Rook – A New Tool for Visualising Word Frequency Changes

Niklas Zechner
Språkbanken Text, University of Gothenburg
niklas.zechner@gu.se

Abstract

A common task in corpus linguistics and digital humanities is finding out how the frequency of a word has changed over time. There are many applications – lexicography, sociolinguistics, historical studies, teaching, etc. Several tools already exist which can do this for Swedish, but they may be slow, complex, or based on unsuitable data. We present a fast and accessible tool for this specific task.

1 Introduction

Finding word frequencies was one of the earliest tasks in corpus linguistics, and it remains an important issue today. It can be used to decide which words are common enough to enter a dictionary, to choose a core vocabulary for language learners, to find the most common spelling for a word, or as a point of comparison for characterising a specific text based on its specific word frequencies, among many other things. If we can also see how those frequencies have changed over time, it opens up further avenues for research. We can track historical language changes in themselves, look at social changes and events through which words were used, estimate when a word became established, or provide a baseline for identifying the age of unknown texts.

There are already tools that can do this, but they are often complex and inaccessible to the general public. To our knowledge, none have the simplicity and scope of the tool presented here. Particularly for users who are not themselves computational linguists – whether they be researchers in other fields, students, journalists, or the general public – there is a need for a tool which fills certain requirements:

- Simple and fast. The user should only need to type in a word, and get an immediate result.
- Providing clear results. The output should be in a format which is easy to read for a person with no scientific background, and there should also be machine-readable output for those who need it.
- Able to compare different words. In many cases, the relevant question is the difference between two words, so it is important that we can easily see both at the same time.
- Based on a large and uniform corpus. There should be as much data as possible for each time slot, and the data should come from similar sources.

The last point may be somewhat controversial. In many cases, researchers strive for a balanced corpus, one which contains texts of many different types, to give a more representative view of the language. This would have some advantages, for example when looking for the first occurrences of a word or spelling change. But at least in this case, we argue that a uniform corpus is preferable. Firstly, in a corpus where different sources dominate at different times, we can never be sure that a change in frequency is due to an actual change in language use, and not a change in the balance of sources. By using the same or similar sources, we can more confidently draw conclusions about change. Secondly, even the most balanced corpus cannot be claimed to represent "the true language", only an arbitrary subset of it. By keeping to a particular type of texts, in this case newspapers, we can at least with some authority claim that a change has happened in that specific category of language.

With this in mind, we have created Rook, a simple web-based tool for looking up and visualising word frequencies and how they have changed over time. It uses a fixed database of pre-calculated word frequencies, based on newspaper data from 1850 forward, and shows the result as a graph. The tool is available at spraakbanken.gu.se/prototyp/rook.

2 Interface

Basic usage of the Rook interface is straightforward: Enter one or more words, press the button, and a graph is shown. You can set limits to which years are shown, and choose to show a table with the underlying data.

Since a curve might have a lot of data points, and there can be more than one curve, the result may sometimes look cluttered. For this reason, we can choose to use a smoothing algorithm: block, sliding, or gaussian. "Block" simply combines the points up to a given width – by default, 12 months – and shows them as one point, with the average of the frequencies. "Sliding" takes the sliding average for each month – that is, with a smoothing width of 12 months, the average for the 12 months around the given point. "Gaussian" uses a one-dimensional Gaussian blur, also known as Weierstrass transform – that is, each point is replaced by a normal distribution curve with the given width as standard deviation, and then these are summed to create the new curve. This acts as a low-pass filter, creating a smooth curve that ignores all changes over shorter periods than approximately the smoothing width.

Since the data only consists of stored word frequencies, it is not possible to perform any further analysis based on the text itself – parsing, sense disambiguation, and so on. That also means that you cannot search for a lemma, only the word as written. The search is not case-sensitive.

You can combine two or more words in a search using a plus sign. For example, if we want to search for all forms of *hatt* ('hat'), we can type *hatt+hattar+hatten+hattarna*.

A perhaps surprising discovery is that many words differ in frequency based on the time of year. It is not unexpected that a word like *sommar* ('summer') is more common in the summer than in the winter, but we also find differences in common words which would not seem to be seasonal; for example, *från* ('from') is also noticeably higher in the summer. In order to study this phenomenon, the interface has an option to make the graph periodic. In the periodic graph, we see a curve from January to December, showing the average over years of the frequency for each month. The individual values are shown as disconnected dots.

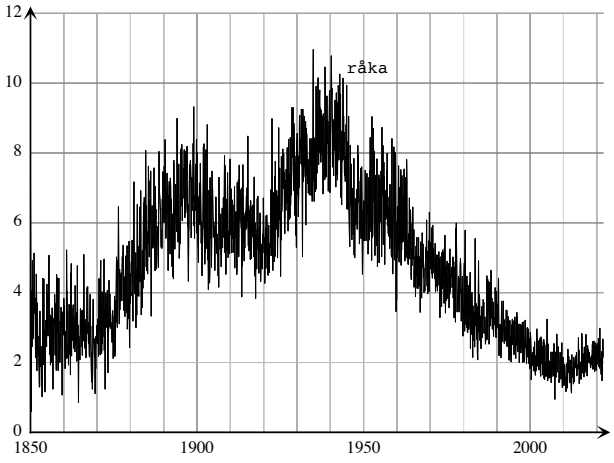

Figure 1: Search result for *råka* ('rook (bird)' or 'happen to'), without smoothing. Y axis shows occurrences per million words.

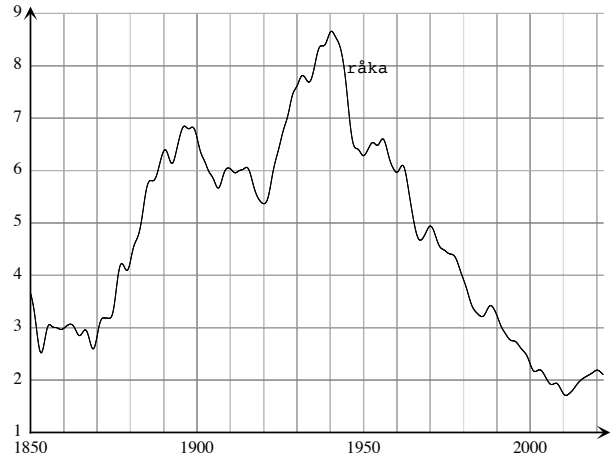

Figure 2: Search result for *råka*, with smoothing (Gaussian, width 12).

Finally, additional parameters for modifying the appearance of the graph can be specified as part of the search string, for example for changing the line width or colour, adding labels, and many other things. These have not been added as separate fields to avoid cluttering the interface. More information about them can be found in the manual.

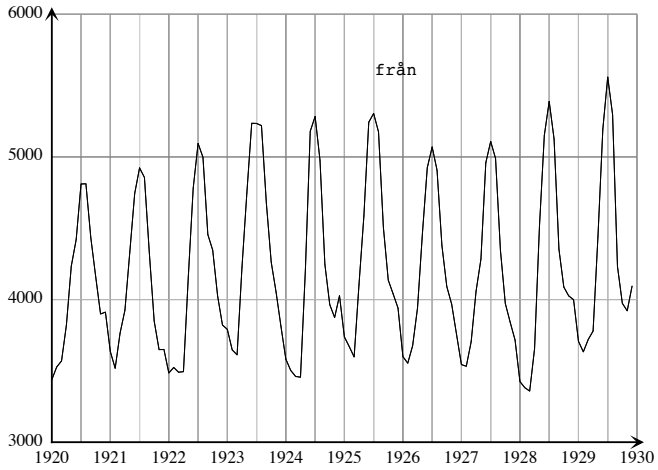

Figure 3: Search result for *från* ('from'), 1920s.

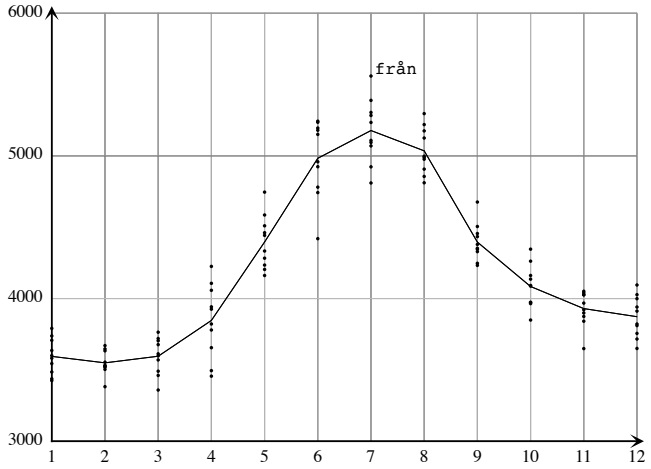

Figure 4: Search result for *från*, 1920s, periodic. X axis shows months (so 1 is January, etc.).

3 Data

The underlying text data used in Rook comes from the digitised newspaper collection of the National Library of Sweden (KB), probably the largest uniform digital corpus of Swedish. This material is already partially available to the public through Svenska dagstidningar (tidningar.kb.se). Here, you can look up a word and see it in context from newspapers. Newspapers older than 115 years can be viewed in their entirety, whereas for newer texts, only a small context is shown due to copyright. This website also shows you a graph over how many hits there were for each year, but because of the very uneven distribution of data, this does not tell us about changing word frequencies (and it counts the number of pages containing the word, not occurrences).

We have used all available data from 1850 up to and including 2021 (so far). The limit 1850 was set because the available material before 1850 is smaller, and because this was around the time the newspapers changed from blackletter to modern fonts, greatly improving the OCR quality. The data contains over 50 000 000 000 tokens.

The amount of data varies greatly over the currently 172 years. From 1850 to 1907, it increases from about 5 to 50 million words per month. This includes most of the published newspapers from that time. From 1907, only a fraction of the published newspapers are included, and the data per month varies between roughly 10 and 35 million words, up until 2013. After that there is a large increase: 2013 has 80 million words per month on average. From 2014, all published newspaper data is included, which gives an average around 200 million words per month and a minimum of 150, except for the last two months of 2021.

Since we calculate (relative) frequencies, the amount of data should not directly impact the results, but drastic shifts in which publications are included may still have an effect. We also see some indications that the number of OCR errors differs between periods.

4 Implementation

From the data received from KB, we have extracted the word frequencies for each month. By only keeping the frequencies, we avoid any copyright restrictions, and reduce the size of the database and the time to search it. We have for each month discarded those words which occur only once; looking at a sample of them shows that they are almost all OCR errors, and if we did not remove them, they would make up 95% of the database. An advantage of doing the cutoff by month instead of globally is that we can then employ the same rule for data added in the future, without needing to redo the whole process. The words are separated by a hash code, and then stored in a table format with their frequencies for each month. Although lists based on time would have been easier to create and update, this makes the lookup faster, since we are usually looking for few words and many months. When doing a search, each word is looked up by its hash code, and values for the appropriate months are read from the data file. The coordinates for each data point,

along with any additional parameters, are passed to our graphing program. It calculates the physical coordinates, applies smoothing as needed, chooses the limits of the coordinates, and outputs the resulting graph as SVG code, which is shown as an image on the webpage.

5 Future work

In future versions, we hope to add access to bigrams (sequences of two words). This will allow us to look for words used together, fixed expressions, persons, and more. Longer sequences may be difficult to implement due to the large number of possible combinations, as well as copyright limitations. We are also considering alternative corpora, including some which have been pre-processed to include lemmatisation and other forms of analysis. There is currently an ongoing project at KB and Språkbanken to analyse the same newspaper data we have used here.

6 Conclusion

The tool presented here provides a way for researchers as well as the general public to look up word frequencies and frequency changes. It is easy to use, and gives fast and clear results. The uniformity of the corpus means that we can make well-defined statements and avoid interference. We can search for multiple words, either adding the results or comparing them, making it easy to see when one word has gone out of fashion and another taken over. We hope that this will be helpful both for those in science studying language and history, and those who are generally curious about how the Swedish language has changed over time.