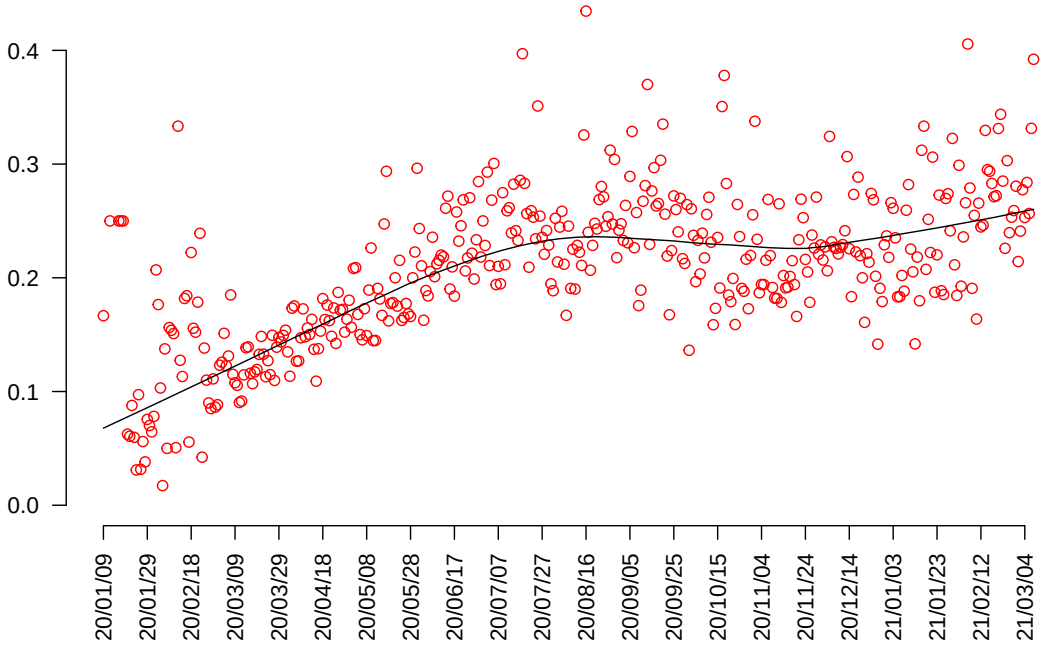

Figure 6. Variation (types / tokens) per day.

The type count per day, as shown in Figure 5, at first sight seems to roughly follow the increase in the token count. However, as Figure 6 shows with a trend line (fitted with local polynomial regression), the ratio of types to tokens per day is not constant, but increasing, with some flattening from the late summer of 2020. Initially the ratio was around 0.1, which means that every word was used on average ten times per day. Near the end of the studied period, the ratio had risen to around 0.25, which means that every word was used on average only about four times per day. This evolution suggests that the variation in subtopics in the discourse not only increased markedly during the initial few months of the pandemic, but also remained high until the end of the observation period.

Another measurement is the number of *new* words per day, i.e. types which had not been recorded on earlier dates during this period (and not even before 2020, for practically all the words). A list was made of all types, with their first date of occurrence and the first newspaper in which they were observed. Cumulative counts of these new types, per day, shown in Figure 7, show the speed of the vocabulary growth. In January and February 2020, the number of new compounds increased very slowly, but a sharp acceleration can be observed around February 26, 2020, when the virus had reached Norway. This steep increase continues throughout March 2020, before it flattens out slightly in April and a bit more in May, but after May, the vocabulary growth remains remarkably strong and linear until the end of the period with observations. This can be seen as an indication that the discourse needed more and more descriptive words as the effects of the pandemic continued to affect more and more aspects of our society.

As expected, most compounds were nouns, e.g. *koronapsyken* ("the corona psyche"), some were verbs, e.g. *koronastenge* ("close down due to corona"), some were adjectivally used participles, e.g. *corona-stanset* ("stopped by corona") and some were adjectives, notably including *koronafast* ("stuck due to

[9] A frequency list and a list of items by date of first occurrence can be accessed at https://github.com/clarino/corona. In these lists, types are reduced to their final parts.

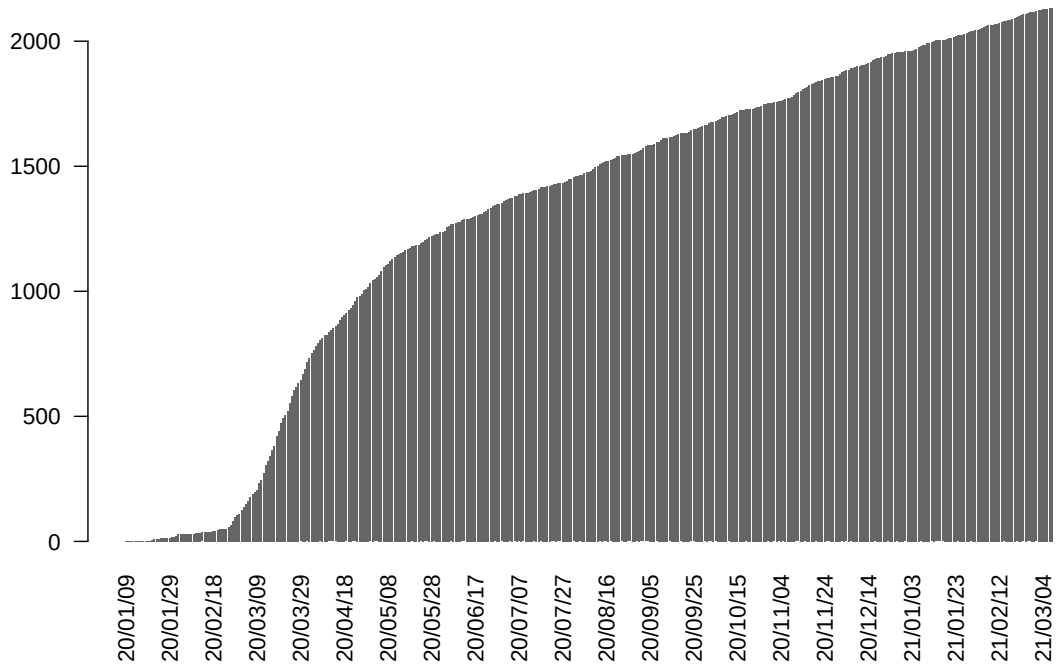

Figure 7. Cumulative increase of the *corona* compound vocabulary.

corona"), the latter modeled after the existing *værfast* ("stuck due to bad weather") and the relative newcomer *askefast* ("stuck due to the ash cloud") in 2010 (De Smedt, 2012).

Most of the new compounds that appeared in the current study are not quite transparent. Indeed, before 2020 it would have been difficult to interpret, for instance, *korona-telt* ("corona tent"), *koronautsettelsene* ("the corona postponements"), *coronalov* ("the corona law") and *corona-kompensasjon* ("corona compensation"). Taken as a whole, the semantic contribution of *corona/korona* in such compounds is a broad contextualization of the meaning and can be paraphrased as "related to the virus, the disease, the epidemic, or the measures to combat all of these." Several of the compounds are metaphorical and have emotional connotations, such as the final parts *knekken* ("the breakdown"), *knipen* ("the pinch"), *spøkelset* ("the ghost"), *tsunamien* ("the tsunami") and *tabu* ("taboo").

4 Discussion

This paper presents a timely use case demonstrating the potential of a newspaper monitor corpus, i.e. a corpus which is regularly updated with fresh newspaper articles, for the purpose of tracing changes to the language in almost real time. In particular, it tracks and analyzes new compounds with *corona/korona* in the Norwegian Newspaper Corpus. An earlier study (in Norwegian) with a similar objective and method covered a period of only 139 days, ending on May 26, 2020 (De Smedt, 2020). In contrast, the present study represents a considerable extension in time, as it covers 425 days, i.e. a period more than three times as long, stretching until March 8, 2021, when the current data were collected.

It was found that the majority spelling had changed in the course of about one month, although further convergence on the new standard, which could have been expected, is not borne out by the data. The huge number of occurrences (tokens) of compounds with *corona/korona* in the studied period may not be entirely surprising, but the dynamics of vocabulary growth and diversity are noteworthy. The sharp acceleration in the creation of compounds from about February 26, 2020 started to slow in April and May 2020; if that slowing had continued, the vocabulary might have flattened out at around 1500 words. Instead, the vocabulary growth continued in a surprisingly linear climb, adding at least 600 more words from June 2020. Another noteworthy result was that the variation of compounds in use, measured as the ratio of types vs. tokens per day, not only increased in the initial phase, as already observed by De Smedt (2020), but remained high during the entire period. These dynamics are certainly driven by the continuing

need to report on a widening range of situations and events that were consequences of the pandemic, but perhaps just as much by journalists' willingness to continue exploiting the salience of a new word and bombard their readership with attention-drawing compounds at an average pace of about five new ones per day.

The present findings show similarities and differences with a previous study on compounds with *aske* ("ash") following the volcanic eruption in Iceland in 2010 (De Smedt, 2012). That study found a sharp increase in variation but it was less broad in scope and it flattened out after half a month. In comparison, the presently reported increase in variation did not rise as quickly, but accelerated after about a month. Also, the creation of new compounds did not decrease quickly, but was sustained until the end of the studied period. A possible explanation for these differences might be that the effects of the current pandemic are not only lasting longer, but are also having more widespread and long-lasting effects on society. Nevertheless, both in the earlier study of *aske* and the current study of *corona/korona*, creative compounding seems to be contagious among journalists who appear intent on outdoing each other with ever more creative neology.

There are other systems for tracking news or tracing new words. Among news trackers, the European Media Monitor[10] has a long-standing reputation. However, its interface is oriented towards topic tracking and alerts rather than detecting neologisms. The use of the Norwegian newspaper archive Atekst Retriever[11], based on daily harvesting from even more media sources than the Norwegian Newspaper Corpus, was briefly considered. However, Retriever is less suitable for linguistic research, as it returns webpages with summaries, from which it is more difficult and less reliable to extract keywords and relevant information (i.e. at least the date and source). Furthermore, the earliest mention of a relevant compound with *corona/korona* in Retriever was December 28, 2019, which turned out to be a mistake in dating: in reality it was an article from February 28, 2020. It must be added that also the Norwegian Newspaper Archive had some problems with dates, so that two occurrences dated January 2, 2020 were considered unreliable and were removed by the script. More generally, such issues show that although correct dating is paramount in studies like these, there is always a risk that errors go under the radar; for a related discussion of hidden dangers in digitized text, see Nunberg (2009).

For the detection of neologisms as such, several systems are useful in their own right, such as The Word Spy[12] for English and Die Wortwarte[13] for German, the latter developed in the context of CLARIN-D and using monitor corpora. However, these sites seem to offer neither regular expression search, nor output as a complete list of observations tagged with sources and dates. For those reasons, they are less suitable for the kind of data aggregation and analysis of compounds on a time line as presented here. In fact, the Norwegian Newspaper Corpus also features a separate automated system which every day identifies new words that have been added to the corpus. However, the goal of the present study is not so much to spot new words, but rather to trace both the creation and the protracted use of compounds based on a specific stem through a given period.

The dynamics exposed in this study may, on the one hand, serve to illustrate collective journalistic practice and contagious tendencies towards salient and eye-catching terms. On the other hand, the dynamics may also provide clues as to how fast, how broadly and for how long salient events are affecting our society, whether an event may initially have been underestimated, and so on. This kind of information may in turn be used in applications, such as the automatic detection of significant "bursts" of words in information streams (Kleinberg, 2002, e.g.), which, in cases like the current one, would benefit from compound analysis.

Further studies, including investigations of *corona* creativity across languages, might be interesting. However, despite the advantages of the availability of many newspaper corpora through CLARIN, the above-mentioned CLARIN resource family overview of newspaper corpora shows a lack of up-to-date monitor corpora. Almost all of the newspaper corpora in the CLARIN list consist of fairly dated materials and their different periods do not always overlap. Furthermore, the corpora are not easily interoperable;

[10] https://emm.newsbrief.eu
[11] https://web.retriever-info.com/services/archive, accessed March 17, 2020
[12] https://www.wordspy.com
[13] https://wortwarte.de

they do not share the same annotation and formatting, and they are not searchable through the same interface. Addressing these issues might call for a kind of multilingual Federated Content Search on this CLARIN resource family, and by the compilation of more monitor corpora that allow the study of vocabulary linked to current events.

A final remark is that the Norwegian Newspaper Corpus may not be sustainable in its current form. Current agreements with the newspaper publishers allow scraping of the public newspaper websites, but more and more articles are being hidden behind paywalls. Over the years, the annual corpus accrual of 100 million words for the ten major newspapers has sunk to about 50. Some newspapers that provided materials in an early phase of corpus collection have no open articles at all now. Furthermore, changing webpage formats affect the quality of materials obtained by scraping. Clearly, it would be better to obtain complete newspapers, appropriately encoded and licensed, directly from the publisher. This requires new agreements, which are currently being discussed with Norwegian media corporations, through cooperation with The Norwegian Language Bank at the National Library of Norway, with the aim of building a new Norwegian Media Corpus.

Acknowledgements

Thanks to Knut Hofland for implementing and maintaining the Norwegian Newspaper Corpus and to Paul Meurer for importing the corpus in Corpuscle. Thanks to Mikkel Ekeland Paulsen, Carina Nilstun, Margunn Rauset, Victoria Rosén, Sturla Berg-Olsen and anonymous reviewers for information and comments that were helpful in preparing this paper.

References

Andersen, G. 2012. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Studies in Corpus Linguistics 49. John Benjamins, Amsterdam/Philadelphia.

Andersen, G. and Hofland, K. 2012. Building a Large Corpus Based on Newspapers from the Web. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. Studies in Corpus Linguistics 49. John Benjamins, Amsterdam/Philadelphia, 1–28.

De Smedt, K. 2012. Ash Compound Frenzy: A Case Study in the Norwegian Newspaper Corpus. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. Studies in Corpus Linguistics 49. John Benjamins, Amsterdam/Philadelphia, 241–255.

De Smedt, K. 2020. Smittsomme Koronaord. *Oslo Studies in Language* 11(2):59–73.

De Smedt, K., Samdal, G. I. L., Kyrkjebø, R., Al Ruwehy, H. A. H., Gjesdal, Ø. L., Rosén, V., and Meurer, P. 2016. The CLARINO Bergen Centre: Development and Deployment. *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland*. Linköping Electronic Conference Proceedings. Linköping University Electronic Press, 1–12.

Dimitrov, D., Baran, E., Fafalios, P., Yu, R., Zsu, X., Zloch, M., and Dietze, S. 2020. TweetsCOV19 – A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic [Preprint]. 29th ACM International Conference on Information & Knowledge Management (CIKM2020), Resource Track. Association for Computing Machinery.

Evert, S. and Hardie, A. 2011. Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium. *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK.

Kleinberg, J. 2002. Bursty and Hierarchical Structure in Streams. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 91–101.

Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Xin Ru Wang, N., Wilhelm, C., Xie, B., Raymond, D., Weld, D. S., Etzioni, O., and Kohlmeier, S. 2020. CORD-19: The Covid-19 Open Research Dataset. arXiv: 2004.10706.

Meurer, P. 2012. Corpuscle – a New Corpus Management Platform for Annotated Corpora. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. Studies in Corpus Linguistics 49. John Benjamins, Amsterdam/Philadelphia, 31–49.

Meurer, P. 2020. Designing Efficient Algorithms for Querying Large Corpora. *Oslo Studies in Language* 11(2):283–302.

Nunberg, G. Aug. 31, 2009. Google's Book Search: A Disaster for Scholars. *The Chronicle of Higher Education*.

OED Editorial. Apr. 15, 2020. *Corpus Analysis of the Language of Covid-19*. OED blog. URL: https://public.oed.com/blog/corpus-analysis-of-the-language-of-covid-19/ (visited on 09/14/2020).

Paton, B. Apr. 9, 2020. *Social Change and Linguistic Change: The Language of Covid-19*. OED blog. URL: https://public.oed.com/blog/the-language-of-covid-19/ (visited on 09/14/2020).