6 The Word Rain visualisation technique applied to digital history: How to visualise, explore and compare texts using semantically structured word clouds

Maria Skeppstedt
Stockholm University

Magnus Ahltorp
Nakajima Koen Research Institute

Kostiantyn Kucher
Linköping University

Gijs Aangenendt
Uppsala University

Matts Lindström
Uppsala University

Ylva Söderfeldt
Uppsala University

The Word Rain text visualisation technique aims to retain the simplicity of the classic word cloud, while addressing some of its limitations. In particular, the Word Rain visualisation uses word embeddings to automatically give the visualised words a semantically meaningful position along the horizontal axis. In this handbook chapter, we showcase how this novel approach for word positioning makes the Word Rain technique suitable for exploring, analysing and comparing texts. More specifically, we show how the Word Rain Python module can be used to visualise longitudinal changes in periodicals published by the Swedish Diabetes Association, and how the Word Rain web service can be used to create visualisations that compare the patient organisation periodicals to journals published by the Swedish Medical Association.

1 *Introduction*

The classic word cloud[1] is a popular technique for providing a quick visual overview of the most important content of a text or text collection. The

1 The terms "word cloud" and "tag cloud" are often used interchangeably, but we will here consistently use the term "word cloud", since the main application scenario targeted is the visualisation of words automatically extracted from a running text, and not the visualisation of tags (Torres Parejo et al. 2021).

HUM INFRA Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, Gijs Aangenendt, Matts Lindström & Ylva Söderfeldt. 2025. The Word Rain visualisation technique applied to digital history: How to visualise, explore and compare texts using semantically structured word clouds. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 147–182. University of Tartu Library. DOI: 10.58009/aere-perennius0175

technique is built on the idea of extracting the most frequent words from the text and displaying them in a graph with a font size that corresponds to their frequency (Cao & Cui 2016: 57–59).

While the result is a visually more appealing—and more spatially dense—alternative to a simple sorted word frequency list, it does not provide any additional analytical power. Neither the frequency list, nor the word cloud, assists you in the task of semantically sorting the most prominent words, or in the task of comparing the most prominent words in two different word clouds/word frequency lists. The layout of the word cloud makes it possible to display a large number of words in a small area by giving the less prominent ones a font size too small to read without zooming in. However, the technique does not provide any guidance to where in the cloud to zoom in to reach potentially interesting words, and therefore does not provide any advantage compared to reading a long word-frequency list.

There are many different types of more analytically powerful developments of the classic word cloud that build on interaction and/or animation (Collins et al. 2009, Koh et al. 2010, Lohmann et al. 2012, Liu et al. 2015, Wang et al. 2018, Xie et al. 2024), and, of course, even more examples of analytically powerful text visualisation techniques in general (Kucher & Kerren 2015), many of them relying on a dynamic and interactive user interface. None of them have, however, been close to achieving the same general level of popularity as the static word cloud.

We believe that one of the key factors for the popularity of the word cloud is the elegance that comes with simplicity. This hypothesis is, for instance, supported by the classic word cloud being recommended (in an information visualisation conference paper) as the first and most important text visualisation technique to teach within digital humanities (Jänicke 2022), as well as by the characterisation of the word cloud by Cao & Cui (2016): "There are reasons for word clouds being user friendly in general. Unlike other visualizations, word clouds need no additional visual artifacts and are self-explanatory. Users read the words and compare their sizes to know which one is more salient." We, therefore, set out to develop a text visualisation technique that—as far as possible—would retain this simplicity of the static word cloud, while also providing functionality that would be practically useful for analysing and comparing the content of text collections. This work resulted in the Word Rain[2] text visualisation technique.

We have previously introduced, explained and evaluated the Word Rain technique (Skeppstedt et al. 2024), and we have also exemplified how the technique can be used practically. We have previously focused on three

2 https://github.com/CDHUppsala/word-rain (last accessed: February 14, 2025)

different use cases: i) to compare texts from different text genres (Skeppstedt et al. 2024), ii) to study longitudinal corpora changes (Skeppstedt & Aangenendt 2024), and iii) to study the content of a text in relation to a dictionary/controlled vocabulary (Ahltorp et al. 2025). In this handbook chapter, we will further explore the first two of these use cases, by also providing concrete instructions—together with links to programming code—for how to generate and analyse the word rain visualisations. We will also provide a more advanced example, by adding a third use case, also for the task of exploring longitudinal corpora change. Each use case has an increasing level of complexity, both when it comes to generating the word rains and to interpreting the results. For the first case, we will use the Word Rain web service. This option provides less flexibility, but instead offers the large advantage of not requiring any programming knowledge at all. Instead, the texts are simply uploaded to the website in order to generate word rains. We will, thereafter, describe a use case that requires a minimum amount of programming code, and then the third one which requires more programming skills.

The chapter is organised as follows: We start by briefly discussing limitations associated with the classic word cloud, and we thereafter describe how these limitations are addressed by the Word Rain technique. In the section that then follows, we describe the concrete example use cases. We provide a short description of the corpora used in the examples, and then—for each example—we describe how the word rains are generated and how to read the resulting word rain graphs. Thereafter, we contrast the Word Rain visualisation technique with other similar techniques, by comparing it to a word frequency list and to the classic word cloud, as well as to previous efforts for improving the classic word cloud. The chapter is then concluded by a short summary.

2 *Limitations of the classic word cloud*

There are many versions of the classic word cloud. For instance, the words in the cloud could either be displayed in a random order or—with a small development of the algorithm—be positioned in alphabetical order (Cao & Cui 2016: 57–59). Instead of just displaying the most frequent *words*, it is also possible to include frequent n-grams (i.e., short word sequences) in the visualisation (Cao & Cui 2016: 43). Similarly, raw frequency is a standard measure used for determining prominence/font size, but there are also other options, such as TF-IDF, i.e., term frequency–inverse document frequency (Xu et al. 2016).

However, there are limitations associated with all these standard versions of the word cloud which decrease the usefulness of the word clouds for many analytical tasks. This is problematic, since the word cloud is a frequently used text visualisation technique (Brath & Banissi 2016).[3] Word clouds are not only used to provide a visually appealing illustration, but are also used for analytical purposes. For instance, Hicke et al. (2022) studied the purpose of including word clouds in 351 news articles and academic articles within digital humanities, and found that in 37% of the articles studied, the purpose of the word cloud was to enable an "analytic exploration".

When empirically comparing the efficiency of different standard design options for word cloud-like representations, it has been shown that the optimal design choice can be different depending on which analytical task is to be carried out (Felix et al. 2018). Design trade-offs can thereby be expected when aiming to support too many tasks. We have therefore mainly focused our attention on a few common analytical tasks. Hicke et al. (2022) provide a taxonomy for describing such tasks where word clouds are used today. From this taxonomy, we focused on the tasks of *finding the most prominent words* in a text, *finding "topics" or semantic word clusters*, and *gist-forming* (i.e. getting the gist of a text). Hicke et al. also found that users of word clouds were often expected to perform *inter-cloud comparison tasks*, i.e., tasks that included a comparison of two, or several, word clouds.

For performing these four tasks, we have identified three important limitations associated with the classic word cloud:

1. The first limitation regards the lack of a meaningful interpretation of how the words are positioned in the word cloud (Barth et al. 2014). The lack of a meaningful positioning might result in the reader incorrectly inferring a meaning to where the words are positioned in the graph. In addition, eye tracking studies have shown that a meaningful placement of the words increases the efficiency in how the graphs are used. More specifically, for finding prominent words, a position based on word prominence is most effective.[4] A thematically clustered layout was, instead, most effective for the aim of finding words belonging to a certain topic (Lohmann et al. 2009). Similarly, for an analytical task-based evaluation, word clouds that grouped words into semantic categories—either separated by different colours, or separated by space—made it easier to solve the task (Hearst et al. 2020). This possibility to employ

3 Among 106 peer reviewed text visualisations investigated, 39 were versions of the word cloud (Brath & Banissi 2016).

4 In the respective study, a circular layout with decreasing word prominence towards the edges was used.

word positioning to support analytical tasks is, however, not used in a classic word cloud.

2. Another limitation related to the lack of a semantically meaningful word positioning, is that the word cloud technique does not provide any guidance regarding where in the cloud to zoom in. This makes it very difficult to explore words rendered in a small font size. The use of a small font size even affects readability negatively when the font size is large enough to be *possible* to read without zooming in (Ren et al. 2024). Thereby, in practice only the most prominent words in the cloud can be used for forming a gist of the text content.

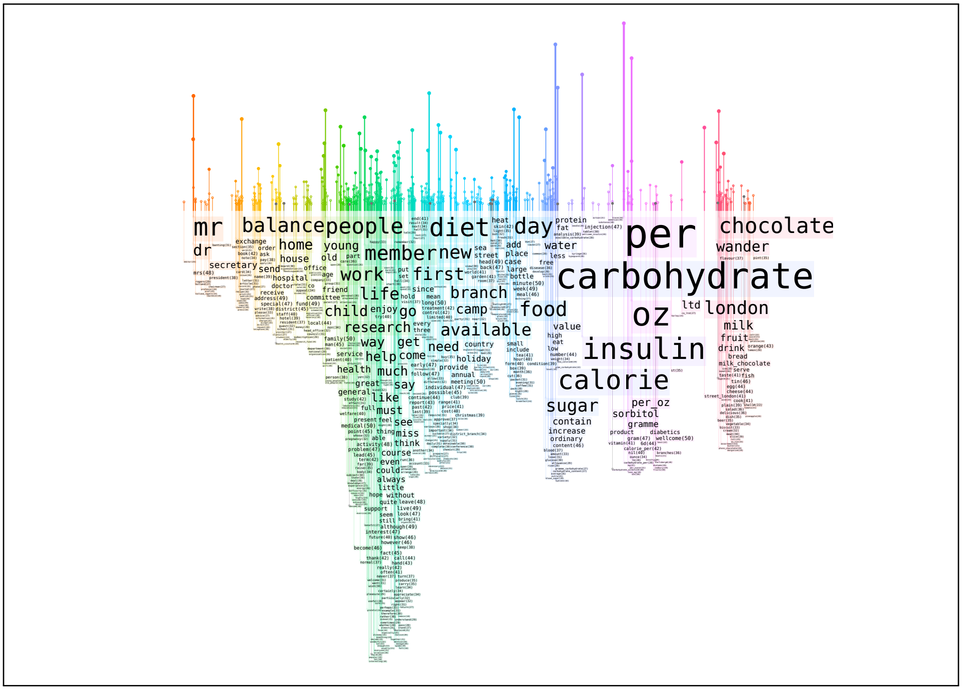
3. The third limitation regards the difficulty to compare word clouds, e.g., when tracking longitudinal changes in a series of word clouds or comparing two word clouds representing two text genres. As expressed by Cao & Cui (2016: 74): "[...] tracking words between different word clouds is the biggest pain for this type of methods". The main reason is the varying position of the same word between different word clouds (Cao & Cui 2016: 69).

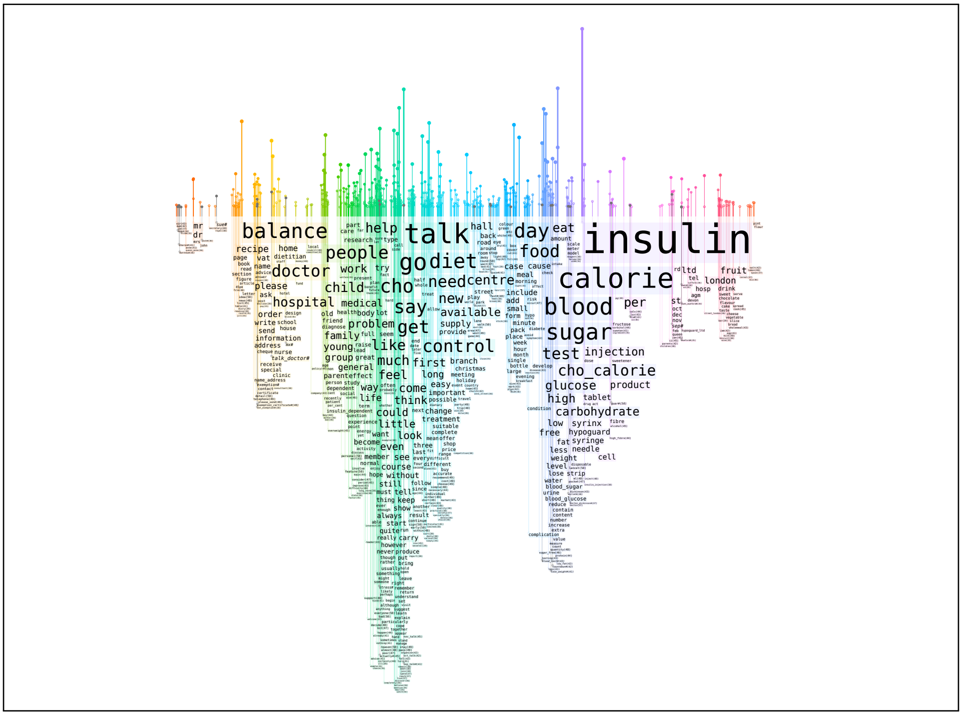
Critique has also been directed towards the use of font size as a prominence indicator, since this might result in longer words incorrectly being perceived as more important, as they occupy a larger area in the graph (Viégas & Wattenberg 2008). However, a large font size is a self-explanatory signal for prominence and the use of a small font size enables the creation of dense graphs with many words. We will, therefore, address this limitation by alleviating it, not completely removing font size as a prominence indicator. In addition, results from Alexander et al. (2018)—where an accuracy of over 90% was achieved for a font size comparison task—show that there are situations in which users are able to compensate for length differences.

While there are previous extensions of the classic word cloud that aim to solve *one* of the limitations listed above (see Section 5.2), we are not aware of any text visualisation technique that—as the Word Rain technique—addresses all the described limitations.

3 *The Word Rain text visualisation technique*

The aim of the Word Rain visualisation remains the same as that of the classic word cloud, i.e., to create a static visualisation that quickly provides an understanding of text content by displaying the most prominent words in a text. However, by creating the Word Rain visualisation technique, we address important limitations associated with classic word cloud, thereby


mr
dr
balance
people
diet
day
per
chocolate
home
young
member
new
water
wander
secretary
work
first
carbohydrate
life
branch
oz
london
child
go
food
research
available
milk
way
get
need
insulin
help
come
health
much
calorie
say
like
must
sugar
sorbitol
a. *1965*


mr
dr
balance
help
talk
day
insulin
people
food
doctor
go
diet
calorie
cho
need
centre
hospital
say
blood
per
problem
get
sugar
like
control
test
injection
much
first
cho_calorie
feel
glucose
come
high
think
carbohydrate
b. *1985*

◀ Figure 1: The top 600 most frequent words in patient periodicals from the British Diabetes Association.

making the visualisation more practically usable for analysing, exploring and comparing texts.

3.1 *The word positioning of the Word Rain technique*

The main method for achieving the additional analytical capability of the Word Rain visualisation is the way in which the words are positioned.

Particularly important is the feature of letting the *horizontal position* of a word be determined by the *meaning of the word*. Before the algorithm for positioning the words can be applied, each word to be visualised in a series of word rain graphs is assigned a fixed value on the horizontal axis (i.e., a fixed value on the x-axis), which represents its meaning. In the currently available implementation, this is achieved by projecting word embeddings—more specifically, multidimensional word-vectors from a word2vec-model—onto one dimension. Thereby, a word is assigned the point on the x-axis that is determined by projecting its word vector from a multi-dimensional space onto the one-dimensional x-axis.

This horizontal word positioning has the consequence that words with a similar meaning are positioned close to each other on the x-axis. Note that the x-axis does not represent a scale of meaning—it would be impossible to represent the general meaning of words on a scale. Instead, the word positioning could rather be viewed as depicting loosely defined clusters of meaning, and words with similar meaning—i.e., those belonging to the same semantic cluster—are positioned close to each other on the x-axis. For instance, in the upper graph in Figure 1, to the right (in violet), there are a number of words related to measurements, e.g., *oz*, *per*, *gramme*, and a bit to the left (in blue) there are many words related to *carbohydrate* and *sugar*.

When several word rains are generated as a series of graphs, as is the case for the two graphs in Figure 1, a word receives the same horizontal position in all word rains generated. That is, one and the same projection of word embeddings on the x-axis is used for every graph in a series. The words *insulin*, *calorie* and *balance*, which are prominent in both graphs, provide evident examples of this consistent word positioning.

While the horizontal position for a word is the same for all graphs included in a visualisation series, the vertical position (i.e., position on the y-axis) is not. The *vertical position* indicates *word prominence* in the text visualised, and it also adapts to the position of the other words in the graph.

The Word Rain algorithm positions the words at their pre-determined x-coordinate in a decreasing order of prominence. The algorithm starts by trying to position a word at the very top of the graph. If the extension of the word overlaps with that of a more prominent one, the word has to yield, and "rains down" in the graph until there is a free spot. In general, more prominent words are therefore positioned above less prominent ones. "Word prominence" can have different meanings (as will be exemplified by the use cases), but in Figure 1, it is equivalent to word frequency in the text visualised.

In the first graph in Figure 1, *per* is the most prominent word, and is therefore positioned first. The horizontal extension of the second most prominent word *carbohydrate* overlaps with *per*, and *carbohydrate* is therefore moved downwards until there is no longer an overlap. In contrast, the word *protein* does not overlap with a more prominent one, and is therefore positioned at the very top of the graph. A more compact graph can thereby be created by allowing the vertical position to not strictly follow word prominence.

The word positioning algorithm often results in vertical clusters of semantically similar words being created, where the more prominent ones generally (but not always) are positioned above the less prominent ones. The assigned x-coordinate of a word is the position where the word starts, i.e., in the example here, the point just to the left of the word.[5] In the first graph in Figure 1, a vertical cluster is, for instance, formed around the x-coordinate where the word *carbohydrate* starts. At almost the same x-coordinate, other words related to nutrition also start, i.e., the words *calorie* (below), and *fat* (above). When looking at a slightly larger portion of the x-axis, other related words can be found, such as *protein*, *eat*, and words related to the level of sugar (in blood/urine).

3.2 *The pinhead bars*

Similar to a classic word cloud, the font size used for the words correspond to the word's prominence. Above the area of the graph containing the words, there is a bar chart, consisting of a bar associated with each word in the graph. Their height also corresponds to prominence values of the words. These bars provide a prominence indication that is independent of the word lengths (thereby partly addressing the above-mentioned criticism of using font size as the sole prominence indicator). The pinhead at the top of each bar makes it easier to see where the bar ends.

5 For right-to-left languages, e.g., Yiddish, Arabic, Hebrew, this can be changed to the upper-right corner by a configuration option.

The bars also provide the additional functionality of making word clusters, as well as absences of words in certain semantic regions, more evident. For instance, when looking at the words in the right-hand side of the second graph in Figure 1, they might give the impression of populating the entire area. However, when looking at the bars, it is evident that there is a gap between the violet and the pink/reddish areas.

The graphs in Figure 1 show a colour configuration, where a gradient along the horizontal axis is formed by a wide spectrum of colours. It is, however, possible to configure the bars to be displayed using another colour scheme, for instance, to use a monochrome bar chart or to use fewer colours, as will be shown below. It is also possible to configure the word rain graphs to emphasise certain words, either by providing a pre-defined set of words to emphasise, or to configure the word rains to emphasise words occurring for the first time in a visualisation series (see use cases below). This is reflected in the bar chart by a grey colour. In the upper graph in Figure 1, there is for instance a cluster of emphasised words (e.g., *sunlight* and *exposure*) among the blue bars.

Each bar is connected to its corresponding word by a thin line. These thin vertical lines are drawn from the bottom of the bar to the starting position of the word with which the bar is connected. The length of these lines does not carry any meaning.

3.3 *Determining the horizontal word positions*

So far, we have used expressions such as "similar meaning" and "semantically close" rather carelessly. What we formally mean is the concept which is referred to as *paradigmatically related words* within distributional semantics, i.e., words that typically can be exchanged for each other in a sentence (Cuba Gyllensten 2023). Paradigmatically related words thus often appear in the same semantic contexts in a text. The words *fruit* and *bread* in the cluster of words to the very right (in red) in Figure 1 forms one example of paradigmatically related words. Both of them would, e.g., fit in a sentence such as "I eat a lot of (bread/fruit)".

To use a word2vec model is one, among many, methods to encode which words are paradigmatically related, and this is the method supported in the current implementation of the Word Rain algorithm. There are also many different methods available for reducing a multi-dimensional space to fewer dimensions. The default dimensionality reduction method supported in the current implementation of the Word Rain algorithm is t-SNE projection (van der Maaten & Hinton 2008).

When generating a word rain, it is either possible to use a pre-trained word2vec model, or to train a model on the same corpus that is visualised. The advantage of the latter option is that the word similarities in the word rain graphs then reflect paradigmatic similarities in the actual texts that are to be analysed. However, when the text corpus used for creating the word2vec model is too small, the quality of the model might not be high enough to be useful for the task of semantically sorting the words to visualise.

Also when the corpus used for training the word2vec model is relatively large, there might be words receiving a horizontal position in the word rain graphs that does not necessarily align with a human perception of paradigmatic similarity. This could, for instance, be the case for words that occur infrequently in the corpus used for training the model, or for words that occur in unexpected contexts in this corpus. For instance, the red food cluster in the first graph in Figure 1 contains the word *wander*, which does not seem to belong in this context.

It should also be noted that the effect of projecting the multi-dimensional information of a word2vec model onto just one dimension will result in information loss compared to using the original model. For the purposes of the Word Rain visualisation, it is enough to retain the information necessary for positioning similar words in close proximity, in order to let the user explore semantic neighbours or to identify clusters of similar words. However, in order to more closely study paradigmatic similarity distances—a task that is normally out of scope of the Word Rain visualisation—a method retaining more of the original multi-dimensional model is likely to be more suitable. Useful observations about paradigmatic similarity can, however, still be made based on the horizontal word positioning in a word rain, for instance observations that could form hypotheses to use for exploring semantic similarity in the original word2vec model. To summarise: the main aim of the Word Rain visualisation is *not* to convey detailed information on paradigmatic word similarity. Instead, the visualisation *employs* paradigmatic similarities in order to efficiently convey the most prominent words in a text.

4 *Use cases*

We present three use cases with an increasing level of implementation complexity. In the first one, the Word Rain web service is used and, thereby, no programming at all is needed. For the second use case, a minimum level of programming is required. The code for this use case does not contain any advanced programming logic, but is merely a way of setting different

parameters for the word rain. The third example, in contrast, shows how more complex configurations can be applied in order to tweak the word rain functionality. Also here, only a minimum amount of programming code is required, but the user needs to have some experience in applying programming logic. By offering—and here describing—different methods for generating word rains, we aim to target different kinds of users, both users who prefer an easy method for generating standard word rains without any need for coding, and users who want the flexibility to create word rains that fit specialised visualisation goals.[6]

4.1 *Corpora used for the experiments*

As example texts for describing the Word Rain technique, and for the three use cases, we used text corpora investigated in two projects connected to the Centre for Digital Humanities and Social Sciences Uppsala, CDHU, namely, the ActDisease and Swemper projects.

ActDisease[7] is a research project that studies the emergence, structure, and significance of diagnosis-specific patient organisations. The aim is to determine how certain diseases became subject to organisational efforts by people affected by them, and how this forming of organisations affected the understanding and management of those diseases. Around ten patient organisations from four European countries (Sweden, UK, France and Germany) are part of the study. As part of the project, their periodical publications (member newsletters and magazines) have been digitised, making up a corpus of approximately 125 000 pages from about 1890 to 1990 (Aangenendt et al. 2024). The project uses a combination of digital and traditional historiographic methods, with the digital text analysis aiming to provide insight into longitudinal changes in the texts, which helps guide the researchers in selecting case studies for a closer study of the texts and additional, archival sources. Furthermore, the project explores how to use digital text analysis methods to reveal layers of the text that are not readily visible to the human reader (Jänicke et al. 2017).

We here use two of the corpora digitised within the ActDisease project. For the use cases, periodicals issued by the patient organisation "The Swedish Diabetes Association" (Söderfeldt 2025) are used. This corpus consists of 8 891 pages, written in Swedish and published from 1949 to 1990 (except in 1950 and 1951). For the examples above, when explaining the Word Rain technique, we use periodicals from "The British Diabetes Association".

6 The code for generating the word rains showcased here is available at https://github.com/CDHUppsala/word-rain/tree/main/handbook (last accessed: February 14, 2025)

7 https://www.actdisease.org (last accessed: February 14, 2025)

The Swemper project[8] is a digital infrastructure currently under development at Uppsala University that aims to create a historical full-text database of Swedish medical periodicals produced during 1781–2011. By the end of the project around 550 000 pages (28 titles, ca. 1 100 physical volumes) will have been scanned, digitised and collected in a database. One of the goals of the project is to make this database available through a web interface (The Swedish Medical History Portal), which is also under development by the project. Another project goal is to explore and disseminate how state-of-the-art machine learning techniques can be used to improve the quality of layout analysis and OCR, and thereby provide better metadata to facilitate search and browsing of both images and text-data in the large database. The machine learning techniques used are described in Chapter 8 of this handbook (Ortiz Pablo et al. 2025).

Also from the corpora digitised in Swemper, we used one of the sub-corpora: "Läkartidningen". This is a Swedish medical journal published by the Swedish Medical Association. We here focus on content from the 1960s that had been digitalised and was available when this study was carried out.

For the first use case, the texts from the Diabetes periodical and from Läkartidningen were used as as-is, i.e., as provided from the digitisation process described by Aangenendt et al. (2024) and Ortiz Pablo et al. (2025), respectively. For use case 2 and 3, we used texts from the Diabetes periodical which had been lemmatised using Efselab (Östling 2018).[9]

4.2 *Comparing two text genres using the Word Rain web service*

The task for the first use case consists of creating an overview of—and comparison of—two corpora from two medical genres, both from the 1960s. The first corpus consists of periodicals from the Swedish Diabetes Association, and the second consists of texts from the journal Läkartidningen, as described above.

We created one text file consisting of pages from periodicals from the Swedish Diabetes Association published during the 1960s, and another text file consisting of the pages from Läkartidningen from the 1960s that so far have been digitised within the Swemper project. From the periodicals, we used all texts from the 1960s (2 239 pages), and from Läkartidningen we extracted pages containing either the substring *diabet* or the substring *sockersjuk* (1739 pages).

8 Communicating Science: Swedish Medical Periodicals, 1781–2011 (SweMPer) is a collaboration between Uppsala University Library, The Centre for Digital Humanities and Social Sciences (CDHU) and The Department of History of Science and Ideas at Uppsala University.

9 https://github.com/skogsgren/pefselab (last accessed: February 14, 2025)

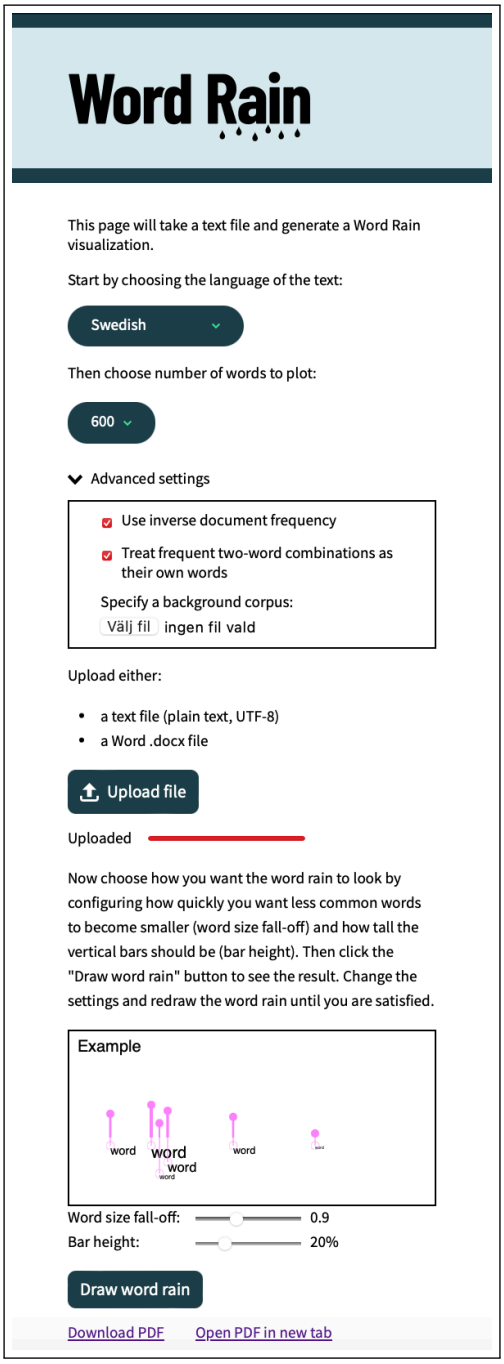


Figure 2: The Word Rain Web Service, i.e., a web page where texts can be uploaded to produce word rains.

The two text files were then uploaded to the Word Rain web service[10] (Ahltorp & Skeppstedt 2025), with parameters as shown in Figure 2. Parameters available include a slider for determining the height of the bars and how fast the font size should decrease for less prominent words. In the dialogue box that appears when "Upload file" is pressed, it is possible to select several files.

The web service produced the word rains shown in Figure 3, where the first graph shows the content from the patient periodical Diabetes, and the second graph shows the content from Läkartidningen.

When Swedish is chosen as the language of the corpora, words included in the standard Swedish stop word list from NLTK (Bird 2002) are used to filter which words are included in the visualisation. A pre-trained word2vec model, trained on Swedish text, is used for positioning the words. We configured the Word Rain web service to use TF-IDF as its prominence measure, i.e., a prominence measure that slightly exaggerates the differences between the corpora compared.

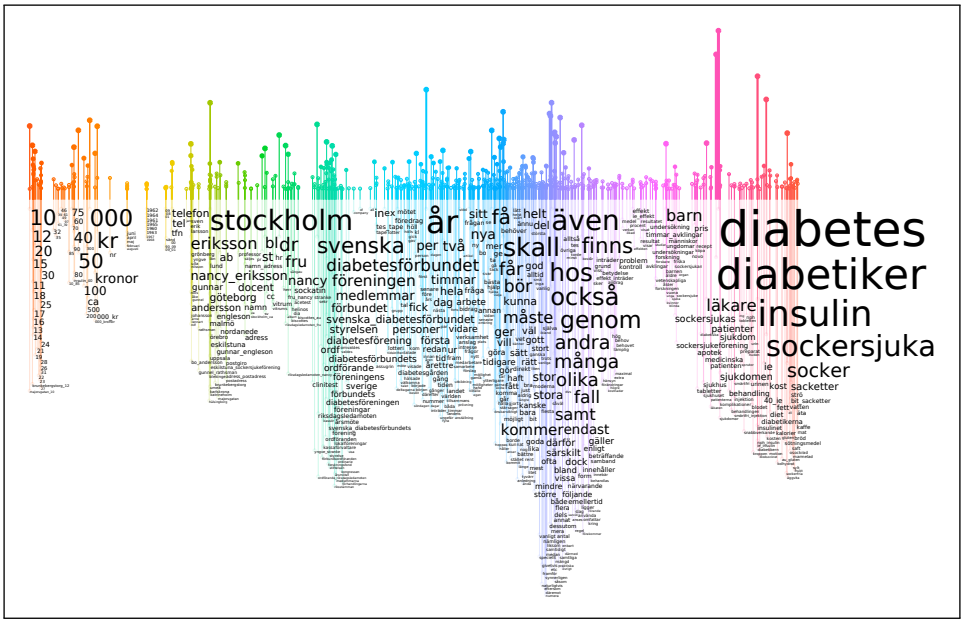
4.2.1 *Reading the word rains*

Despite using TF-IDF, it is still possible to observe many similarities between the two graphs in Figure 3. To the left, both graphs display different kinds of words and bigrams with numbers in orange/yellow, with more emphasis on years for Läkartidningen. To the right, in red/cerise, there are words closely related to diabetes for both texts, e.g., *diabetes* and *insulin*.

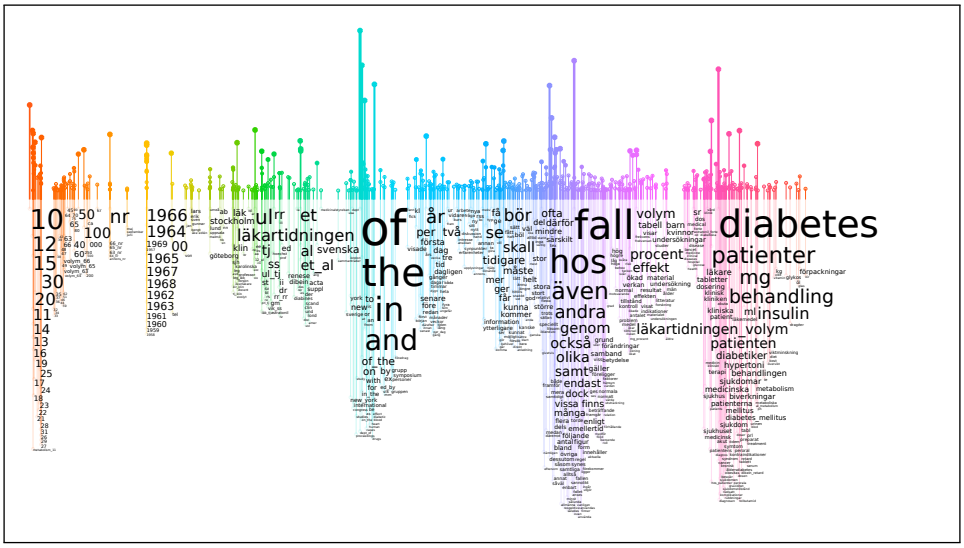
There are also differences between the two corpora; the most obvious being the vertical cluster of English words shown in turquoise in the Läkartidningen graph. While this cluster is very prominent in the Läkartidningen graph, the corresponding horizontal position is almost empty in the visualisation for the Diabetes periodical. That is, there are frequent English words in the Läkartidningen corpus, but not in the Diabetes corpus. Conversely, just to the left of the English words cluster, there is instead an empty segment in the Läkartidningen graph. At the same horizontal position in the Diabetes periodical graph, words related to the Swedish Diabetes Association, as well as words related to organisational matters in general, are located.

Another difference between the Diabetes periodical (directed towards patients) and the Läkartidningen journal (directed towards the medical profession) is the representation of places and people among the most common words. In the left part of both graphs in Figure 3, there is a vertical cluster of Swedish cities, more specifically at the horizontal position where the yellow colour turns into green. In Läkartidningen, only the three largest

10 https://wordrain.org (last accessed: October 7, 2025)



a. *Swedish Diabetes Association*



b. *Läkartidningen*

Figure 3: Comparing periodicals from the Swedish Diabetes Association (a), with content from Läkartidningen, a journal published by the Swedish Medical Association (b), for the first use case. Both corpora stem from texts on diabetes published during the 1960s.

cities in Sweden—as well as two university cities—have made it into the top 600 words on the pages discussing diabetes. In the Diabetes periodical, in contrast, there is a wider spread with some smaller cities included, as well as other types of places: two street names and a mention of the organisation's retreat *Nordanede*.

Names, and other words related to persons, are positioned directly to the left, as well as directly to the right, of the place cluster. In Läkartidningen, there is no full name with a frequency ranking high enough to be included among the top 600 most frequent words/bigrams. There is only one instance of a family name, and there are five instances of given names, all male. In Diabetes, in contrast, there are five different full names, mentioned frequently enough in the 1960s to be included among the most frequent words. Both female and male first names appear among the most prominent words. Frequently used titles in Diabetes include Mrs., Mr., doctor and professor, as well as titles in the form of functions within the patient organisation, such as president. In Läkartidningen, the titles are connected to the medical profession or academia, e.g., (associate) professor, licensed physician, chief physician and doctor.

These contrasts between the two word rains reflect the differences in content and audience of the corpora compared. However, they might not represent general differences between the two publications, since we used the entire content of the Diabetes periodical, while selecting a subset of Läkartidningen containing words related to diabetes. Nevertheless, the comparison of the two corpora form an illustrative example of how word rains can be read to detect differences between text collections.

The pink word cluster in the right-hand side of the graph contains words that could be described as "medical words". Here, there are differences related to vocabulary use. The word *diabetes* is frequent in both texts. The word *sockersjuka* 'sugar sickness', on the other hand, is very frequently used in the Diabetes periodical, but does not appear among the top 600 most frequent words in Läkartidningen. Conversely, in Läkartidningen the full medical term *diabetes mellitus* is used, but does not appear among the frequent words in the Diabetes periodical. Another difference regards what persons occur in the medical cluster. In both texts, the only medical profession frequently referred to in the 1960s (apart from references through titles) is *läkare* 'physician'. In the Diabetes periodical, the most important person in the medical word cluster is *diabetiker* 'diabetic', whereas *patienter* 'patients' is the most important person in Läkartidningen. Possibly, both words might often refer to a person with diabetes, but the vocabulary differs depending on the authors/target readers of the texts.

4.2.2 *Advantages/disadvantages of using the web service*

As previously mentioned, the large advantage of the Word Rain web service is that no programming is required to produce the word rains. The web service is therefore suitable both for those who do not have any need for the more detailed configuration options available when programming, as well as for those who first want to quickly try the visualisation technique on their corpus, to decide whether it would be useful to start using the Word Rain Python module.

A disadvantage of the web service, in addition to it offering fewer configuration options, is that not all corpora are possible to upload to an external organisation, e.g., corpora containing sensitive information. The code for the web service is, however, released as open source code.[11] It is therefore possible to set up the web service at any organisation where it is going to be used, either as a web service that can be accessed externally or as a local service within an organisation/at one specific computer.

4.3 *Visualising longitudinal trends*

The second use case focuses on only one of the corpora, the periodicals from the Swedish Diabetes Association. The task for this use case still includes comparison, but the aim of the word rains created is to visualise longitudinal trends. That is, instead of comparing two different text genres, as for the first use case, the aim is to make it possible to compare the content of each year to the content of the other years in the corpus. We therefore created a series of visualisations, with one word rain graph per year for the years 1949 and 1952–1990.[12]

The Word Rain Python module provides a number of configuration options for deciding which words to include in the graphs, for instance different types of word frequency cut-off configurations and different prominence measures. We have previously investigated and described the effects of varying the cut-off parameters with the aim of highlighting different aspects in the visualisation of longitudinal trends (Skeppstedt & Aangenendt 2024), and this will therefore not be the focus here. Instead, in use case two and three, we showcase two different *prominence measures* for visualising longitudinal trends.

As prominence measure in this second use case we employ raw word frequency. We also apply two different word frequency cut-offs: the word must occur at least ten times in the corpus and at least four times for the year

11 https://github.com/sprakradet/wordrain-service (last accessed: February 14, 2025)
12 There were no issues of the Diabetes periodical from the years 1950 and 1951.

visualised. For each year in the visualisation series, a graph is generated, which includes the top 600 most frequent words that fulfil the cut-off requirements. With these settings, we aim to showcase a longitudinal visualisation series with a configuration that is straight-forward to interpret.

Instead of using a pre-trained word2vec model, as in the first use case, a model was trained[13] on the same corpus that was visualised. Thereby, semantic similarity as expressed on closeness on the x-axis reflects semantic similarity in the corpus visualised, in contrast with general semantic similarity in the language, which was visualised in the first use case.

For the word rains in Figure 6, we have configured the graphs to include dashed vertical lines (and numbers associated with the lines). These lines have no other function than to offer a possibility to refer to different parts of the graphs, without having to rely on colour coding. We also used another colour gradient than in the previous example. The numbers in parentheses show the word frequency per 1000 words in the corpus.

We also configured the visualisation series to mark new words, i.e. words that have not appeared in the graphs for any of the previous years. These words are marked with a #-character and, when using the default font configuration, marked by an italic font. In addition, the bars corresponding to these new words are displayed in a grey colour.

4.3.1 *Stop words*

There are different ways to avoid very frequent words taking up a lot of space in the graphs, for instance, excluding words that occur all of the years in the series, or including general-language background corpora and using TF-IDF as the prominence measure. For this aim, it is also possible to use word rain configuration options for font size and for "word size fall-off", i.e., for how quickly the font size of less prominent words is to be decreased.

Another possibility, which is the method used here, is to employ a manually created stop word list, i.e., a list of words to exclude from the visualisations. We first employed the standard Swedish stop word list included in NLTK (Bird 2002), and then manually extended this list with words that took up a lot of space in the graphs. Most of these words were either semantically bleached in nature—e.g., *kunna* 'can' and *komma* 'come'—or were semantically rich words with a frequency that did not vary much over time—e.g., *insulin* and *diabetes*.[14]

13 For training the model, a window size of three and a vector size of 50 were used.

14 The use of *diabetes* actually *did* vary over time, as *sockersjuk* 'sugar-sick' was a more common term during the first six years visualised. However, if we treat these two terms as alternative ways to refer to the same concept, references to that concept were relatively stable over time.

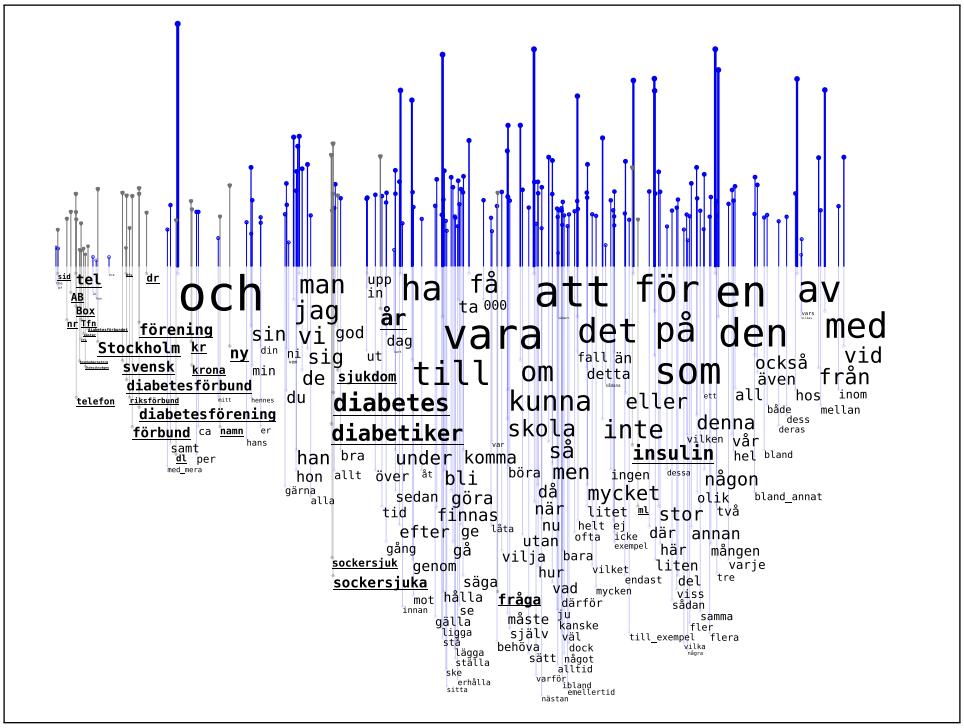
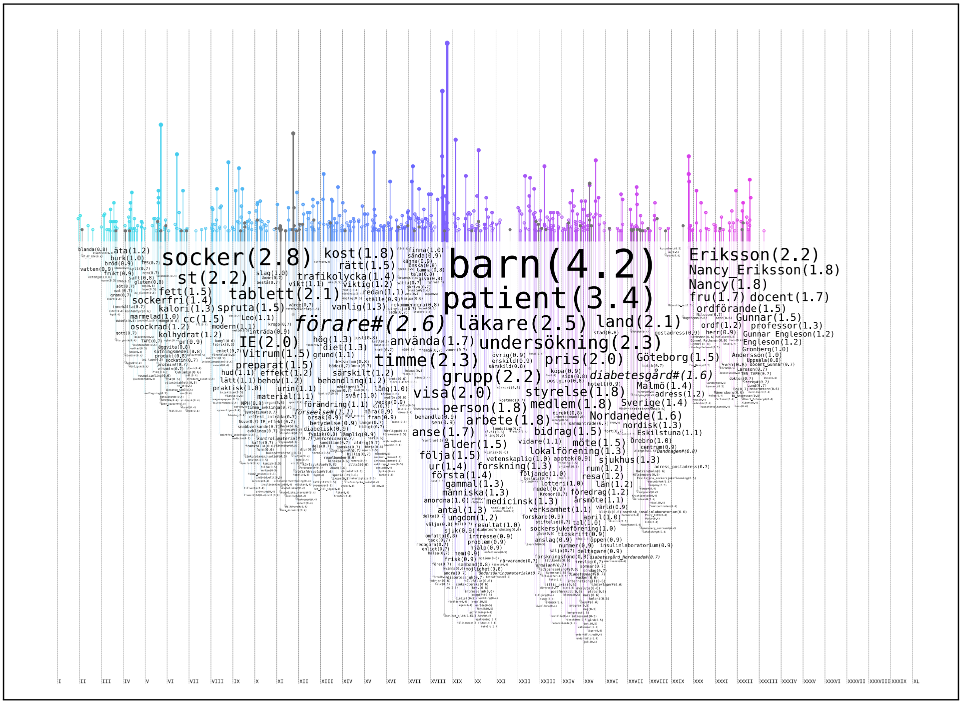


Figure 4: The stop words used in the second use case.

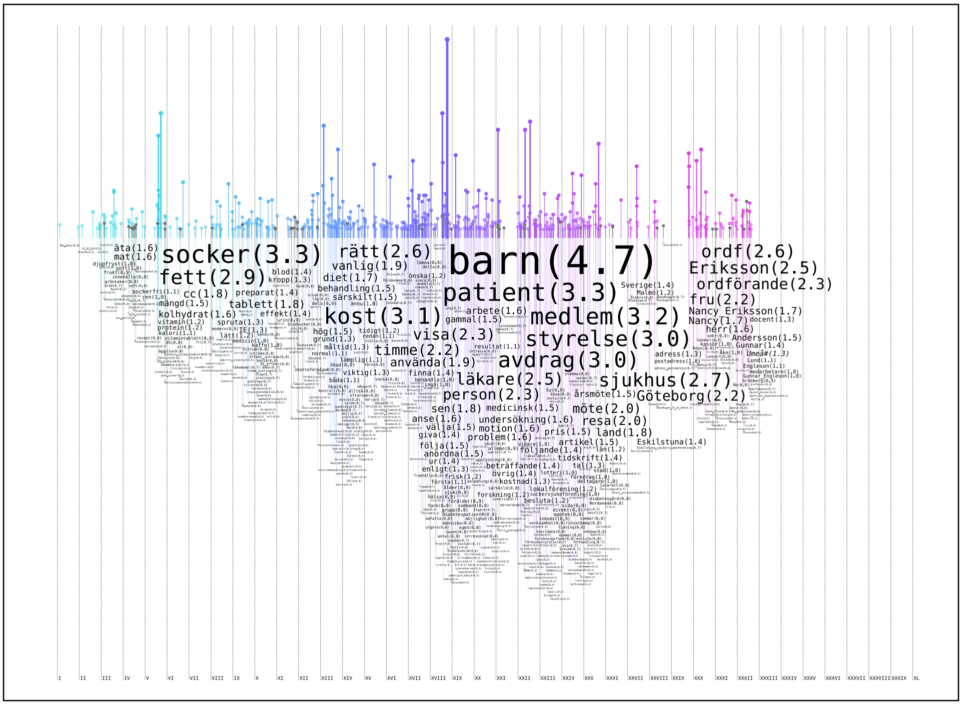
The stop words used are visualised in Figure 4, with the more semantically rich words underlined. Note that the visualisation of the stop words in Figure 4 is created as a completely separate graph from the series of graphs shown in Figures 5 and 6, with another projection of the words on the x-axis.

In addition to visualising the stop words, the graph in Figure 4 also showcases a configuration possibility of the Word Rain code, where the user can define a set of pre-defined words that are the only ones to be included in the graph. That is, instead of specifying a set of words to *exclude*, as for the stop words used for Figures 5 and 6, we here configure the word rain to *only include* a set of pre-defined words. The possibility to specify sets of words to exclude and include, thus, forms an option for easily tailoring the content of the word rains to specific visualisation needs. We have, for instance, previously experimented with compiling a list of named entities found in a text set, and both produced word rains that exclude the named entities (using the stop word functionality for excluding words), and word rains that only include the named entities (using the inclusion list).

In addition to showcasing the possibility to specify a pre-defined set


barn(4.2)
patient(3.4)
socker(2.8)
förare#(2.6)
läkare(2.5)
undersökning(2.3)
timme(2.3)
grupp(2.2)
st(2.2)
Eriksson(2.2)
tablett(2.1)
land(2.1)
IE(2.0)
visa(2.0)
pris(2.0)
Nancy_Eriksson(1.8)
Nancy(1.8)
kost(1.8)
person(1.8)
medlem(1.8)
styrelse(1.8)
arbete(1.8)
fru(1.7) docent(1.7)
ordförande(1.5)
Göteborg(1.5)
diabetesgård#(1.6)
Nordanede(1.6)
ålder(1.5)
möte(1.5)
bidrag(1.5)
ta(1.2) äta(1.2)
rätt(1.5)
trafikolycka(1.4)
fett(1.5)
Vitrum(1.5)
spruta(1.5)
preparat(1.5)
sockerfri(1.4)
kalori(1.3)
cc(1.5)
använda(1.7)
anse(1.7)
följa(1.5)
ur(1.4)
första(1.4)
Sverige(1.4)
Malmö(1.4)
Eskilstuna(1.1)
sjukhus(1.3)
forskning(1.3)
lokalförening(1.3)
I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII XVIII XIX XX XXI XXII XXIII XXIV XXV XXVI XXVII XXVIII XXIX XXX XXXI XXXII XXXIII XXXIV XXXV XXXVI XXXVII XXXVIII XXXIX XL

a. *1965*


barn(4.7)
socker(3.3)
patient(3.3)
medlem(3.2)
kost(3.1)
styrelse(3.0)
avdrag(3.0)
fett(2.9)
sjukhus(2.7)
rätt(2.6)
ordf(2.6)
Eriksson(2.5)
läkare(2.5)
ordförande(2.3)
visa(2.3)
person(2.3)
timme(2.2)
Göteborg(2.2)
fru(2.2)
möte(2.0)
resa(2.0)
vanlig(1.9)
använda(1.9)
cc(1.8)
tablett(1.8)
sen(1.8)
land(1.8)
äta(1.6)
mat(1.6)
kolhydrat(1.6)
arbete(1.6)
undersökning(1.6)
Nancy_Eriksson(1.7)
Nancy(1.7) docent(1.3)
herr(1.6)
Andersson(1.5)
spruta(1.3)
preparat(1.4)
behandling(1.5)
särskilt(1.5)
mängd(1.5)
diet(1.7)
effekt(1.4)
medicinsk(1.5)
årsmöte(1.5)
anse(1.6)
välja(1.5)
följa(1.5)
anordna(1.5)
pris(1.5)
artikel(1.5)
Eskilstuna(1.4)
I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI XVII XVIII XIX XX XXI XXII XXIII XXIV XXV XXVI XXVII XXVIII XXIX XXX XXXI XXXII XXXIII XXXIV XXXV XXXVI XXXVII XXXVIII XXXIX XL

b. *1966*

◀ Figure 5: Comparing two adjacent years, 1965 and 1966, for the second use case.

of words to include, Figure 4 also showcases i) the possibility to provide a user-defined list of words to underline, and ii) the possibility to create a monochrome word rain.

There is a trade-off between avoiding a very frequent, consistently occurring, word to take up a lot of space in the graph, and removing potentially interesting content from the visualisation. This is particularly the case when also removing semantically rich words, as for the example case presented here. Therefore, it might be relevant—as was done here—to also produce a visualisation of the content that was excluded.

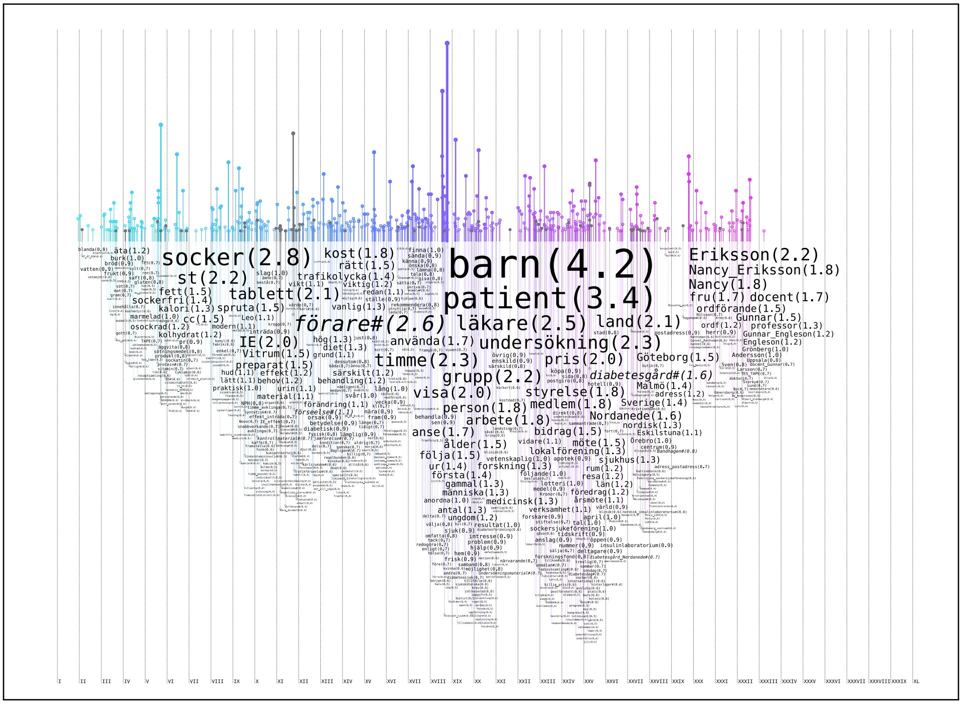
4.3.2 *Reading the word rains*

As examples from the visualisation series (which in total includes one graph per year, for the years 1949, 1952–1990), we compared graphs from the two adjacent years 1965 and 1966 (Figure 5), and graphs from two years far apart, 1965 and 1985 (Figure 6).

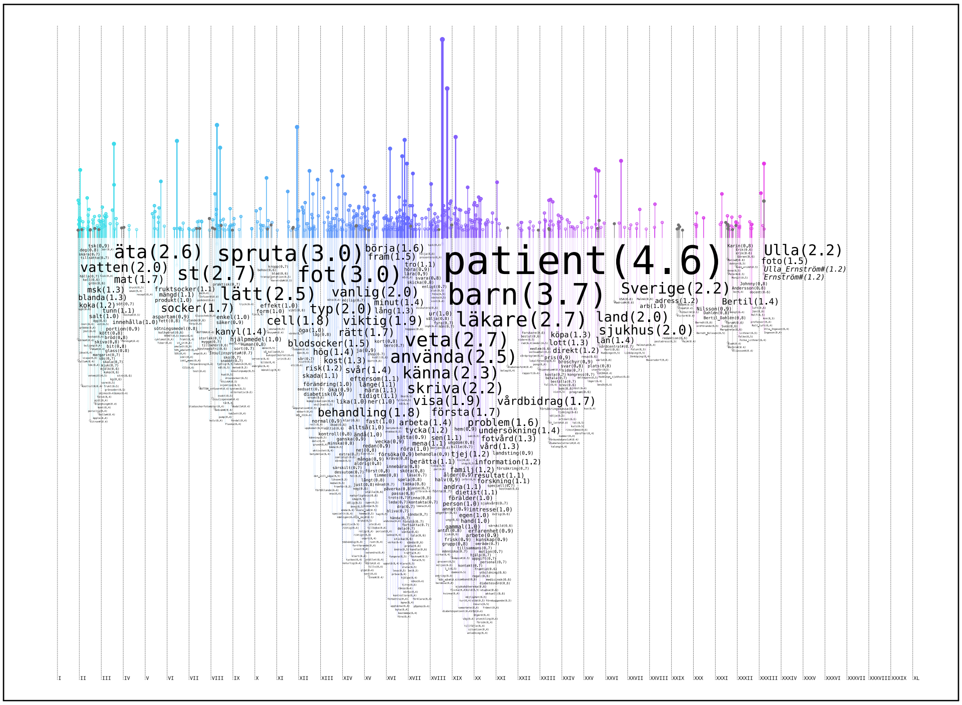
Table 1 contains a summary of the semantic word categories we were able to detect in the graphs in Figures 5 and 6. The roman numbers indicate an approximate position for the words in the graph. The categories within parentheses are only marginally represented in the graphs for the years 1965, 1966 or 1985, typically with only one or two words. However, these categories are prominent other years in the visualisation series—i.e., in the series created for the years 1949–1990—and we therefore included these categories in the table. The year in parentheses is the first year the semantic category appeared in the visualisation series.

The differences between the two adjacent years, 1965 and 1966, in Figure 5 are less obvious than between the two different corpora compared above. Instead, what stands out is rather the similarity between year 1965 and year 1966. There are, however, still examples of immediately evident differences. For instance, words from 1965 related to traffic and driving, e.g., *förare* 'driver' and *trafikolycka* 'traffic accident', are no longer present among the top 600 most frequent words in 1966. Instead, some words related to food have become more prominent in 1966, e.g., *kost* 'food/diet' and *fett* 'fat', as well as words related to organisation administration, e.g., *medlem* 'member' and *styrelse* 'board' and economics, e.g., *avdrag* 'deduction'.

With the grey colour marking, it is possible to spot semantic regions of new words in the graph. For instance, close to line XI in the graph for 1965,


barn(4.2)
patient(3.4)
socker(2.8)
förare#(2.6)
läkare(2.5)
undersökning(2.3)
timme(2.3)
st(2.2)
grupp(2.2)
Eriksson(2.2)
tablett(2.1)
land(2.1)
IE(2.0)
pris(2.0)
visa(2.0)
Nancy_Eriksson(1.8)
Nancy(1.8)
kost(1.8)
styrelse(1.8)
medlem(1.8)
person(1.8)
arbete(1.8)

a. *1965*


patient(4.6)
barn(3.7)
spruta(3.0)
fot(3.0)
st(2.7)
läkare(2.7)
veta(2.7)
äta(2.6)
lätt(2.5)
använda(2.5)
känna(2.3)
Ulla(2.2)
skriva(2.2)
Sverige(2.2)
vatten(2.0)
vanlig(2.0)
typ(2.0)
land(2.0)
sjukhus(2.0)

b. *1985*

◀ Figure 6: The most frequent words/bigrams (after stop word removal) in periodicals from the Swedish Diabetes Association, from 1965 and 1985.

Table 1: Semantic categories in Figures 5 and 6.

Semantic category	Label	Appears
Food/recipes	II–V	1953
(Self tests for sugar)	V–VII	1954
Nutrition	VI	1953
Artificial sweeteners and types of sugar	VI	1954
Insulin and other medication	VIII–X	1949
Blood and urine	XI	1953
(Processes, body parts/substances involved in the cause of diabetes)	X–XI	1953
Medical complications, risks and related body parts	XII–XIII	1953
Traffic and driving	XIII	1965
Diet/weight/exercise/treatment	XIII–XIV	1949
Words for discussions and feelings	XVII–XVIII	1955
Children/youths, other people, professions	XIX	1949
Work (employment)	XX	1949
Research, advice, development, education	XX–XXI	1949
Economical matters	XXI	1949
Organisation words, e.g., member/board	XXII–XXIII	1949
Lottery, fundraising	XXIII	1957
Research or research funding	XXIII	1953
Organisational activities	XXIV–XXV	1954
Published/printed material	XXIV	1949
Summer camp/Nordanede/trips	XXV–XXVI	1949
Places where healthcare is taking place	XXVI	1954
Places outside Sweden	XXVI–XXVII	1954
Swedish cities and addresses	XXVII–XXIV	1949
Organisation functions, titles and names	XXX–XXXI	1949
Academic titles, names and university cities	XXXII–XXXIII	1949

there are a number of words not previously included in the series of graphs.

In Figure 6, which instead compares two years far apart, 1965 and 1985, the differences are larger. Many (but not all) semantic categories are present in both years, but they often differ in content. For instance, the insulin/medication category in 1985 reflects the spread of new technology around this time (Bradwel 2023: pp. 80–82, 85–87), with new words, such as *insulinpump* 'insulin pump' and *insulinpenna* 'insulin pen'. There is also more focus on the

Figure 7: Comparing two adjacent years, 1965 and 1966, with the frequency difference compared to previous year as the prominence measure for the third use case. ▶

body affected by diabetes, with *fot* 'foot' as one of the most prominent words in 1985, as well as *öga* 'eye' and *sår* 'wound' being prominent words. The food category is also more prominent, and it includes more verbs frequently used in recipes, such as *koka* 'boil' and *blanda* 'mix'.

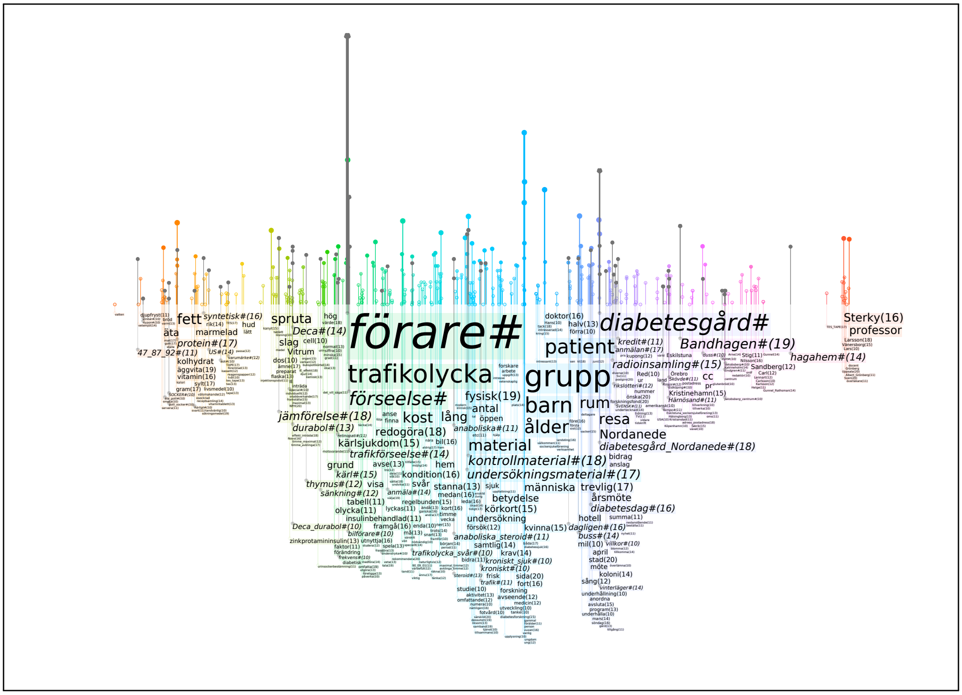
The graphs also show a change in common organisational activities, with many different kinds of activities in 1965—e.g., lectures, meetings, diabetes day, singing—while only congresses, national meetings and courses remain in 1985. Activities related to the diabetes retreat Nordanede, which was acquired in 1963 and closed in 1983 (Swedish Labour Movement's Archives and Library, Personal archives of Nancy Eriksson n.d.), are also not any more present in the graph for 1985. Organisation-related words in general receive much less attention in 1985, e.g., with words such as *styrelse* 'board' and *medlem* 'member' being very prominent in 1965, but not at all in 1985.

4.4 *Longitudinal trends, but with a focus on novelty*

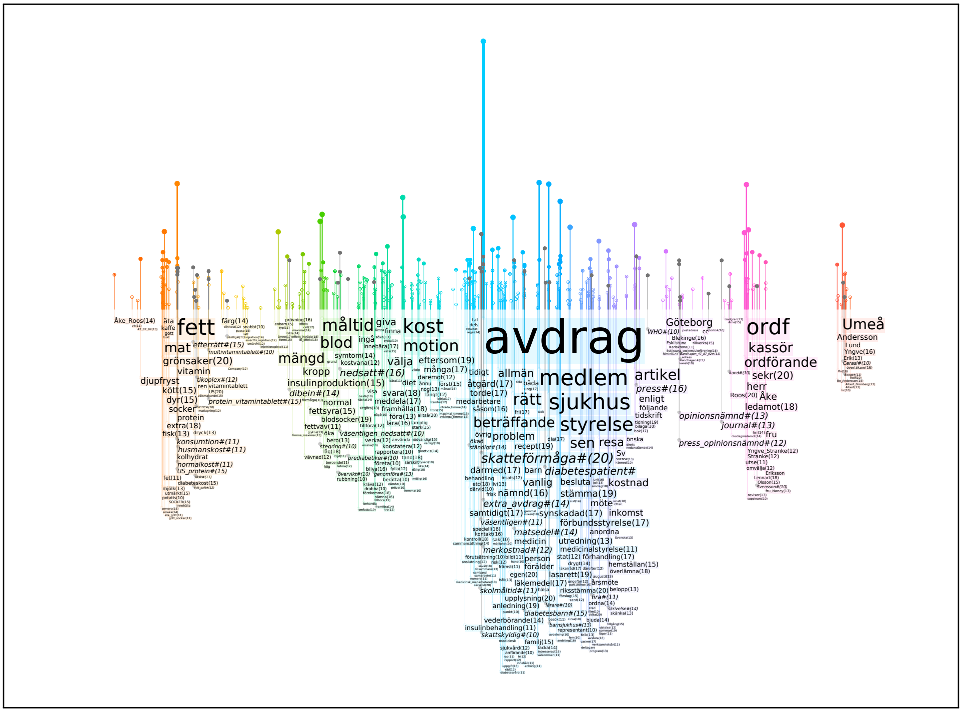
For the third use case, we also show visualisation examples from the two adjacent years, 1965 and 1966. Here, we do, however, not apply raw word frequency as the prominence measure for selecting which words to show. Instead, we use the word frequency *difference* between the year that is visualised and the previous year. That is, the graph for 1965 in Figure 7 shows the words for which the frequency has increased the most compared to 1964. Similarly, for the graph for 1966, the prominence measure applied is the increase in frequency from 1965 to 1966. Practically, we use the differences in relative frequency.

In contrast with the previous use case, there is no ready-made word rain configuration setting for using differences in relative frequency as the prominence measure. Instead, the Word Rain code offers a way for the user to submit their own Python function for defining word prominence. This functionality provides the user with a large degree of flexibility in defining the prominence measure, but, as already mentioned, also requires more programming skills.

The point of the prominence measure for the third use case is to make it possible to detect words that show a large increase in usage for the year visualised. This prominence measure is not only more difficult to technically implement, but also slightly more difficult to interpret than the standard fre-


förare#
trafikolycka
förseelse#
grupp
patient
diabetesgård#
barn
ålder
material
resa
Nordanede
kost
fett
spruta
Bandhagen#(19)
Sterky(16)
professor

a. *1965*


avdrag
fett
måltid
kost
blod
motion
mängd
medlem
sjukhus
rätt
styrelse
beträffande
artikel
ordf
kassör
ordförande
Umeå
Göteborg
skatteförmåga#(20)
mat
grönsaker(20)

b. *1966*

quency measures. When analysing these results, it is important to remember that novelty/frequency increase is the sole focus of the visualisation. When frequency for a word instead has *decreased* compared to the previous year, the word is therefore not included in the graph for that year, despite that this word might be a frequently occurring one also that year. When analysing visualisations using differences in relative frequency as the prominence measure, it is thereby useful to also look at word rain visualisations that use the raw word frequency measure. For instance, in Figure 7, the prominence measure focusing on novelty makes it easier to detect the large frequency increase for the two words *diabetesgård* 'diabetes retreat' and *Nordanede* 'name of a retreat place' (both shown in violet) in 1965, than to detect frequency changes by comparing two graphs that visualise raw word frequency. However, to study the general development of these two words, e.g., how much they then decreased in 1966, standard word frequency is a better prominence measure.

Apart from using a different prominence measure, we altered a few additional configuration settings compared to the previous use case, in order to showcase other possible word rain settings. For instance, the top 400 most prominent words are shown (instead of the top 600), the minimum word frequency in the text visualised is set to ten, the labelled lines are not included, and the number shown in parentheses is the word's absolute frequency in the text visualised (and this is only shown for a frequency of 20 and less).

For this use case, we also showcase a configuration option that allows the user to somewhat influence the result of the projection on the x-axis. This configuration was also used for the visualisation of the British Diabetes Association texts in Figure 1. With this configuration option, the user defines a k number of clusters, to use for clustering the words, and more distinct word clusters along the x-axis are thereby created. When comparing the graphs in Figures 5 and 6 to the graphs in Figures 1 and 7, it is possible to see that the words in the latter figures are positioned in more distinct clusters along the x-axis. In Figure 7, we configured the projection to use 50 clusters, and in Figure 1 to use 30 clusters.

The same wide colour scale as in Figure 1 was used for the colour gradient. For both Figure 1 and Figure 7, the colours and clusters were emphasised even further by using a configuration option which provides the words with a coloured background box. With the purpose of reminding about the possibility to vary the parameters for the word2vec model used, we here trained a new word2vec model on the corpus, slightly altering[15] the parameters used.

15 More specifically, increasing to a window size of five words and a vector size of 100.

4.4.1 *Reading the word rains*

In the graph for 1965, the previously mentioned prominent semantic categories are easier to detect when word frequency difference is used as prominence indicator. In green, under the word *förare* 'driver', there are a number of prominent words related to traffic and driving, mainly regarding traffic accidents. Similarly, there are many words related to the diabetes retreat Nordanede (between blue and violet), that have increased in 1965 compared to 1964. This is also the case for words related to nutrition (to the left in orange). There are also some individual words that stand out. For instance, among organisational activities (between blue and violet), there is the new word *diabetesdag* 'diabetes day', an annual campaign day with among other things a nationwide fundraising effort through the radio introduced in 1965 (Pehrson 1965). Consequently, among the lottery/fundraising words (in violet) *radioinsamling* 'fundraising through the radio' stands out.

In 1966, words related to nutrition continue to increase in frequency (to the left in orange), and organisational words also increase (in pink). Most prominent here, is the economic category (in blue, under the word *avdrag* 'deduction'). Also here, there are individual words that have increased, e.g., *kost* 'food/diet' and *motion* 'exercise'.

5 *Word Rain in relation to other text visualisation techniques*

We will end this chapter by positioning the Word Rain technique in the context of other text visualisation techniques. The content of this section is not required for understanding, generating or reading the word rains, but might be useful for comparing the technique to other possible choices for visualising or listing the most prominent words in a text or corpus (Felix et al. 2018).

We will start by comparing the Word Rain technique to the techniques it aims to replace, i.e., to word frequency lists and the classic word cloud. Thereafter, we will provide some other examples of word cloud extensions.

5.1 *Word Rain compared to a word frequency table and a word cloud*

By creating a word frequency list (Table 2) and a classic word cloud (Figure 8) for the same texts we illustrated above in Figure 6, we aim to emphasise the advantages of the Word Rain technique compared to the other two methods.

Similar to the word clouds, the word rains in Figure 6 also immediately show which are the most important words in the texts. However, when the Word Rain technique is used for visualising the texts, much more information

Table 2: A table showing the 30 most frequent words/bigrams (after stop word removal) in periodicals from the Swedish Diabetes Association, from the year 1965 and the year 1985. Frequency is shown by the number of occurrences per 1000 words (‰) in the periodicals from that year.

	Year 1965		Year 1985	
Rank	Word	‰	Word	‰
1	*barn*	4.2	*patient*	4.7
2	*patient*	3.4	*barn*	3.7
3	*socker*	2.8	*spruta*	3.0
4	*förare*	2.6	*fot*	3.0
5	*läkare*	2.5	*läkare*	2.8
6	*undersökning*	2.3	*veta*	2.7
7	*timme*	2.3	*st*	2.7
8	*st*	2.2	*äta*	2.6
9	*grupp*	2.2	*lätt*	2.5
10	*Eriksson*	2.2	*använda*	2.5
11	*land*	2.1	*känna*	2.3
12	*tablett*	2.1	*Sverige*	2.2
13	*visa*	2.0	*skriva*	2.2
14	*pris*	2.0	*Ulla*	2.2
15	*IE*	2.0	*land*	2.0
16	*kost*	1.8	*vatten*	2.0
17	*Nancy Eriksson*	1.8	*vanlig*	2.0
18	*Nancy*	1.8	*typ*	2.0
19	*styrelse*	1.8	*sjukhus*	2.0
20	*person*	1.8	*visa*	1.9
21	*arbete*	1.8	*viktig*	1.9
22	*medlem*	1.8	*cell*	1.8
23	*krona*	1.8	*behandling*	1.8
24	*docent*	1.7	*socker*	1.8
25	*fru*	1.7	*vårdbidrag*	1.7
26	*använda*	1.7	*första*	1.7
27	*anse*	1.7	*rätt*	1.7
28	*diabetesgård*	1.6	*mat*	1.7
29	*Nordanede*	1.6	*krona*	1.6
30	*ålder*	1.5	*problem*	1.6

a. *1965*

b. *1985*

Figure 8: Two word clouds showing the most frequent words/bigrams (after stop word removal) in periodicals from the Swedish Diabetes Association, from the year 1965 and the year 1985.

can be derived from the graphs. As exemplified in the use cases, the word rains make it possible to find semantic word categories, compare texts, and zoom in into interesting areas. We argue that it would have been very difficult to find prominent semantic word categories—i.e., to produce the content of Table 1—from the classic word clouds or from the word frequency lists (Table 2 and Figure 8). Similarly, the analyses of differences between

the graphs would also have been very difficult to produce based on the information provided by the classic word clouds or frequency lists.

5.2 *Previous improvements of the classic word cloud*

There are many other previously applied approaches, addressing some of the limitations of the word cloud. To address the absence of a meaningful word positioning, different types of semantically motivated word positioning algorithms for static word clouds have been proposed (Wu et al. 2011, Barth et al. 2014), including those based on word embeddings (Xu et al. 2016). The methods typically keep a layout similar to the classic word clouds, but position semantically related words close to each other, and emphasise semantic clusters/relations by colour coding the words.

Also the problem associated with the difficulties of comparing word clouds has been addressed through solutions visually very similar to a classic word cloud. For instance, word clouds more suitable for comparison can be achieved by assigning words a fixed position in a series of word cloud graphs, and only allowing sets of words that never occur in the same word cloud to occupy the same space (Herold et al. 2019). Another possible approach is to apply restraints on the word positioning to keep the contexts of the words similar in each graph produced in a series of word cloud graphs (Cui et al. 2010). A very different approach for addressing the difficulty of comparing word clouds is to produce a single word cloud for all the texts that are to be compared, and then use colours (Diakopoulos et al. 2015, Burch et al. 2014) or trend lines (Lee et al. 2010) to indicate in which texts the word occurs or how the word has varied over time.

There are also solutions to the problem of using font size as prominence indicator. For instance, Alexander et al. (2018) provided the idea of supplementing font size with a padded bounding box around the words, which has the effect of short words taking up as much space as longer ones.

The use of further typographic attributes—such as italic font or to underline text—has also been suggested, as a means beyond font size or colour to indicate word properties in word clouds (Brath & Banissi 2016).

There are thus many different approaches for addressing the limitations associated with the classic word cloud. As previously mentioned, however, these approaches typically focus on addressing only *one* limitation. With the Word Rain technique we, instead, aim to address *several* of the limitations that make the word cloud unsuitable as a tool for analysing text.

6 *Conclusion*

We have in this chapter described limitations of the classic word cloud, as well as how the Word Rain technique addresses these limitations. We have then provided three use case examples, with an increasing level of difficulty, which show how the word rains can be used for exploring, analysing and comparing the most prominent words in different texts. Finally, we have compared the Word Rain technique to using a word frequency list and a classic word cloud, as well as described previous approaches for improving the word cloud.

With this description of the Word Rain technique—and of how it can be practically applied—we hope to have convinced you to use a Word Rain visualisation, where you previously would have used a word frequency list or a word cloud. We also hope that our practical examples have shown you how a series of semantically structured word graphs very quickly can provide you with an overview understanding of the content of a text collection, as well as with an understanding of how this content varies over time. We would argue that after spending a short time analysing the visualisations generated here, even a person with no prior knowledge of the Swedish Diabetes Association would be able to possess an overview understanding of the content of its member publication, for the years discussed.

In many usage scenarios for the Word Rain technique, the word rains could be employed as a first step when approaching a new text collection to study. The technique provides a content overview of the texts, as well as lets the user explore and structure more specific, and potentially interesting, content. This overview understanding might be enough, but it might also be relevant to take the word rains as a point of departure, and dig deeper into the content elsewhere. For instance, to produce frequency trend lines for interesting words/word clusters found in the visualisations, or to use the word rain as a basis for selecting a subset of texts to read more closely. These further analyses are, however, out of scope for this handbook chapter. This is also the case for different types of text pre-processing techniques—e.g., named entity recognition or topic modelling—for further tailoring the visualisation to different user requirements. Also a discussion of how to best generate word rains to use as graphic design elements on printed material is left to future publications.

Through the three use cases with an increasing level of complexity, we have shown how the Word Rain technique can support both the user who wants to upload texts on a web page and create a standard word rain, as well as the user who wants to adapt the word rains to their specific visualisation requirements. The three use cases presented here are, however, just the

beginning. There are still many unexplored possibilities to further tweak the word rains, for instance, by using different text pre-processing techniques before generating the word rains, or by applying different word filtering options, different colour schemes and/or different prominence measures. Even larger are the possibilities in the form of different text collections on which it would be relevant to apply the Word Rain visualisation. We therefore look forward to following how the Word Rain web service—as well as the Word Rain Python module—will be applied by future users to generate novel types of visualisations for different kinds of text collections.

Acknowledgments

The work described in this handbook chapter has mainly been conducted within the ActDisease project, which is funded by the European Union (ERC ActDisease ERC-2021-STG 101040999). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The development of the Word Rain Python module and the Word Rain web service is also funded by the Swedish Research Council: Huminfra (2021-00176, 2023-00171), InfraVis (2021-00181) and Swe-CLARIN/The National Language Bank of Sweden (2017-00626).

Finally, Communicating Science: Swedish Medical Periodicals, 1781–2011 (SweMPer) is funded by Riksbankens jubileumsfond.

References

Aangenendt, Gijs, Maria Skeppstedt & Ylva Söderfeldt. 2024. Curating a historical source corpus of 20th century patient organization periodicals. In *Proceedings of the Huminfra Conference (hic 2024)*, 76–82. DOI: 10.3384/ecp205011.

Ahltorp, Magnus, Jean Hessel, Gunnar Eriksson & Maria Skeppstedt. 2025. Visualisering av ett lexikons täckning av olika textgenrer: Experiment med en jiddischordbok [Visualising the coverage of dictionary for different text genres: Experiments with a Yiddish dictionary]. In *Nordiske studier i leksikografi*. Accepted for publication.

Ahltorp, Magnus & Maria Skeppstedt. 2025. Word Rain as a service: Making semantically structured word clouds available to everyone. In Vincent Vandeghinste & Thalassia Kontino (eds.), *Selected papers from the CLARIN annual conference 2024* (Linköping Electronic Conference Proceedings 216). DOI: https://doi.org/10.3384/ecp216.

Alexander, Eric, Chih-Ching Chang, Mariana Shimabukuro, Steven Franconeri, Christopher Collins & Michael Gleicher. 2018. Perceptual biases in font size as a data encoding. *IEEE Transactions on Visualization and Computer Graphics* 24(8). 2397–2410. DOI: 10.1109/TVCG.2017.2723397.

Barth, Lukas, Stephen G. Kobourov & Sergey Pupyrev. 2014. Experimental comparison of semantic word clouds. In *Experimental algorithms*, 247–258. Springer International Publishing. DOI: 10.1007/978-3-319-07959-2_21.

Bird, Steven. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL workshop on effective tools and methodologies for teaching natural language processing and computational linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bradwel, Stuart. 2023. *Insulin: A hundred-year history*. Cambridge: Polity.

Brath, Richard & Ebad Banissi. 2016. Using typography to expand the design space of data visualization. *She Ji: The Journal of Design, Economics, and Innovation* 2(1). 59–87. DOI: 10.1016/j.sheji.2016.05.003.

Burch, Michael, Steffen Lohmann, Fabian Beck, Nils Rodriguez, Lorenzo Di Silvestro & Daniel Weiskopf. 2014. RadCloud: Visualizing multiple texts with merged word clouds. In *Proceedings of the international conference on information visualisation* (IV '14), 108–113. DOI: 10.1109/IV.2014.72.

Cao, Nan & Weiwei Cui. 2016. *Introduction to text visualization*, vol. 1 (Atlantis Briefs in Artificial Intelligence). Atlantis Press. DOI: 10.2991/978-94-6239-186-4.

Collins, Christopher, Fernanda B. Viégas & Martin Wattenberg. 2009. Parallel tag clouds to explore and analyze faceted text corpora. In *Proceedings of the IEEE symposium on visual analytics science and technology*, 91–98. DOI: 10.1109/VAST.2009.5333443.

Cuba Gyllensten, Amaru. 2023. *Quantifying meaning*. KTH Royal Institute of Technology. (Doctoral dissertation).

Cui, Weiwei, Yingcai Wu, Shixia Liu, Furu Wei, Michelle Zhou & Huamin Qu. 2010. Context-preserving, dynamic word cloud visualization. *IEEE Computer Graphics and Applications* 30(6). 42–53. DOI: 10.1109/MCG.2010.102.

Diakopoulos, Nicholas, Dag Elgesem, Andrew Salway, Amy Zhang & Knut Hofland. 2015. Compare Clouds: Visualizing text corpora to compare media frames. In *Proceedings of the IUI workshop on visual text analytics*.

Felix, Cristian, Steven Franconeri & Enrico Bertini. 2018. Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE Transactions on Visualization and Computer Graphics* 24(1). 657–666. DOI: 10.1109/TVCG.2017.2746018.

Hearst, Marti A., Emily Pedersen, Lekha Patil, Elsie Lee, Paul Laskowski & Steven Franconeri. 2020. An evaluation of semantically grouped word cloud designs. *IEEE Transactions on Visualization and Computer Graphics* 26(9). 2748–2761. DOI: 10.1109/TVCG.2019.2904683.

Herold, Elisa, Marcus Pöckelmann, Christian Berg, Jörg Ritter & Mark M. Hall. 2019. Stable word-clouds for visualising text-changes over time. In *Digital libraries for open knowledge*, 224–237. Springer International Publishing. DOI: 10.1007/978-3-030-30760-8_20.

Hicke, Rebecca M. M., Maanya Goenka & Eric Alexander. 2022. Word clouds in the wild. In *Proceedings of the IEEE workshop on visualization for the digital humanities*, 43–48. DOI: 10.1109/VIS4DH57440.2022.00015.

Jänicke, S., G. Franzini, M. F. Cheema & G. Scheuermann. 2017. Visual text analysis in digital humanities. *Computer Graphics Forum* 36(6). 226–250. DOI: 10.1111/cgf.12873.

Jänicke, Stefan. 2022. A research-teaching guide for visual data analysis in digital humanities. *Computer vision, imaging and computer graphics theory and applications*. 205–222. DOI: 10.1007/978-3-030-94893-1_9.

Koh, Kyle, Bongshin Lee, Bohyoung Kim & Jinwook Seo. 2010. ManiWordle: Providing flexible control over Wordle. *IEEE Transactions on Visualization and Computer Graphics* 16(6). 1190–1197. DOI: 10.1109/TVCG.2010.175.

Kucher, Kostiantyn & Andreas Kerren. 2015. Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proceedings of the IEEE Pacific visualization symposium*, 117–121. DOI: 10.1109/PACIFICVIS.2015.7156366.

Lee, Bongshin, Nathalie Henry Riche, Amy K. Karlson & Sheelash Carpendale. 2010. SparkClouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics* 16(6). 1182–1189. DOI: 10.1109/TVCG.2010.194.

Liu, Xiaotong, Han-Wei Shen & Yifan Hu. 2015. Supporting multifaceted viewing of word clouds with focus+context display. *Information Visualization* 14(2). 168–180. DOI: 10.1177/1473871614534095.

Lohmann, Steffen, Michael Burch, Hansjörg Schmauder & Daniel Weiskopf. 2012. Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds. In *Proceedings of the international working conference on advanced visual interfaces*, 753–756. DOI: 10.1145/2254556.2254701.

Lohmann, Steffen, Jürgen Ziegler & Lena Tetzlaff. 2009. Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Human-computer interaction – interact 2009*, 392–404. DOI: 10.1007/978-3-642-03655-2_43.

Ortiz Pablo, Dalia, Sushruth Badri, Gijs Aangenendt, Mo von Bychelberg & Matts Lindström. 2025. A machine learning pipeline for digitalising historical printed materials – from data collection to a searchable database. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 207–250. University of Tartu Library. DOI: 10.58009/aere-perennius0177.

Östling, Robert. 2018. Part of speech tagging: Shallow or deep learning? *Northern European Journal of Language Technology* 5. 1–15. DOI: 10.3384/nejlt.2000-1533.1851.

Pehrson, Birger. 1965. 1964 blev ett händelserikt år. *Diabetes* (7).

Ren, Hui, Yuan Liu, Gaowa Naren & Junyi Lu. 2024. The impact of multidirectional text typography on text readability in word clouds. *Displays* 83. 102724. DOI: 10.1016/j.displa.2024.102724.

Skeppstedt, Maria & Gijs Aangenendt. 2024. Using the Word Rain technique to visualize longitudinal changes in periodicals from the Swedish Diabetes Association. In *Proceedings of the workshop on visualization for natural language processing*. DOI: 10.2312/vis4nlp.20241132.

Skeppstedt, Maria, Magnus Ahltorp, Kostiantyn Kucher & Matts Lindström. 2024. From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization* 23(3). 217–238. DOI: 10.1177/14738716241236188.

Söderfeldt, Ylva. 2025. Joint efforts in the Swedish model: The Swedish Diabetes Association under Nancy Eriksson (1956–1978). In Alexander Dunst, Chantal Marazie, Despo Kritsotaki, Matthew Smith & Nicolas Henckes (eds.), *Mobilising medicine: Health and social movements in global perspective, 1950s-2020s*. (In press). Manchester University Press.

Swedish Labour Movement's Archives and Library, Personal archives of Nancy Eriksson. N.d. *SE/ARAB/1611/4/1/4: Data om Diabetesgården i Nordanede; Diabetesgården i Nordanede.*

Torres Parejo, Úrsula, Jesús R Campaña, M Amparo Vila & Miguel Delgado. 2021. A survey of tag clouds as tools for information retrieval and content representation. *Information Visualization* 20(1). 83–97. DOI: 10.1177/1473871620966638.

van der Maaten, Laurens & Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9. 2579–2605.

Viégas, Fernanda B. & Martin Wattenberg. 2008. Tag clouds and the case for vernacular visualization. *Interactions* 15(4). 49–52. DOI: 10.1145/1374489.1374501.

Wang, Yunhai, Xiaowei Chu, Chen Bao, Lifeng Zhu, Oliver Deussen, Baoquan Chen & Michael Sedlmair. 2018. EdWordle: Consistency-preserving word cloud editing. *IEEE Transactions on Visualization and Computer Graphics* 24(1). 647–656. DOI: 10.1109/TVCG.2017.2745859.

Wu, Yingcai, Thomas Provan, Furu Wei, Shixia Liu & Kwan-Liu Ma. 2011. Semantic-preserving word clouds by seam carving. *Computer Graphics Forum* 30(3). 741–750. DOI: 10.1111/j.1467-8659.2011.01923.x.

Xie, Liwenhan, Xinhuan Shu, Jeon Cheol Su, Yun Wang, Siming Chen & Huamin Qu. 2024. Creating emordle: Animating word cloud for emotion expression. *IEEE Transactions on Visualization and Computer Graphics* 30(8). 5198–5211. DOI: 10.1109/TVCG.2023.3286392.

Xu, Jin, Yubo Tao & Hai Lin. 2016. Semantic word cloud generation based on word embeddings. In *Proceedings of the IEEE Pacific visualization symposium*, 239–243. DOI: 10.1109/PACIFICVIS.2016.7465278.

List of abbreviations

TF-IDF	Term frequency–inverse document frequency
t-SNE	t-distributed stochastic neighbor embedding
OCR	Optical character recognition
NLTK	Natural Language Toolkit

Corresponding authors

Maria Skeppstedt
Department of Linguistics
Stockholm University
maria.skeppstedt@ling.su.se

Magnus Ahltorp
Nakajima Koen Research Institute
magnus@nakajimakoen.org

Kostiantyn Kucher
Department of Science and Technology
Linköping University
kostiantyn.kucher@liu.se