Using Wiktionary revision history to uncover lexical innovations related to topical events: Application to Covid-19 neologisms

Franck Sajous

Franck Sajous

Using Wiktionary revision history to uncover lexical innovations related to topical events: Application to Covid-19 neologisms

1 Introduction

In April and July 2020, two extraordinary updates of the *Oxford English Dictionary* (OED) focused on the neologisms related to the Covid-19 pandemic. The responsiveness of the OED was made possible by the ability of its team to monitor, analyse and report quickly a sudden inflow of lexical changes. This ability, while not unique, is not prototypical in the lexicographic landscape. Corpus lexicography obviously requires corpora, but also tools to process and query them and sufficient person-hours. Fulfilling these standard requirements simultaneously, however, is no trivial task. The tools are not an issue as far as lexical creations are concerned. Building a headword list is indeed not considered a "hard part of lexicography" (Kilgarriff 1998) and detecting formal neologisms to update a nomenclature only requires "simple maths" (Kilgarriff 2009). Identifying semantic changes is more challenging. Clustering algorithms have been devised by Cook et al. (2013) while recent approaches use diachronic word embeddings (Fišer/Ljubešić 2018). These methods enable the detection of cultural shifts and linguistic drifts (Hamilton et al. 2016) but error rates are generally high. Another issue is that prediction-based models are appropriate for the detection of semantic changes over long time spans (decades or centuries) in very large corpora but they rarely perform well with shorter time units and smaller corpora (Kutuzov et al. 2018). On the corpus side, appropriate text collections to be used as input for the tools (i.e. diachronic corpora updated on a regular basis) are – sadly – not publicly available for most languages. Lastly, corpus lexicography also requires substantial manpower – ideally, trained lexicographers – to analyse vast amounts of data in a reasonable timeframe. Most institutions however, whether private or public, rarely have the manpower and the time they would like. The limitations are bound to the conditions of dictionary production rather than being intrinsic to corpus-based or corpus-driven approaches, as Landau (2001: 323) explained:

Acknowledgements: My thanks go to Basilio Calderone for checking the statistical analyses. The parsing of Wiktionary revision logs was performed using the OSIRIM platform, which is administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government and ERDF.

Franck Sajous, CLLE – CNRS & Université de Toulouse 2, Maison de la Recherche – 5, allées Antonio Machado – F – 31058 Toulouse Cedex 9, e-mail: franck.sajous@univ-tlse2.fr

https://doi.org/10.1515/9783110798081-014

> Dictionaries are not written in a vacuum, but by people working under the pressure of time. It sometimes seems to me that as technology has improved the speed and power with which we can examine the language, the pressures to produce quickly and with fewer staff have kept pace, so that on balance nothing is accomplished any faster or better. The expectations of management seem to rise at the same rate as the speed and power of the computer increase [. . .] Corpora can be used well or they can be used badly. Time pressures too often push the lexicographer to cut corners to avoid time-consuming analyses. It really doesn't do much good having a good corpus with marvelous analytical tools if they aren't used.

Time pressure and manpower are conversely not an issue in collaborative projects such as Wiktionary, which relies on massive online contributions performed by crowds of amateurs, not on corpus-driven analysis. Despite this questionable approach to lexicography and the resulting weaknesses described, inter alia, by Hanks (2012) and Rundell (2017), the exhaustiveness and the responsiveness of Wiktionary can be leveraged to detect lexical changes. Sajous et al. (2018) showed how swiftly the crowds are likely to detect formal and semantic neology. For example, in 2017, 73% of the entries added to the OED were already recorded in Wiktionary, whose median lead time was 4 years.

In the present contribution, I investigate if and how the English and French editions of the Wiktionary collaborative dictionary can be used as a corpus for real time neology watch. This option is envisaged as a stopgap, when no satisfactory corpus is available. Wiktionary can also prove useful in addition to standard corpus analysis, to minimize the risk of overlooking new coinages and new senses. Since the collaborative dictionary's quest for exhaustiveness makes the manual inspection of the new additions unreasonable (more than 31,000 English lemmas and 11,000 French lemmas entered the nomenclature in 2020), identifying the possibly relevant headwords is an issue. The solution proposed here is to use Wiktionary revision history to detect the (new or existing) entries that received the greatest number of modifications. The underlying hypothesis is that the most heavily edited pages can help identify the vocabulary related to "hot topics", assuming that, in 2020, the pandemic-related vocabulary ranks high. I used two measures introduced by Lih (2004), whose aim was to estimate the quality of Wikipedia articles: the so-called *rigour* (number of edits per page) and *diversity* (number of unique contributors per page). In the present study, I propose to adapt the rigour and diversity metrics to Wiktionary in order to identify the pages that generated a particular stir, rather than to estimate the quality of the articles. I do not subscribe to the idea that – in Wiktionary – more revisions necessarily produce quality articles (more revisions often produce *complete* articles). I therefore adopt Lih's notion of diversity to refer to the number of distinct contributors, but leave out the name *rigour* when it comes to the number of revisions. Wolfer and Müller-Spitzer (2016) used the two metrics to describe the dynamics of the German and English editions of Wiktionary. One of their findings was that the number of edits per page is correlated with corpus word frequencies. The variation in number of page edits should therefore reflect to some extent the variation of corpus word

frequencies. Renouf (2013) established a relationship between the fluctuation of word frequencies in a diachronic corpus and various neological processes. In particular, she illustrated how specific events generate sudden frequency spikes for words previously unseen in the corpus. For instance, *Eyjafjallajökull*, the – existing – name of an Icelandic glacier, appeared in the corpus when the underlying volcano erupted in 2010 and disrupted air traffic in Europe. In order to check if the same phenomenon occurs when using Wiktionary edits instead of corpus frequencies, I manually annotated the most frequently revised entries (according to various ranking scores) with the binary tag: "related to Covid-19" (yes/no). The annotations were then used to test the ability of various configurations to detect relevant headwords from the English and French Wiktionary, namely Covid-19 neologisms and related existing words that deserve updates.

2 Methodology

Scrutinising Wiktionary offers several opportunities for collecting Covid-related neologisms quite easily, depending on the language edition, and one's ability to automatically process the content of the dictionary. First of all, the *Coronavirus* category[1] of the English Wiktionary included 52 headwords on January 1st, 2021 and 124 in June. The English Wiktionary also has a category named *Hot words newer than a year*.[2] These words are described in Wiktionary as "presumably failing the criteria for inclusion on the *spanning less than a year* requirement", but are kept, according to Wiktionary, "because they have become widely used in that short time". Which is precisely the subject of the present study. In January 2021, the category included 94 words, 26 of which were not English. 79% of the English words (54 out of 68) were related to Covid-19. This observation is encouraging in that it suggests that the 2020 hot words are those related to the pandemic. Relying on the two categories mentioned is probably a good start, but by no means a satisfactory solution. First, some headwords that would deserve to be classified in these categories are not. Second, headwords that are related, but not specifically, to Covid-19, do not necessarily fit into these categories. Third, the goal of the present study was to develop a method for discovering topical neologisms that can be adapted to other language editions and other topics. In the French Wiktionary, there is no such thing as a "coronavirus" or a "hot words" category, and such categories will not make it possible to discover neologisms related to other "hot topics" in the future. Looking for some patterns (*covid*, *corona*, etc.) in the headword list, the definitions and the usage

1 https://en.wiktionary.org/wiki/Category:en:Coronavirus

2 https://en.wiktionary.org/wiki/Category:Hot_words_newer_than_a_year The page also contains a link *Hot words older than a year*, to which some of the 2020 hot words have been moved.

examples help to harvest relevant headwords (e.g. ***covid**iot*, ***covid** party*, ***corona**sceptic*, ***corona**viruslike*, *etc.*, and *long-hauler*, defined as 'a **COVID**-19 patient who is suffering from [. . .]'). However, the method fails to retrieve words that are not morphologically derived from the patterns and that are related but not specific to the pandemic, i.e. headwords whose defining words do not match such patterns (e.g. no word matches the patterns in the definition and usage examples of *social distancing*). Wiktionary revision logs are the same for all language editions and the number of editions/editors per page can be extracted regardless of any target topic. Details on the processing required to exploit the logs are given in Section 2.1.

2.1 Data processing

The history dump of Wiktionary is a large file released on a regular basis, which contains every version of all articles, stored after each individual contributor's edition. For each revision, the username of the contributor, or the IP address (for unregistered users) is provided, as well as the revision date. The files released on January 1, 2021 were downloaded[3] and processed for the English and French editions of Wiktionary so as to extract, for each month and for each article, the number of revisions and the number of unique contributors.[4] Several pre-processing steps were performed to discard data irrelevant to the present work:

- discussion pages, user talk pages, etc.
- parts of speech other than common noun and proper noun, verb, adjective and adverb.
- pages related only to inflected forms or entries in a language different from the target language (e.g. the page of the English Wiktionary which describes the Vietnamese headword *siêu vi corona* 'coronavirus' was ignored).
- revisions related only to other language sections. For example, the revision dated 5 November 2020 on the *coronavirus* page of the English Wiktionary, that resulted in the addition of the derived term *coronaviraal* to the Dutch language section was ignored.

After discarding irrelevant pages and revisions, more than 14 million revisions remained for the English Wiktionary and more than 27 million for the French language edition.

Studies that focus on Wiktionary (or Wikipedia) revisions differ in whether they take into account the revisions performed by bots and by anonymous users. A bot is

3 https://dumps.wikimedia.org/

4 The computing is an extension of the work done by Sajous et al. (2020) to produce WIND, a resource which contains the dates of inclusion of Wiktionary headwords.

a program devised to automatically perform specific types of revision targeting a range of articles (mainly formatting, importing audio files, etc.). Such automatic editions amount to 45% of the revisions in the English Wiktionary and 62% in the French edition. A contributor to Wiktionary may be identified by a registered account or an IP address. Regular contributors generally create an account while occasional contributors may edit an article "anonymously". Dismissing or taking into account anonymous revisions (which represent 7.4% of the revisions in the English Wiktionary and 4.7% in the French Wiktionary) is a matter of debate. They are discarded in some studies on the grounds that anonymous users are less experienced or trustworthy than registered users and that identifying Internet users by their IP addresses is a rough approximation. Several contributors can indeed use the same IP,[5] while a given contributor can use several IPs,[6] as was the case during the present study (see the discussion on *myroblyte* in Section 3.1). The objection could however apply to registered accounts too: different contributors sharing the same account is probably an exception, but it happens that a single user owns several accounts. Since the present study is concerned with the tendency of articles to be revised many times by possibly many people (not only experienced or reputable Wiktionarians), I was tempted to argue that there is no a priori reason for ignoring anonymous contributions (while revisions performed by bots are not relevant). The quantitative investigations presented in Section 3.1 show that there is no definitive answer as to whether considering or ignoring anonymous contributions is the best option. Regarding qualitative considerations, discarding anonymous contributions poses a risk of overlooking relevant words. For instance, in 2020, the articles for the synonyms *R0*, *basic reproduction number* and *basic reproduction ratio* were created from the same IP address. Whatever their rankings, these words would have gone unnoticed if anonymous contributions had been ignored.

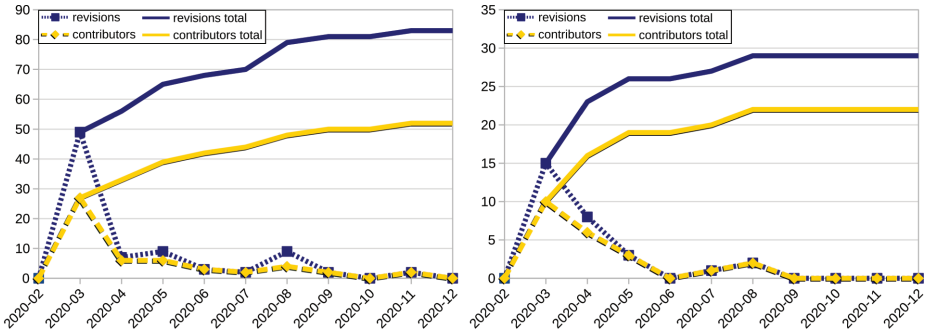
Regarding existing headwords (i.e. those created prior to 2020), it is not difficult to detect new senses added to Wiktionary in 2020 by using the revision log, but the main focus in the present study is on any kind of updates: additions, modifications or replacements of definitions, usage examples, semantic relationships, translations, usage notes, etc. Beyond new meanings, such revisions may indicate the need for article reviews (cf. the examples of *ventilator* in Section 2.4 and *comorbidity*, Section 3.5), which is information that lexicographers may find useful.

5 Especially contributors accessing the Internet from behind an institution firewall.

6 Either intentionally, to deliberately mask one's identity, or unintentionally, as a result of dynamic IP assigning.

2.2 Ranking the new headwords

Looking for the new articles that have the maximum number of revisions and contributors should make it possible to detect headwords related to topical events, notably the Covid-19 pandemic. Figure 1 illustrates the cases of *social distancing* and *flatten the curve*. Dotted lines correspond to the number of revisions/contributors per month, while plain lines represent the total number of revisions/contributors since the creation of the articles.



(a) *social distancing* (b) *flatten the curve*

Figure 1: Number of revisions and contributors for *social distancing* and *flatten the curve*.

The lines for the two words follow the same pattern: a sudden spike of activity when the page is created (typically, when a word comes into usage) and an off-peak period, with occasional contributions (this revision pattern is reminiscent of the pattern described by Renouf (2013) concerning the corpus frequency of *Eyjafjallajökull*, as mentioned in the Introduction). An analysis based on a one-year span is likely to detect the two words *social distancing* and *flatten the curve*, with the former ranking higher (note that the vertical scales in Figure 1 are different). Conversely, words such as, for example, *cognitive bias*, added in April 2020, which received 7 contributions by 4 distinct human contributors are unlikely to be detected. In addition, the two Covid-related words are likely to appear in the first trimester candidate list when performing quarterly analyses.

Wiktionary headwords can be represented in a coordinate system whose axes correspond to the number of revisions and unique contributors of each headword. The 31,107 new headwords added to the English Wiktionary in 2020 are depicted by a scatterplot in Figure 2. The article *COVID-19* was modified 115 times by 63 unique contributors and is therefore represented by the coordinate point (115, 63). A given coordinate point can correspond to several headwords. For example, 17,415 words have not been modified since their creation. They are all represented by the coordinate point (1, 1). A less extreme case, the headwords *self-isolate* and *Wuhan coronavirus*

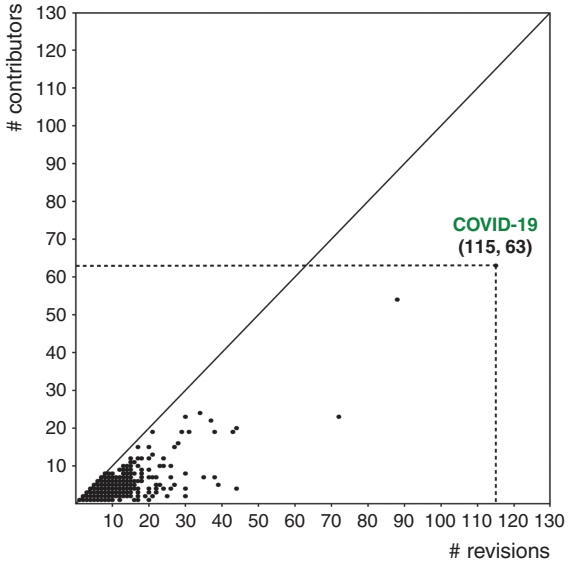


Figure 2: Distribution of the 31,107 lemmas added to the English Wiktionary in 2020.

were each modified 16 times by the same number (11) of distinct contributors. They are therefore represented by the same coordinate point (16, 11). All the points are located along or below the diagonal line (i.e. the *contributors=revisions* line) because there obviously cannot be more contributors than contributions for a given headword. The points along the diagonal line are those for which each revision was made by a distinct contributor. For instance, the words *coronoia*, *Wuhan shake* and *Zoombombing* were modified 8, 7 and 4 times, respectively, each time by a different contributor.

This kind of diagram enables a geometrical interpretation of the headwords' location. The rightmost points of the diagram (i.e. those with the highest abscissa) are those corresponding to the most heavily revised pages. The upmost points (i.e. those with the highest ordinates) are those corresponding to the headwords edited by the greatest diversity of contributors. Given two points having the same abscissa, the upmost point corresponds to the headword revised by a greater diversity of contributors. For example, the two headwords *flatten the curve* and *Medusavirus* 'a virus that infects amoeba' were each modified 30 times in 2020, and have a similar creation date (February and March, respectively). However, the 30 edits of *flatten the curve* were made by 23 distinct contributors, compared to 8 contributors for *Medusavirus*.

Four ranking scores were tested to detect potential Covid-related neologisms. Given a headword *h*, the ranking scores are defined as follows:

1. $revs_h$ = raw number of revisions for *h*
2. $contribs_h$ = raw number of distinct contributors for *h*

3. $dist_h$ = distance from the origin of the coordinate system to the point ($revs_h$, $contribs_h$) = $\sqrt{revs_h^2 + contribs_h^2}$
4. $prod_h$ = product of the number of revisions and the number of contributors[7] = $revs_h x contribs_h$

The geometrical interpretation of the first three ranking scores is depicted in Figure 3 (the product score has no geometrical interpretation). The revisions-based score orders the headwords from right to left. When two headwords have the same abscissa (i.e. the same number of revisions), they are ordered by their ordinate value (their number of contributors), i.e. the upmost headword is ranked first. For instance, the initially equally ranked (4th position) *Wuhan pneumonia* and *Mount Mayon* (a volcano in the Philippines), whose coordinates are (44, 20) and (44, 4), are finally ranked fourth and fifth. Similarly, the contributors-based score orders the headwords from top to bottom. When two headwords have the same ordinate (i.e. the same number of contributors), they are ordered by their abscissa value (i.e. their number of revisions), i.e. the rightmost headword is ranked first. For instance, the equally ranked (4th position) *myroblyte* (see Section 3.1) and *flatten the curve*, whose coordinates are (72, 23) and (30, 23), are finally ranked fourth and fifth, which, in this case, is not the best option. Finally, the distance-based score orders the headwords according to their remoteness from the origin of the coordinate system.

2.3 Ranking the existing headwords

The scores introduced in Section 2.2, devised to rank Wiktionary new entries, are based on raw numbers of revisions and contributors. Using raw numbers to rank existing entries would not make any sense. We can indeed expect the revision rate of Wiktionary articles to depend on the nature of the entry, i.e. whether it is a frequent or a rare word, polysemous or monosemous, belonging to a specialised field or to the general language (knowing that these characteristics are related). For example, the larger spike observed in 2020 for *coronavirus* when compared to that observed for *virus* in Figure 4(a) is all the more noticeable as the article corresponding to the frequent and polysemous word *virus* is regularly revised, while the entry *coronavirus* rarely is. Another telling example is the number of revisions of *masks*, *facemask* and *surgical mask*, as depicted in Figure 4(b). If we consider the 2020 period globally, the three articles received a similar number of revisions (36, 36 and 34, respectively). However, their "usual" yearly revision values are very different.

7 The product can be normalised to values between 0 and 1 by dividing the score by the maximum number of revisions and the maximum number of contributors. Normalising the product, however, is useless since it does not change the ranking order.

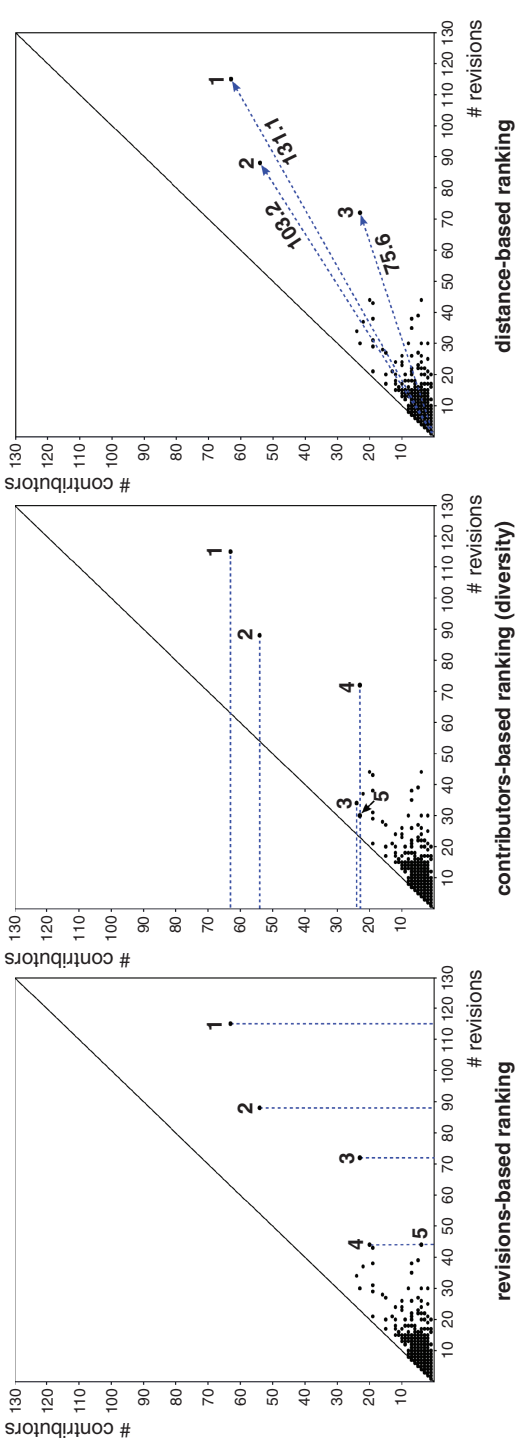
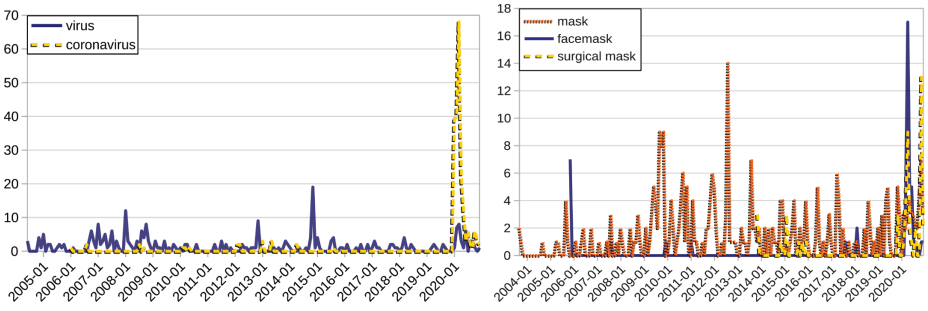


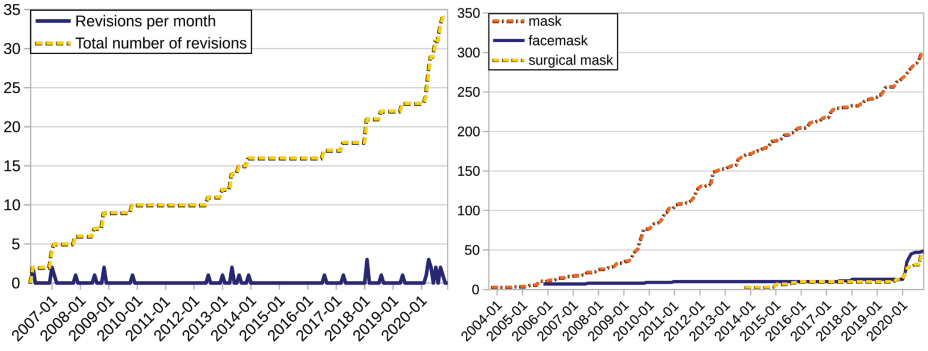
Figure 3: Ranking scores based on revisions, contributors and distance.

Another way to uncover unusual increases in the number of editions is to represent the total number of revisions (or contributors) for a given headword, as depicted in the time-graphs in Figure 5. Figure 5(a) shows that revisions performed over several consecutive months may result in jumps that can be observed for the resulting period. Figure 5(b) shows the total number of revisions for *mask*, *facemask* and *surgical mask*. The increase in the number of revisions is in line with the usual trend for *mask*, while the increases for *facemask* and *surgical mask* are more noticeable.



(a) *virus* and *coronavirus* (b) *mask, facemask* and *surgical mask*

Figure 4: Monthly revision frequencies in the English Wiktionary.



(a) monthly and total revisions for *comorbidity* (b) total number of revisions for *mask, facemask* and *surgical mask*

Figure 5: Monthly and total number of revisions in the English Wiktionary.

The boxplot in Figure 6 statistically confirms these observations: With respective mean values of 3.1, 5.9 and 16.7 (median values of 0, 2 and 15), *facemask* was revised 12 times more than usual, *surgical mask* 5.8 times more and *mask* only 2.2 times more. The two upmost circles in the figure (which represent extreme values) correspond to the 2020 number of revisions for *facemask* and *surgical mask* (another extreme value, observed for *facemask*, correspond to the 7 revisions made

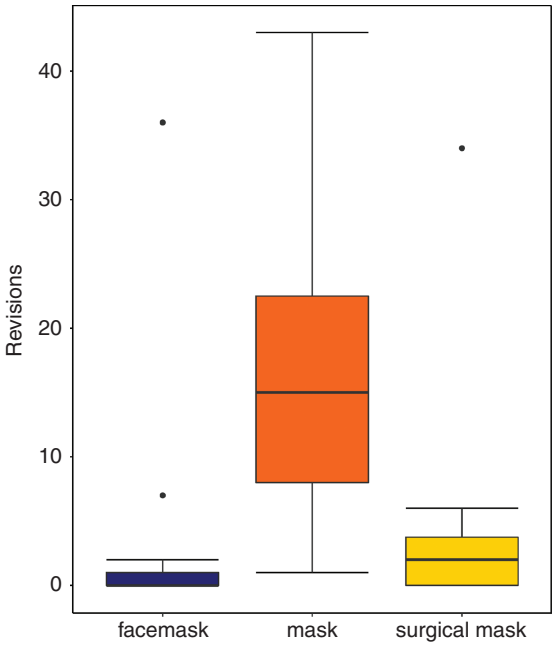


Figure 6: Distribution of the yearly number of revisions in the English Wiktionary.

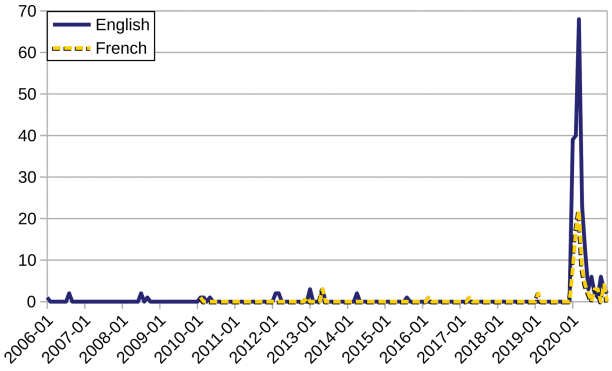


Figure 7: Number of revisions for *coronavirus* in the English and French Wiktionary.

in November 2005 when the article was created). Conversely, the 2020 value for *mask* is not identified as an extreme value.

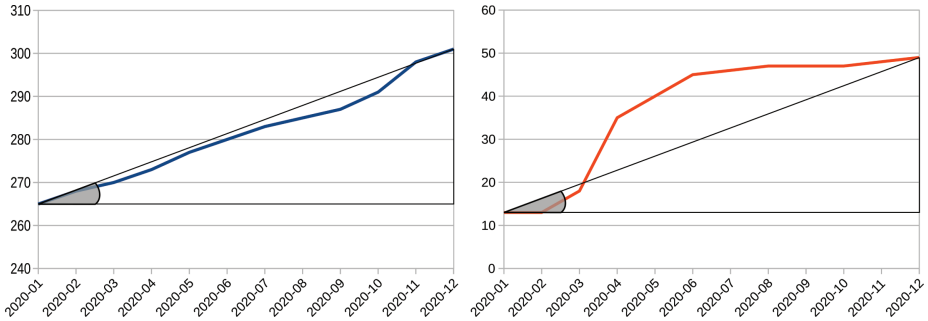
Regardless of the linguistic characteristics of the headwords, the revision rate may differ from one language edition to another. For example, the evolution of the number of revisions for the article *coronavirus* follows a similar trend in the English and French Wiktionary, but with different magnitudes (cf. Figure 7).

Detecting a particular "stir" around a headword is like looking for the extreme values of the boxplot in Figure 6. It therefore requires comparing the number of revisions over a given period to its *usual* revision rate, just as extracting keywords by comparing a focus corpus to a reference corpus requires the use of relative frequencies, not raw frequencies. The scores used to detect the most unusually revised articles compare the number of revisions (or contributors) over a given period to the usual (mean or median) number of revisions (or contributors) over similar time spans. Given a target period p and headword h, the scores are calculated as follows:

1. $avgRevsRatio_p(h) = \frac{1 + revs_p(h)}{1 + avg(revs(h))}$
2. $medianRevsRatio_p(h) = \frac{1 + revs_p(h)}{1 + median(revs(h))}$
3. $avgContribsRatio_p(h) = \frac{1 + contribs_p(h)}{1 + avg(contribs(h))}$
4. $medianContribsRatio_p(h) = \frac{1 + contribs_p(h)}{1 + median(contribs(h))}$

Medians and averages are calculated over the period that spans from the creation of the article corresponding to the headword h to the month before period p begins. A constant (here, 1) is added to the denominator (and to the numerator, for balance) so as to avoid divisions by zero. The median value can be null (as we saw above with *facemask*), but the average value should not be, as all the articles have been edited at least once (when they were created). However, certain revisions (performed by bots or anonymous contributors) are discarded in some of the experiments described below, which makes the addition of a constant necessary.

Ranking the headwords according to the slope of the curve for a given period was tempting. The slope accounts for the increase in the number of revisions (or contributors) over a given time span. For both the English and French language editions, the scores based on slope values performed poorly. As the slope is equal to the ratio between the number of revisions (or contributors) and the length of the time span, its value is proportional to the raw number of revisions (or contributors), and disregards the corresponding *usual* amount, which explains the low results. Figure 8, which consists of two enlargements of Figure 5(b), illustrates the situation for *mask* and *facemask*. Although it is clearly visible in Figure 5(b) that the two words have different usual revision rates, Figure 8 shows that their slope values on the 2020 period are the same. The slope-based score was therefore abandoned and is not further discussed.



(a) *mask* (b) *facemask*

Figure 8: Headwords with similar slope values over the year 2020.

2.4 Annotation of headwords

In order to assess the performances of the ranking scores, several sets of top-ranked headwords were annotated for each metric introduced in Section 2.2 with the binary flag 'related to Covid-19' (yes/no). New headwords were annotated to detect true formal neologisms, or words that already existed but were too rare or too specialised to enter a dictionary before. Existing headwords were annotated to detect semantic neology or articles that potentially deserve an update. The interpretation of the "relatedness to Covid-19" criterion encompasses words whose referents are in a direct or indirect relationship with the virus and the disease, medical care, controlling the spread of the pandemic, statistical analysis, consequences of the pandemic on professional activities and social lives, as well as humorous coinages. For the English and French language editions of Wiktionary, I annotated, for each data source,[8] and each relevant ranking score (cf. Sections 2.2 and 2.3):

- the first 200 new headwords, ranked over the whole 2020 period;
- the first 200 existing headwords, ranked over the whole 2020 period;
- the first 100 new headwords, ranked over each trimester;
- the first 100 existing headwords, ranked over each trimester.

The different (overlapping) sets of headwords represented a total of 3,070 English and 3,168 French words to be annotated. The words were stored in four groups of spreadsheets, setting apart new and existing entries, English and French words. Each word was accompanied by a hyperlink to the online article, along with the definition of the first sense as it stood in Wiktionary. In most cases, the annotation

8 Data sources are discussed in Section 3.1.

was rather self-evident.[9] Conversely, some headwords required further investigation, e.g. reading the definition or looking for additional encyclopaedic knowledge. The definition taken from Wiktionary was intended to help annotate new entries rather than existing, polysemous ones, that require a look at the online article (and, sometimes, at the differences between the versions of articles before/after 2020) or other sources. Encyclopaedic knowledge was especially necessary for annotating words related to the fields of pathology and pharmacology. For example, two drugs related to hydroxychloroquine were positively annotated: *pamaquine* (existing in the English Wiktionary since 2009) and *quinium* (added to the French Wiktionary in 2020). Revisions of entries denoting other drugs may have been motivated by pandemic-related reasons, such as drugs used in the treatment of respiratory diseases (e.g. *bambuterol*, used in the treatment of asthma). However, in the absence of clear evidence of a relation with Covid-19, such headwords were negatively annotated. On the same lines, *extractor fan* may be related to air purification that helps prevent the spread of the virus. However, this entry, created in June 2019, makes no reference to such a meaning. It was deemed too general and was therefore negatively annotated. Conversely, *ventilator* was annotated positively. Although not related to the Covid pandemic when used as a synonym of *fan*, the 2020 updates clearly target the *medical ventilator* sense. The previous synonymic definition '(medicine) A respirator' was changed to '(medicine) A machine that moves breathable air into and out of the lungs of a patient who is unable to breathe sufficiently', with *respirator* now appearing as a hypernym. A picture of a medical ventilator has been added, as well as the derived term *tank ventilator* and numerous translations.

In the case of French borrowings from English, the prior annotation of the English word helped. For instance, it would have been hard to come to a decision in the case of the neologism *doomscrolling* '(informal) The practice of continually reading Internet news about catastrophic events' retrieved from the English Wiktionary by only reading its definition. The definition may refer to one's state of astonishment when following the news after the pandemic outbreak or when the first lockdown was decided. But *catastrophic events* can, sadly, designate numerous other facts. The problem was finally easily solved, due to the presence of the *Coronavirus* category at the bottom of the article. In the French Wiktionary, the article *doomscrolling*, which mentions the borrowing from English, but does not mention the pandemic in the definition or the usage examples, is devoid of any topical category. The previous annotation of the English word led to a positive annotation. In the French Wiktionary, the Anglicism *contact tracing* was added in April 2020, and its annotation did not raise any difficulty. Conversely, *tracking* was debatable. Until

9 Given the number of regionalisms, occasionalisms and dated words, in addition to words of different subcultures, a quick look at the definition was necessary, even for French words. *Self-evident* therefore means non-ambiguous here, rather than *immediate*.

2020, the corresponding article described the English gerund. In April 2020, the description of the French Anglicism *tracking* was added and defined as *surveillance de masse des populations par pistage de tous les citoyens* 'mass surveillance of populations by tracking all citizens'. Though the definition does not refer to controlling the spread of the pandemic, the three usage examples are related to this goal, in particular to the use of cell phones for contact tracing which was then a matter of debate, echoing discussions on other freedom-destroying laws. Knowledge of current events helped annotate the headword positively. In the English Wiktionary, *pastette* was only described as the plural form of the Italian *pastetta*, a variant of *pastella* 'batter' until 2020. The English noun was added in 2020 and described as a synonym of '*Pasteur pipette*'. Although the use of this instrument is not specific to blood sample collection for Covid-19 testing, the addition of the entry to the dictionary is obviously related to the pandemic and the headword was therefore positively annotated.

Some additions and some words already in the dictionary refer to things of the past. For example, *méthode Raspail* entered the French Wiktionary in March 2020 and refers to a hygiene system named after its creator François Vincent Raspail, mainly based on handwashing and dating back to the nineteenth century. Despite the lack of exclusive connection to Covid-19, the 2020 addition of this old preventive measure, simultaneously with the revisions of *gel hydroalcoolique* 'alcogel' and the addition of *geste barrière* 'practice intended to avoid the spread of a virus' argued in favour of a positive annotation. Although the 7 revisions of *quarantine flag* (in the nautical field, the flag that was hoisted by a ship to signal that it had contagious disease aboard) by 5 distinct contributors to the English Wiktionary, which resulted in a rewording of the definition and the addition of four translations and a reference, are striking, the word was deemed too indirectly related to the pandemic to be assigned a positive annotation (the flag is said to have been *formerly* hoisted and the reference dates back to 1916).

Lastly, some words were close to being given a positive annotation they did not deserve. With the videoconferencing software in mind, it was tempting to annotate positively the French *zoom* and *zoomer*, and the English *zoomer*, without checking the corresponding definitions. However, the French words are only related to the camera lens and the revisions of the English *zoomer* are related to the generational designation (*active boomer*, *member of Generation Z*).[10] The British slang *lurgy*, denoting a fictitious, or uncategorised, infectious disease with cold or flu-like symptoms, that renders one unable to work, was a good candidate. However, several occurrences found in newspaper articles dating back to Fall 2019 (i.e. *before times*), whose topic

10 Derivatives of *Zoom* (the videoconferencing software) retain the initial upper case in English, according to the English Wiktionary.

was the ironically named season "of dreaded lurgy" in reference to people seeking excuses for work absenteeism, led to the word being rejected.

The scatterplot in Figure 9 depicts, for the English and the French Wiktionary, the Covid-related words in blue circles and the negatively annotated words in yellow circles. Grey triangles correspond to the superposition of several words, some of which are related and others not related. Empty circles close to the origin of the coordinate system correspond to words that were not annotated because they did not rank high enough in any configuration.

Simple linear regressions were performed for the two language editions and the red lines represent the models fitting the distributions. A first observation is that the points corresponding to positively annotated headwords are mostly above the regression line. Given several articles that received the same number of revisions, the articles related to the Covid pandemic are those edited by the largest number of contributors. This finding is further investigated in Section 3.2.

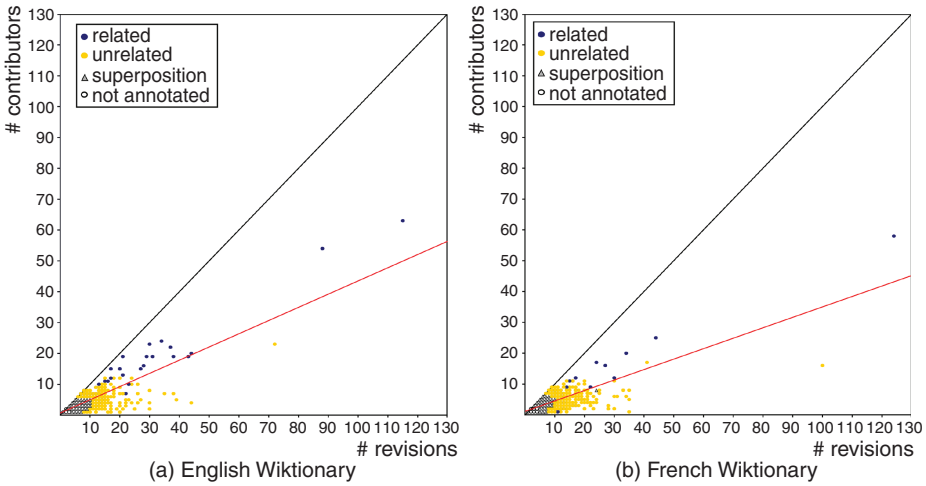


Figure 9: Distribution of the headwords added to Wiktionary in 2020 with respect to revisions and contributors.

3 Results

3.1 Contributor types

The performances of the ranking scores were calculated from different data sources in order to evaluate which kinds of revisions were worth taking into account with respect to the contributor types (cf. Section 2.1). Figures 10(a) to 10(d) show the results

obtained when considering all revisions compared to the results obtained when ignoring revisions performed by bots and/or by anonymous contributors. Figures 10(a) and 10(b) correspond to the English Wiktionary while Figures 10(c) and 10(d) correspond to the French Wiktionary. For the two language editions, the results obtained with the ranking score based on the number of revisions are visible on the left-hand side, i.e. in Figures 10(a) and 10(c). The results obtained with the ranking score based on the number of contributors are visible on the right-hand side, i.e. in Figures 10(b) and 10(d). The line chart shows the percentage of new entries related to the pandemic on the ordinate, as a function of the number of candidates examined (ranging from 1 to 200), on the abscissa.

Regardless of the data sources, the results are better for the English Wiktionary than for the French edition and the ranking score based on the number of contributors performs better than that based on the number of revisions. These observations are further discussed in Section 3.2. The experiments confirm that discarding the revisions performed by bots generally improves the results. Regarding anonymous contributions, conclusions are mitigated. Discarding these contributions significantly lowers the results obtained with the French Wiktionary. It also lowers the results obtained with the English Wiktionary when using the ranking score based on the number of revisions, but it improves those based on the number of contributors. In Figure 10(b), a clear advantage is visible in the top of the list, up to rank 18. The first downshift observed in this figure for the data involving anonymous contributors is due to the headword *myroblyte* 'a saint whose relics or place of burial produce or are said to have produced the Oil of Saints'. With 71 revisions coming from 22 IP addresses, the headword reaches the third rank. A closer look revealed that the addresses were most likely assigned to the same machine.[11]

Given that the ranking score based on the number of contributors outperforms the ranking score based on the number of revisions whoever the authors of the revisions, the experiments described in Section 3.2 were performed with the source of data that produced the best results with the contributors-based score, for each language edition, i.e. the "no bots, no anonymous" option for the English Wiktionary and "no bots, with anonymous" for the French edition. The best choice regarding data sources, however, is unstable. The experiments conducted for each trimester led to better results for the English Wiktionary when the anonymous contributions were taken into account. The results presented in Section 3.3 were produced with the "no bots, with anonymous" option.

11 All of them have the same two left numbers, and are probably due to dynamic IP assigning. A comparable number of revisions stemming from the same addresses is observed for the same headword in the French Wiktionary.

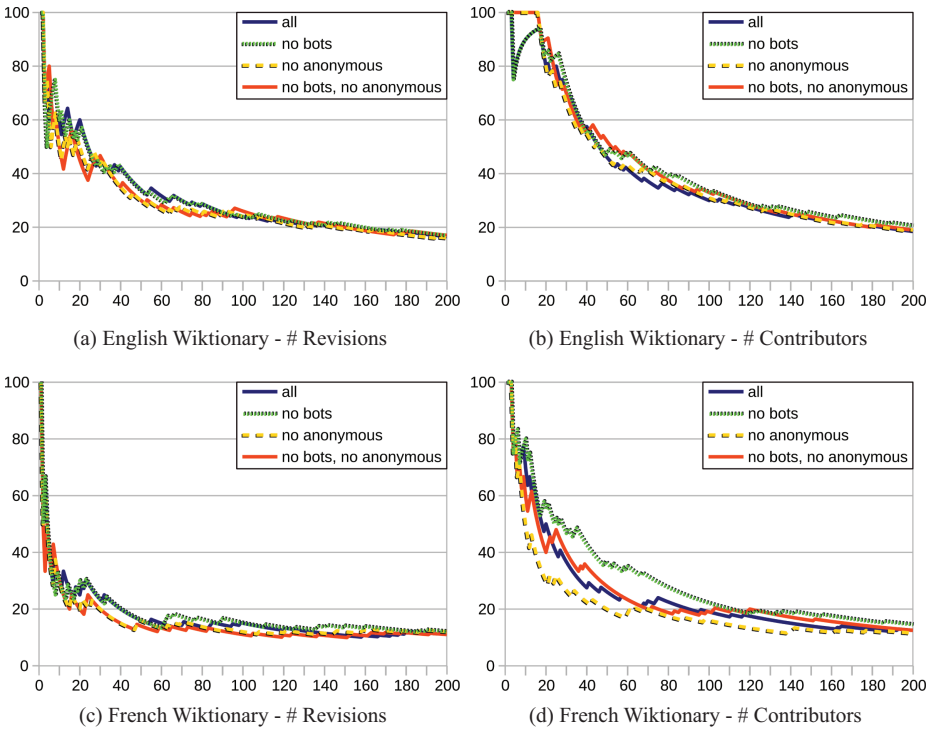


Figure 10: Influence of the contributor types on the ranking scores.

3.2 Yearly ranking of the new headwords

The more up-to-date a dictionary is, and the more exhaustive its list of headwords, the more likely a new headword is to be a neologism. The top-ranked new additions to Wiktionary, according to the metrics introduced in Section 2.2, were therefore inspected to detect formal neologisms. Table 1 reports, for the English edition of Wiktionary, the 20 most heavily edited new entries in 2020, sorted by number of revisions, number of unique contributors, and by the two combinations (product and distance) introduced in Section 2. Grey cells indicate headwords that are not related to the pandemic, while the others are.

When ordered by number of revisions, less than half of the top-ranked entries (9 out of 20) are related to the pandemic. When ordered by number of contributors, 90% of them (18 out of 20) are positively annotated. This result seems to confirm the initial hypothesis: the most frequently edited Wiktionary pages, especially pages edited by many distinct contributors, can help detect topical neologisms.

Looking down the list after the 20th rank of the contributors-based score helps detect the following relevant words:[12] *fever clinic* (21), *self-isolate* (27), *coronoia* (29), *before times* (38), *doomscrolling* (41), *Wuhan shake* (42), *maskne* (43), *SARS-CoV* (48), *rat-licker* (54), *contact trace* (58), *maskhole* (61), *elbow bump* (74), *mascne* (88), *plandemic* (94), *China virus* (96), *Covidtide* (103), *corona virus* (134), *case fatality rate* (144), *coronasceptic* (180), *elbow shake* (182), *antimasker* (238), *covid-19 party* (250), *long-hauler* (272), *corona belly* (484), *community spread* (492), etc.

Table 1: 20 most frequently edited pages in the English Wiktionary 2020 additions, according to different ranking scores (data source: no bots, no anonymous).

Rank	# Revisions	# Contributors	Product	Distance
1	COVID-19	COVID-19	COVID-19	COVID-19
2	social distancing	social distancing	social distancing	social distancing
3	Mount Mayon	covidiot	Wuhan pneumonia	Mount Mayon
4	Wuhan pneumonia	flatten the curve	Wuhan virus	Wuhan pneumonia
5	Wuhan virus	Wuhan virus	covidiot	Wuhan virus
6	Bicol Region	Wuhan pneumonia	flatten the curve	Chinese virus
7	Peja	covid	Chinese virus	Bicol Region
8	Chinese virus	social distance	covid	Peja
9	Berat	SARS-Co V-2	infectious disease specialist	covidiot
10	Medusavirus	Chinese virus	social distance	flatten the curve
11	Vlorë	infectious disease specialist	SARS-CoV-2	Berat
12	Kizilsu	contact tracing	contact tracing	Medusavirus
13	covidiot	self-isolation	Kung Flu	covid
14	infectious disease specialist	Kung Flu	Wuhan flu	infectious disease specialist
15	flatten the curve	Wuhan flu	Medusavirus	Vlorë

12 Whether such neologisms should be added to a dictionary headword list is not the focus of the present research and depends on the editorial policy. The final decision is up to the lexicographer, and is not discussed here.

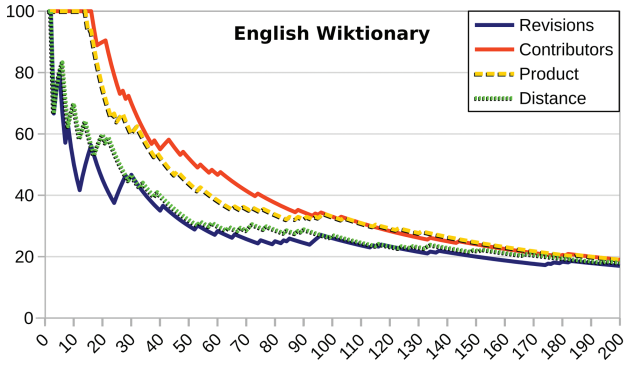
Table 1 (continued)

Rank	# Revisions	# Contributors	Product	Distance
16	covid	COVID	self-isolation	Kizilsu
17	world map	Spleef	world map	world map
18	Accompong	Trumpster fire	Bicol Region	social distance
19	Arlberg	self-quarantine	Mount Mayon	SARS-CoV-2
20	sinoatrial node	Wuhan coronavirus	wokefish	contact tracing

The same ranking score applied to the French Wiktionary provides the following Covid-related words: *Covid-19* (1), *covid* (2), *COVID-19* (3), *covidiot* (5), *déconfinement* 'deconfinement, lockdown removal' (6), *distanciation sociale* 'social distancing' (8), *covidé* 'sick from Covid-19' (9), *déconfiner* 'deconfine, remove lockdown' (10), *reconfinement* 'reconfinement, new lockdown' (12), *masque chirurgical* 'surgical mask' (18), *coronavirus 2 du syndrome respiratoire aigu sévère* 'SARS-CoV-2' (19), *méthode Raspail* 'hygiene system based on handwashing' (21), *télétravaillable* '(work) that can be done by teleworking' (25), *covidien* 'related to, or sick from Covid-19' (27), *cas contact* 'contact case' (31), *distanciation physique* 'physical distancing' (34), *gel hydroalcoolique* 'alcogel' (35), *doomscrolling* (49), *infodémie* 'infodemics' (53), *hydroxychloroquine* (58), *Covid* (60), *pneumonie de Wuhan* 'Wuhan pneumonia' (67), *syndémie* 'syndemic' (121), *antimasque* 'antimask' (124), *démerdentiel* '(informal) activity performed with the means available' (133), *coronasceptique* 'coronasceptic, who denies the reality or the aftermath of the coronavirus' (136), *autoconfinement* 'self-isolation' (142), *Covid positif* 'Covid positive' (156), *coronapiste* 'temporary cycle lane built during the Covid-19 pandemic' (186), *raoultiste* 'supporter of Pr. Raoult' (214), *candidat-vaccin* 'vaccine candidate' (308), *coronavirussé* 'sick from Covid-19' (361), *tempête immunitaire* 'cytokine storm' (432), etc.

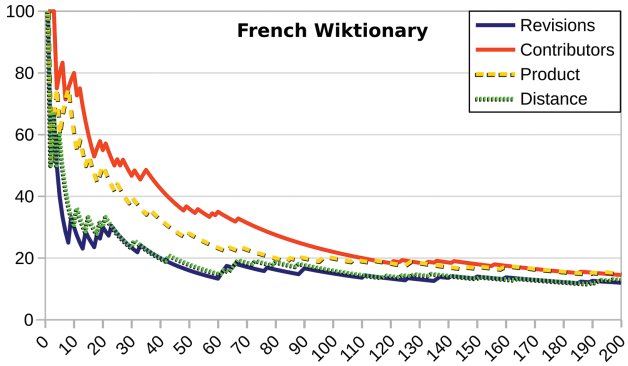
The performances of the different ranking scores are further illustrated in Figures 11 and 12 for the English and French language editions. For both languages and for the four ranking scores, the line charts are similar to those in Section 3.1 and show the percentage of new headwords that are related to the pandemic on the ordinate, as a function of the number of candidates considered (ranging from 1 to 200) on the abscissa.

The same observation can be made for both language editions: the ranking based on the number of unique contributors performs markedly better than the ranking based on the number of revisions. The number of contributors alone even outperforms the combinations of the two measures (with the "product" score only slightly improving the results locally from ranks 115 to 179 and equalling the results of the contributors-based score from rank 187 onwards).



	Ranking score			
Rank	Revs	Contribs	Prod	Dist
5	80	**100**	**100**	80
10	50	**100**	**100**	70
25	40	**76**	64	52
50	30	**52**	44	32
100	26	**33**	**33**	27
200	17	**19**	**19**	18

Figure 11: Performance of the ranking scores for the English Wiktionary new headwords.



	Ranking score			
Rank	Revs	Contribs	Prod	Dist
5	40	**80**	60	60
10	30	**80**	60	60
25	28	**52**	30	30
50	16	**36**	28	18
100	15	**22**	20	16
200	12	14.5	**15**	13

Figure 12: Performance of the ranking scores for the French Wiktionary new headwords.

In order to explain the lower results obtained with the French language edition, a simple linear regression was performed, as depicted in Figure 9 (Section 2.4), where the regression lines appear in blue. Linear regressions are usually performed to confirm that two variables are significantly related. In the case of the number of revisions and contributors, we already know that this is the case, but we are interested in the regression coefficients. With a value of 0.43,[13] the slope of the regression line for the English Wiktionary is greater than that for the French edition (slope value of 0.34).[14] This means that, given two articles selected at random in the English and French editions, that have the same number of revisions, the article from the English Wiktionary is likely to have been modified by a greater number of contributors than that in the French Wiktionary. This finding, combined with the better results obtained for the English language edition, is an argument in favour of the relevance of the "diversity"

13 $F(1, 31104) = 6.64e^{+4}$, p-value < 0.001.
14 $F(1, 11619) = 1.545e^{+4}$, p-value < 0.001.

measure. To go further in this direction, the coordinates of the positively and negatively annotated headwords originating from the English Wiktionary were set apart in two distinct scatterplots, as shown in Figure 13.

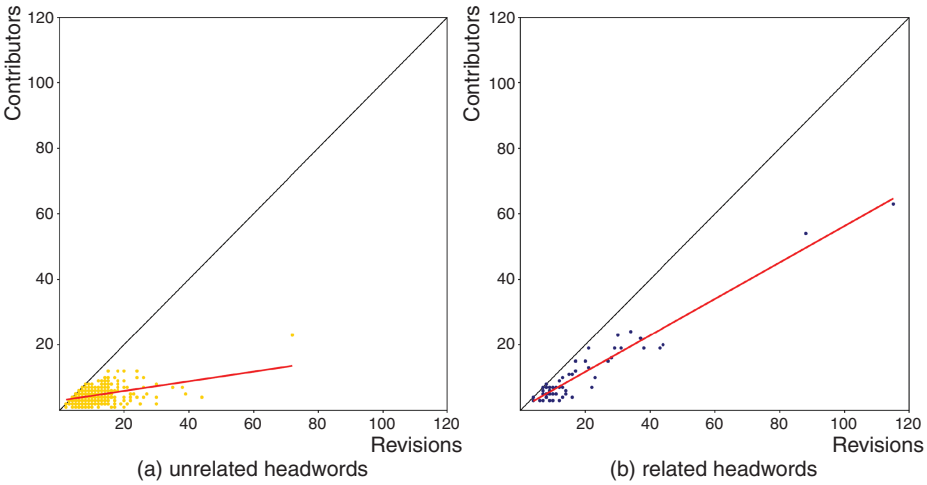


Figure 13: Distribution and regression lines for the related vs unrelated English headwords.

For each distribution, a simple linear regression was performed. The slope coefficients are 0.56[15] for the Covid-related words and 0.15[16] for the words that are not related (regression lines are depicted in red in Figure 13). This means that, given a number of revisions, a positively annotated headword is likely to have been revised by a larger number of contributors than a negatively annotated headword with the same number of revisions.

For each headword of the two annotation sets, the ratio between the number of contributors and the number of revisions was calculated. This ratio can be understood as the "local" diversity for individual headwords: diversity(h) = contributors(h) / revisions(h). Its maximum value is 1 when all the revisions were made by distinct contributors, i.e. when contributors(h) = revisions(h). The ratio is low when all the revisions were made by the same contributor, and especially when the number of revisions is high. The boxplots in Figure 14 represent variations of this ratio. Figure 14(a) shows the difference in diversity between related and unrelated headwords in the English Wiktionary. The median value is 0.62 for the positively annotated words and 0.56 for the words annotated negatively. A Welch two-sample t-test shows that the difference

15 $F(1, 51) = 730.2$, p-value < 0.001.
16 $F(1, 633) = 119.7$, p-value < 0.001.

is statistically significant.[17] The same experiment was conducted on the French Wiktionary. For this language edition, the diversity is also higher for the positively annotated words than for words annotated negatively. This time, however, the difference was not statistically significant.

To conclude on the importance of diversity, we compared the diversity ratio for the 1598 annotated new headwords (688 originating from the English Wiktionary and 910 from the French Wiktionary), regardless of the annotation. The variation in diversity is depicted in Figure 14(b). The diversity is greater in the English Wiktionary (median value of 0.57 compared to 0.5 for the French Wiktionary) and the difference is statistically significant.[18] Once again, this result, together with the lower performances observed for the French Wiktionary, drives home the importance of the diversity measure.

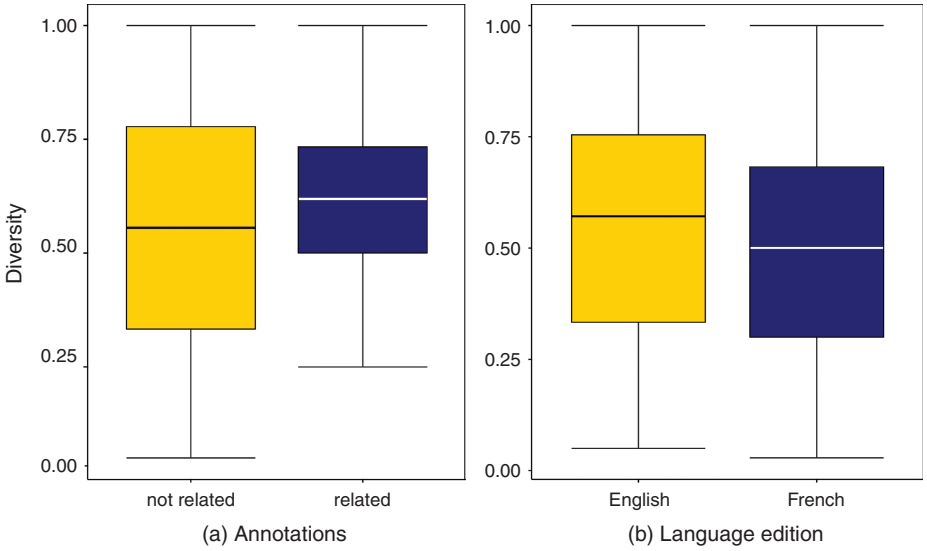


Figure 14: Variation in the diversity ratio depending on annotations and the language editions.

3.3 Quarterly ranking of the new headwords

The method proposed above is based on the analysis of Wiktionary revisions on a whole year basis. However, retrospectively identifying neologisms one year after the Covid-19 outbreak (and after lists of neologisms have proliferated) might seem like making a weather forecast for the day before. For dictionaries such as the

17 t(71) = -2.0905, p-value < 0.05.

18 Welch two-sample t-test: t(1423) = 3.7912, p-value < 0.001.

French *Petit Robert* and *Petit Larousse*, which are updated once a year, the method based on yearly analyses makes sense.[19] However, some online dictionaries are updated more or less continuously and, among them, the OED is ordinarily updated quarterly. In order to assess the validity of the proposed method for updating such dictionaries under conditions closer to reality, I examined what would have been the top-ranked entries by the end of the four 2020 trimesters (thereby mimicking the quarterly updates that usually occur in the OED). The number of pandemic-related neologisms that would have been detected among the first 100 candidate headwords by the end of each trimester, according to the contributors-based score, is reported in Table 2.

Table 2: Number of Covid-related additions, depending on the number of candidates inspected.

# candidates	English Wiktionary				French Wiktionary			
	T1	T2	T3	T4	T1	T2	T3	T4
5	4	5	2	2	2	4	0	3
10	8	9	4	3	3	6	1	4
25	12	12	12	6	3	9	4	6
50	18	14	14	8	3	11	6	8
100	20	18	16	13	4	14	9	16

The relevant words retrieved from the English Wiktionary are listed below. The headwords in regular type font are those which were already detected during the previous trimesters while headwords in boldface indicate previously undetected words:[20]

- **T1: social distancing**, **COVID-19**, **Wuhan pneumonia**, **flatten the curve**, **Wuhan coronavirus**, **SARS-CoV-2**, **Kung Flu**, **Wuhan virus**, **COVID**, **social distance**, **Wuhan flu**, **SARS-CoV**, **case fatality rate**, **Chinese virus**, **self-isolate**, **covid**, **community spread**, **self-isolation**, **self-quarantine**, **noncoronaviral**
- **T2:** COVID-19, social distancing, **infectious disease specialist**, **covidiot**, SARS-CoV-2, flatten the curve, Wuhan virus, self-isolation, COVID, Chinese virus, social distance, **corona virus**, self-quarantine, **elbow shake**, Kung Flu, SARS-CoV, **Rona**
- **T3:** COVID-19, covid, social distancing, **maskne**, **mascne**, **plandemic**, Chinese virus, **contact tracing**, **covid-19 party**, covidiot, **doomscrolling**, Wuhan shake, coronoia, **antimasker**, social distance, SARS-CoV-2

19 No Covid-related neologism was added to the printed *Petit Robert* in 2020 (i.e. to the 2021 edition) but some words were added to the online version, which, for the first time, became out of sync with the paper version. See: https://orthogrenoble.net/mots-nouveaux-dictionnaires/entrees-petit-robert-2021/ (last access: 1 June 2021).

20 Discarding previously detected words and upshifting words of lower ranks only results in the addition of *coronapocalypse* at the end of the list of the second trimester.

– **T4: fever clinic**, **rat-licker**, COVID-19, **before times**, China virus, **Covidtide**, Wuhan flu, **coronasceptic**, **long-hauler**, covidiot, Wuhan virus, **long Covid**, **long covid**

In the French Wiktionary, the relevant words detected are the followings:
– **T1: Covid-19, COVID-19, distanciation sociale, pneumonie de Wuhan**
– **T2:** Covid-19, **déconfinement**, **covidiot**, **covid**, **covidé**, **méthode Raspail**, **infodémie**, **déconfiner**, distanciation sociale, **coronavirus 2 du syndrome respiratoire aigu sévère**, **gel hydroalcoolique**, COVID-19, **distanciation physique**, **masque chirurgical**
– **T3: doomscrolling**, covid, **démerdentiel**, COVID-19, Covid-19, **coronapiste**, **taux d'incidence** 'incidence rate', **coronavirussé**, déconfiner
– **T4:** Covid-19, covid, **télétravaillable**, **reconfinement**, **syndémie**, covidiot, déconfinement, **candidat-vaccin**, **cas contact**, coronavirus 2 du syndrome respiratoire aigu sévère, **infectivité** 'infectivity', **antivaccinisme** 'antivaccinism, opposition to vaccination', **télétravaillabilité** 'ability (of a work) to be done by teleworking', **antimasque**, **covidien**, **coronasceptique**

Some of the words detected are true formal neologisms (e.g. *COVID-19*, *Wuhan flu*). Conversely, numerous new headwords already existed before their addition to the Wiktionary list of headwords, as happens in commercial dictionaries. For example, *case fatality rate* has been mentioned by the Office québécois de la langue française in the terminological record *taux de létalité* of its *Grand Dictionnaire Terminologique* since 2009,[21] and probably had long been used by specialists of epidemiology and statistics before it was inventoried in the term bank. The sudden spread of the word in the general language, due to the – no less sudden – spread of the pandemic, motivated the creation of the corresponding article.

Reviewing the lists produced by varying the ranking scores and the data sources is a good idea, as it retrieves headwords that the "globally better" configuration misses. For example, with 6 revisions performed by only 2 registered contributors in the third trimester, *supercontaminateur* 'superspreader' does not rank high enough to be detected by the contributors-based score but ranks 19th with the revisions-based score applied to the "no bots, no anonymous" dataset.

3.4 Yearly ranking of the existing headwords

The same experiments were conducted for existing headwords as for new headwords, by varying the source of data and the ranking scores. Only the main results are reproduced here, while others are summarised.

21 http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=100408 (last access: 10 June 2022).

Just as for the analysis related to new headwords, discarding anonymous contributions slightly improved the results for the English dictionary (the "no bots, no anonymous" option was therefore chosen). As for the new headwords, the scores based on the number of contributors provide better rankings than those based on the number of revisions in the English Wiktionary, as shown in Figure 15. Regarding median and mean values, using one or the other alternatively improves or lowers the results locally. The ratio using the average number of contributors reaches the best result, with 18% of the 100 best-ranked headwords being related to Covid: *coronavirus, hydroxychloroquine, rona, lockdown, superspreader, surgical mask, pandemic, facemask, corona, herd immunity, MERS, MERS-CoV, Coronavirus, Zoom, ventilator, chloroquine, facial mask, SARS*. The other ranking scores produce the same words (but fewer) in different orders. They provide only two extra words – *respirator* and *syndemic* – that are further down in the list (ranks 191 and 1353, respectively).

In the French Wiktionary, considering or discarding anonymous contributions provides quantitatively similar results, with the set of relevant retrieved words differing according to the score used. The contributors-based score performs better, but the difference with the score based on the revisions is less pronounced than when experimenting with the English Wiktionary. Overall, the proportion of relevant words identified in the 100 best-ranked words is lower in the French Wiktionary. The best configuration (ratio involving median values of contributors and no anonymous contributions), which reaches 9%, is half of that obtained in the English Wiktionary. This configuration made it possible to retrieve the words *confinement, coronavirus, pandémie, chloroquine, cluster, télétravail* 'teleworking', *distanciation, contagiosité* and *présentiel* 'in-person activity'. Other configurations retrieved three additional words: *SRAS* 'SARS', *quatorzaine* 'two weeks quarantine' and *corona*.

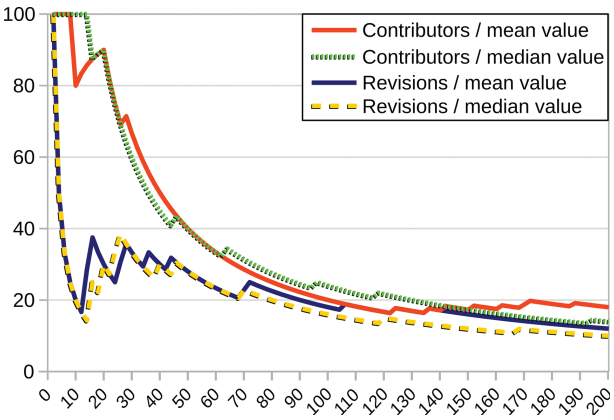


Figure 15: Percentage of Covid-related existing headwords in the English Wiktionary, depending on the ranking scores.

3.5 Quarterly ranking of the existing headwords

For the two languages, discarding the activity of bots improved the results but discarding anonymous contributions barely improved them. Scores based on the number of contributors produced, again, better rankings than those based on revisions, which, sometimes, help detect few extra relevant words. The numbers of relevant existing headwords retrieved from the English and French Wiktionary, as a function of the number of candidates examined, are given in Table 3.

Table 3: Number of identified existing words related to Covid-19, as a function of the number of candidates inspected.

# candidates	English Wiktionary				French Wiktionary			
	T1	T2	T3	T4	T1	T2	T3	T4
5	5	5	0	2	2	1	1	0
10	6	8	0	3	3	2	2	0
25	8	9	2	5	4	4	2	1
50	11	10	2	5	6	8	2	3
100	14	15	3	7	7	9	3	5

Once again, the method achieved better results with the English Wiktionary. For this language edition, the score based on the current/usual contributors ratio (using mean values) performed best. The true positives retrieved by this score applied each trimester to the "no bots, no anonymous" data source are:

- **T1: coronavirus, pandemic, lockdown, corona, Coronavirus, quarantine, Wuhan, rona, herd immunity, panic buying, respirator, superspreader, MERS-CoV, ventilator**
- **T2:** coronavirus, corona, lockdown, **facemask**, rona, **hydroxychloroquine**, **surgical mask**, **Corona**, herd immunity, **pulmonologist**, pandemic, **Zoom**, superspreader, **quar**, MERS-CoV
- **T3:** coronavirus, lockdown, **intensive care unit**
- **T4:** hydroxychloroquine, coronavirus, superspreader, pandemic, **virulent**, **chloroquine**, rona

Varying the data source (i.e. retaining anonymous contributions) retrieved the additional *severe acute respiratory syndrome* (T1) and *viral load* (T3). Changing from mean to median (still with the contributors ratio) retrieved *disinfection* (T1) and *coronary* (T2) while switching to the revisions-based score added *antigen* (T1), *pulmonology* (T2), *immunology* (T3) and *syndemic* (T4).

The same ranking score (contributors ratio, mean value) applied to the same type of revisions ("no bots, no anonymous" option) retrieved the following words from the French Wiktionary:

- **T1: coronavirus**, **confinement**, **chloroquine**, **grippe espagnole** 'Spanish influenza', **pandémie**, **SRAS**, **cluster**
- **T2:** confinement, coronavirus, **vidéoconférence**, **pangolin**, **masque** 'mask', pandémie, **corona**, **skypero** 'online apéritif using Skype', **quatorzaine**
- **T3:** coronavirus, **présentiel**, **jauge** 'capacity'
- **T4: vaccin, antivaccin, vaccinal, distanciation, écouvillon** 'swab'

Changing from mean to median only added *stop and go* (T4), while using the revisions-based score additionally produced *virus* (T1), *télétravail* 'teleworking' (T2) and *infectiosité* 'infectivity' (T4).

The better results observed for the English Wiktionary are related to the greater number of revisions/contributors for some articles. The variation observed for comparable words (e.g. translational equivalents that may have comparable frequencies, degrees of polysemy and specialisation) can simply be due to the number of contributors on the lookout. Another explanation is the possible different degrees of completeness of the articles. For example, the existing headword *comorbidity* ranks high in the second trimester for having been revised several times, as shown in Figure 5(a) (Section 2.3). Although the article was quite up-to-date (the definitions were not modified in 2020), a usage example was replaced by another, more recent, one with an explicit reference to the coronavirus. A pronunciation, an alternative form (*co-morbidity*), synonyms and related words were added, as well as numerous translations. In French, no alternative form exists and the pronunciation was already mentioned in the article before the pandemic. The only revision in 2020 (addition of a recent citation) did not enable the word to be detected.

3.6 False negatives

All the experiments above demonstrated that the proposed method uncovers relevant neologisms, or indicates entries that possibly deserve a review. What the experiments do not say, however, is what the method missed. I therefore examined the ranking of the words included in the two OED updates dedicated to Covid-19 vocabulary. The words added to the OED in April were generally at the top of the list of the headwords detected in Wiktionary, e.g. *Covid-19, social distancing, flatten the curve, Covid, self-isolation, contact tracing, self-quarantine, self-isolate*. Most of the words added in July were already in Wiktionary before 2020. Some of these existing headwords rank quite high, e.g. *corona, surgical mask, MERS, Zoom, dexamethasone, comorbidity*. Others rank much lower down, meaning that the articles received very few revisions, either because they were overlooked by contributors, or

because they were already up-to-date. For example, once *Kawasaki's* is defined as 'synonym of Kawasaki disease', with a hyperlink to the corresponding article, there is not very much left to say. The most noticeable words that the method failed to detect are *shelter-in-place*, added to Wiktionary as a run-on entry under *shelter* in 2020,[22] and words whose ranking scores are very low, e.g. *RO* (dated March 2020), which received 5 revisions made by 3 contributors and *frontliner*, unmodified since its creation in April 2020. Other undetected words or words missing from the nomenclature are those having more popular derivatives (e.g. *contact tracer* ranks relatively low while *contact tracing* ranks high on the list) or equivalents (*community transmission* is absent from Wiktionary, but *community spread* ranks high).

Undetected words are those having too few revisions or contributors. One could speculate that these words are too rare or too specialised to catch contributors' attention. Another variable is Wiktionarians' idiosyncratic contribution patterns. Some contributors make numerous successive revisions as soon as they add a few words, which may result in a large number of revisions made by a single contributor (the coordinate points are those along the x-axis, on the right side of the scatterplots in Figure 9). Others contribute significant editions. For instance, the article *aéroportage* 'air transportation' in the French Wiktionary contains two senses ('transport by air' and 'airborne spread of a disease') with definitions and examples, a pronunciation, inflected forms, a synonym and a related word. The article was created in November 2020 in a one-shot edition and has not been modified since. Located at the (1, 1) coordinate point, it is undetectable. Future experiments will consist in taking into account the nature and length of contributions and possibly lead to a refinement of the method presented here.

I investigated above the presence of the OED Covid-related headwords in Wiktionary. A reverse question is: are Wiktionary's most heavily modified Covid-related headwords in the OED? The top-ranked ones are, except the various stigmatizing appellations *Wuhan virus, Wuhan pneumonia/flu* and *Chinese virus* that were used before the virus and resulting disease were officially named *Covid-19*. Humorous coinages such as *corona belly* or the derogatory *maskhole* may not be good candidates for OED inclusion. Other words such as *syndemic* and, maybe, *doomscrolling*, could be considered. Regarding semantic neology, the 2020 revisions of *antimask* in Wiktionary could draw attention to the possible need to update the OED article which only describes the grotesque dance.

4 Conclusion

The present study was based on the hypothesis that Wiktionary's most heavily modified articles can help detect new and existing headwords that are related to topical

22 Wiktionary's run-on entries were not taken into account in the present study.

events. Experimenting on the 2020 revisions and targeting Covid-related vocabulary proved successful and validated the hypothesis. One finding is that using only the number of unique contributors performs better than relying on the number of editions. In other words, Wiktionary's "crowd" of contributors is an asset for the task at hand. It does not mean that the number of revisions is not relevant. The conclusion to be drawn is rather that, given a set of articles having a similar number of revisions, the articles modified by the greatest diversity of contributors are the most likely to be related to topical events. Varying the ranking scores is also a good idea as it retrieves additional true positives.

Using Wiktionary's revision logs was considered a stopgap when no satisfactory diachronic corpus is available. When such a corpus exists, cross-checking the results of corpus-driven analyses and Wiktionary's history mining is certainly a good option.

A strength of the proposed method is that it is language and topic independent. Regarding language, the method is likely to perform well with the editions of Wiktionary that have the most active online communities. Regarding topics, one has to keep in mind that an event such as the Covid-19 pandemic is extraordinary, as were the two OED updates – and that *unprecedented* was the Oxford Languages word of the year 2020. Whether the suggested method is able to detect lexical innovations related to topical events that are less massive is an open question and the subject of future analyses. Trawling through the lists of candidates for the present study made me confident on that point. Other topics emerged, related for example to the US presidential election, identity and discrimination questions, police brutality – 2020 was also the year of the killing of George Floyd that brought the (pre-existing) Black Lives Matter movement to the forefront, with the BLM acronym ranking high in Wiktionary in the second trimester. Similar topics emerge from the French Wiktionary. In this language edition, a large number of feminine agent nouns were added. Though not precisely related to a timestamped event, and even though most of these nouns are feminized job titles related to forgotten professions, this trend is noteworthy.

Wiktionary revision logs give the opportunity to predict the past. A possible assessment of the suggested method consists in reiterating the experiments on the revisions of previous years and analysing what vocabulary emerges, related to which topic. In the meantime, the current study led me to examine the revision rate of *quarantine*, for which I observed a jump in 2020, and another back in 2009 (cf. Figure 16). Calculating the most frequently modified articles for that year enabled the detection of *swine flu*, which ranked first among the new articles (eclipsing the equally new *H1N1*), while, regarding existing headwords, *epidemic* ranked 111th globally and 16th in the second trimester; *quarantine* ranked 141st globally and 7th in the second trimester; and *mask* ranked 307th globally and 142nd in the third trimester. Hopefully, the future will allow for the detection of more enjoyable neologisms to be included in the dictionary. The present is apparently not a time for complacency. In the French Wiktionary, *vaxxie* 'a selfie taken while getting a Covid-19 vaccine', *centre de vaccination*

'vaccination centre', *Covid long* 'long Covid', *passeport vaccinal* 'vaccine pass' and *vaccino-sceptique* 'sceptical about the usefulness or the efficiency of vaccines' are among the most frequently modified neologisms during the first trimester of 2021 (with respective ranks of 4, 9, 17, 63 and 118). Regarding existing words, *vaccinodrome* 'large capacity vaccination centre' that was coined in 2020 and entered Wiktionary in March 2020 was not revised enough to be detected that year, but ranks 24th in the first semester of 2021, while *couvre-feu* 'curfew' ranks 15th in the second trimester. Which is a good point in favour of the suggested method, if not a lighthearted final note.

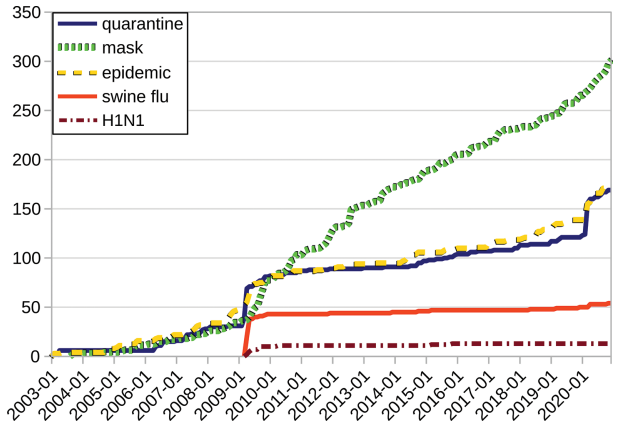


Figure 16: Neologisms and existing words showing notable increases in revisions in 2009.

Bibliography

Cook, Paul/Lau, Jey Han/Rundell, Michael/McCarthy, Diana/Baldwin, Timothy (2013): A lexicographic appraisal of an automatic approach for detecting new wordsenses. In: *Proceedings of the eLex 2013* conference, Tallinn, Estonia. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut, 49–98.

Fišer, Darja/Ljubešić, Nikola (2018): Distributional modelling for semantic shift detection. In: *International Journal of Lexicography* 32(2), 163–183.

Hamilton, William L./Leskovec, Jure/Jurafsky, Dan (2016): Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In: *Proceedings of the Conference on Empirical Methods in Natural Language* Processing *(EMNLP'2016)*, Austin, Texas. Association for Computational Linguistics, 2116–2121.

Hanks, Patrick (2012): Corpus Evidence and Electronic Lexicography. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, 57–82.

Kilgarriff, Adam (1998): The hard parts of lexicography. *International Journal of Lexicography* 11(1), 51–54.

Kilgarriff, Adam (2009): Simple maths for keywords. In: Mahlberg, Michaela/González-Díaz, Victorina/Smith, Catherine (eds.): *Proceedings of Corpus Linguistics Conference*, Liverpool, UK. University of Liverpool.

Kutuzov, Andrey/Øvrelid, Lilja/Szymanski, Terrence/Velldal, Erik (2018): Diachronic word embeddings and semantic shifts: a survey. In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico. Association for Computational Linguistics, 1384–1397.

Landau, Sidney (2001): *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Lih, Andrew (2004): Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. In: *Proceedings of the 5th International Symposium on Online Journalism*, Austin, Texas.

Renouf, Antoinette (2013): A finer definition of neology in English: The life-cycle of a word. In Hasselgård, Hilde/Ebeling, Jarle/Ebeling, Signe Oksefjell (eds.): *Corpus Perspectives on Patterns of Lexis*. Amsterdam/Philadalepia: John Benjamins, 177–208.

Rundell, Michael (2017): Dictionaries and crowdsourcing, wikis, and user-generated content. In: Hanks, Patrick/de Schryver, Gilles-Maurice (eds.): *International Handbook of Modern Lexis and Lexicography*. Berlin/Heidelberg: Springer, 1–16.

Sajous, Franck/Josselin-Leray, Amélie/Hathout, Nabil (2018): The Complementarity of Crowdsourced Dictionaries and Professional Dictionaries viewed through the Filter of Neology. In: *Lexis* 12.

Sajous, Franck/Calderone, Basilio/Hathout, Nabil (2020): ENGLAWI: From Human- to Machine-Readable Wiktionary. In: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, 3016–3026.

Wolfer, Sascha/Müller-Spitzer, Carolin (2016): How many people constitute a crowd and what do they do? Quantitative analyses of revisions in the English and German Wiktionary editions. In: *Lexicos* 26, 347–371.