Vis4NLP'24 – Visualization for Natural Language Processing (2024)
T. Yousef and K. Alkhatib (Editors)

Using the Word Rain Technique to Visualize Longitudinal Changes in Periodicals from the Swedish Diabetes Association

Maria Skeppstedt[1] and Gijs Aangenendt[1,2]

[1]Uppsala University, Centre for Digital Humanities and Social Sciences, Department of ALM, Uppsala, Sweden
[2]Uppsala University, Department of History of Science and Ideas, Uppsala, Sweden

Abstract

The Word Rain visualization technique is a development of the classic word cloud that aims to retain the simplicity of the word cloud, while at the same time making it possible to use the text visualization for exploring and comparing corpora. We here showcase how the Word Rain technique can be used for visualizing longitudinal changes by displaying the most prominent words for each (user-defined) time period in the corpus. "Most prominent" for a time period can, however, have different meanings depending on how the word rain is configured. We here list a number of possible configuration types, and we also provide examples of word rains generated using some of these configurations. We showcase the configuration settings on a corpus of periodicals from the Swedish Diabetes Association.

1. Introduction

The Word Rain visualization technique is a development of the classic word cloud. The visualization technique was mainly developed with three different use cases in mind: (i) text genre comparison, (ii) expansion of lexical resources [AHES24], and (iii) the study of longitudinal change in text corpora. We have previously presented all three use cases [SAKL24], but the aim of this paper is to dig deeper into one of them; the study of longitudinal change.

The Word Rain, similar to the standard word cloud, is a visualization of the most prominent words in a document. However, what "most prominent" means when visualizing a time series of documents published over a longer time period is not self-evident. The visualizations depend on the configuration used, e.g., on the word frequency cut-offs applied. By varying the configuration settings, different kinds of longitudinal text changes can be shown, which allows the user to explore the corpus from different angles. Using a corpus of periodicals issued by the Swedish Diabetes Association, we here provide a number of configuration examples. For each configuration setting, we treat all periodicals published during a year as one document to visualize and generate a time series consisting of a word rain for each year in the corpus.

2. The Word Rain visualization technique

We refer to a previous publication [SAKL24] for a more thorough description of the Word Rain technique, and will here only provide a summary. The Word Rain addresses two of the limitations of the classic word cloud (i) that the word positioning lacks a semantic interpretation, which can be confusing [BKP14] and which also makes it difficult to explore a word cloud and compare the content of two word clouds, and (ii) that font size is used as the sole indication of word prominence, which might have the effect that longer words are incorrectly perceived as more important [VW08].

By addressing these two limitations, we have aimed to create a visualization that to a larger extent than the classic word cloud can be used as an analytical tool for exploring and comparing text corpora. While such an aim might require a slightly more complex visualization than the classic word cloud, a design criterion for the Word Rain technique was to, as far as possible, retain the simplicity of the classic word cloud. Most importantly, the visualization should be able to fill the role that the word cloud fills today, i.e., it should consist of a static image that can be included in a paper or printed on a poster, and it should not contain any interactive elements other than zooming. To be able to zoom in on the generated word rains is, however, an important feature of the Word Rain design. The word rains are therefore generated with a resolution high enough to make it possible to zoom in and read the words that are too small to be legible when the image is viewed in its original size. Thereby, the Word Rain technique first provides an overview of the most prominent words in the document, and allows the user to zoom in to inspect interesting areas, supporting the "Overview first [...] details on demand workflow" [Shn96]. This functionality is, of course, better supported when digital word rains are used. But stepping closer to inspect details on a poster that are not legible when the poster is viewed from a distance provides a non-digital take on the "Overview first, details on demand" workflow.

DOI: 10.2312/vis4nlp.20241132

The basic principle of the Word Rain technique is to position the words along the x- and y-axes as follows: The x-axis indicates semantics, i.e. words that have a similar meaning are positioned close to each other on the x-axis, while the y-axis is used (as one of the means) to indicate word prominence. The specific techniques used for determining semantics and prominence could be varied. However, our current implementation uses *word embeddings* to determine semantics (x-axis), and *term frequency* or *term frequency-inverse document frequency* to determine word prominence (y-axis). More specifically, in our implementation, the semantic information from a multidimensional word2vec model [CG23] is projected onto one dimension using t-SNE dimensionality reduction [vdMH08]. The x-position of a word is thereby provided by the value in the vector resulting from this dimensionality reduction. This results in that words occurring in similar contexts in the corpus, on which the word2vec model is trained, are positioned close to each other on the x-axis.

Font size is kept as an additional prominence indicator, despite its limitations. Partly because the prevalent use of font size-based word clouds has made this approach somewhat of a standard, but more importantly because the use of different font sizes enables the "Overview first [...] details on demand workflow". The user can locate interesting areas in the word cloud based on words that are displayed with a font size large enough to be legible without zooming in and – based on these words – decide which areas that might be interesting to explore more closely by zooming in. In contrast to the classic word cloud, however, font size is not the *only* prominence indicator. In addition to font size and position on the y-axis, the Word Rain visualization also indicates prominence with a bar chart, where bar heights are proportional to word prominence.

The reason for not relying solely on the position on the y-axis as the prominence indicator to supplement font size, is that the y-axis position only forms an approximate prominence indicator. This is a consequence of the Word Rain algorithm for determining a word's y-position. The words are given their y-position in order of prominence, starting with the most prominent word. The algorithm for positioning the words always attempts to position a word at the *x-position* given by the t-SNE-produced vector, and at the *y-position* 0. If it then collides with a (more prominent) word that has already been positioned close to this coordinate, the new word is moved downwards in the graph until it no longer collides with any of the previously positioned words. (That is, the word "rains down", i.e. a step in the algorithm which gave the visualization technique its name.) Thereby, more prominent words will generally be given a higher y-position than less prominent ones, but low-prominence words might still be given a high y-position, in cases when they do not collide with a word of larger prominence.

It is important to notice that when producing word rains for a collection of documents that are to be compared — as for the case here, when comparing word rains in a time series — *the same t-SNE projection is used for all visualizations*. Thereby, the visualizations share the same semantic x-axis, and this axis can thereby be used for semantic comparisons between different word rains. While there are word cloud extensions specifically created for the task of corpus comparison and/or exploration of longitudinal changes [DES*15, CVW09, LBSW12, LRKC10, WDN13, JBR*18] as well as those that base word positioning on semantics [JS16, HPP*19], we are not aware of any previous approach that uses dimensionality reduction to create semantically comparable word clouds along one dimension.

With these developments of the classic word cloud, we aim to create a visualization that is practically useful when carrying out distant reading [JFSC15] on a corpus too large for a full manual analysis. The Word Rain technique is meant to support an inductive corpus linguistics approach [Ste20], for exploring word (or n-gram)-frequencies when the set of words to explore is *not* known beforehand. For a deductive approach, if the aim is to study the longitudinal frequency change for a set of pre-defined words, it is better to simply plot these frequency changes, for instance using a line graph.

We provide open source code that implements the Word Rain algorithm [CDH24], which we have further developed based on our previous expert review of the technique [SAKL24]. Using the configuration options implemented, we will here explore the different possibilities available for using the Word Rain for visualizing longitudinal changes in a corpus.

3. Configurations for selecting prominent words to show

The word prominence calculations of the Word Rain implementation relies on the TfidfVectorizer class within the scikit-learn programming library [PVG*11]. The user provides a set of folders where the content of each folder is counted as a "document" by the vectorizer. For instance, in the example provided below, we treat each year as one document, and consequently have one folder for each year in the corpus. Many configuration options available for the Word Rain implementation are built on configurations for the TfidfVectorizer class. These are (i) whether to use raw term frequency *(tf)* or term frequency-inverse document frequency *(tf-idf)* as *prominence metric*, (ii) if a stop word list is to be used (i.e. a list of words to be excluded), (iii) the reverse of the stop word list, i.e. an optional list for selecting which words that *are* to be included, (iv) whether to use single tokens or n-grams (where n is configurable), and (v) the *maximum document frequency* and *minimum document frequency*, i.e., the minimum/maximum number of documents in which a word has to occur in the document collection in order to be included in the visualizations.

We have also implemented a number of additional configuration options not available in the TfidfVectorizer class. These include (i) the *minimum number of occurrences in the entire corpus* for a word to be included, and (ii) the *minimum number of occurrences in the document* for a word to be included when creating the visualization for this document. The first option makes it possible to remove words from the visualization that only occur rarely in the corpus. This measure is similar to the minimum document frequency cut-off, but the total number of occurrences of a word is counted, instead of the number of documents in which it occurs. The second option makes it possible to exclude words that only occur a few times in a document. Words that are very infrequent in the corpus as a whole can still receive a high tf-idf value by only occurring a few times in a document. It might therefore be useful to be able to exclude such words from the visualization.

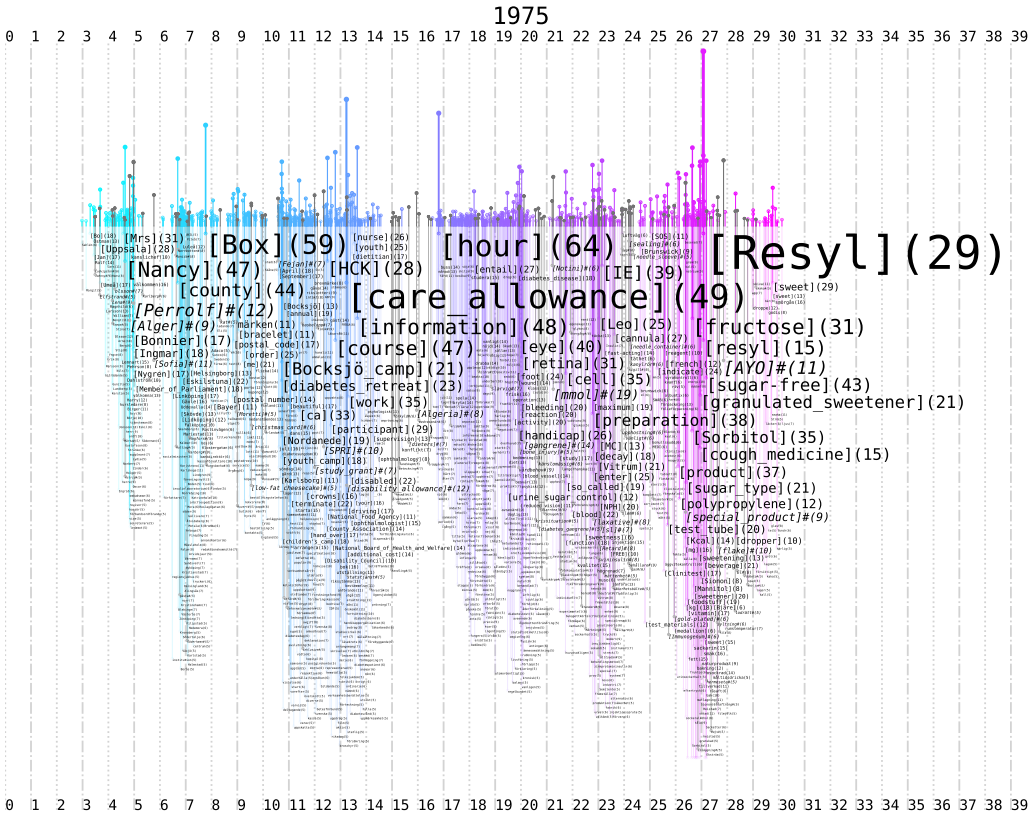

Figure 1: *One year from the series "Time-typical", where the top 1000 most prominent words are visualized. The most prominent words have been translated from the original Swedish into English. The vertical grid and numbers provide guidance when comparing word rains.*

3.1. Incorporating external resources

It is possible to, in several ways, incorporate corpus-external resources when creating the visualizations. The aforementioned "reversed stop word list" (i.e., a list of words which functions as a filter of what words to include in the visualization) can be compiled using external resources, e.g., different types of controlled vocabularies or words included in other corpora. It is also possible to configure the Word Rain implementation to use such word lists, not as inclusion filters, but to emphasize words included in these lists with a contrasting color and by underlining them.

Yet a configuration possibility is to use an external background corpus for calculating the tf-idf value. For instance, to visualize words typical to an IPCC report, in contrast to general language words, we previously used paragraphs from a small subset of "the British National Corpus" as background corpus. Words from the general language, which occur in many documents in the background corpus, receive a high df (document frequency), and thereby a lower tf-idf value than words typical to the IPCC report [SAKL24].

Finally, the word2vec model used for creating the semantic x-axis for the word rain generated can be trained on another corpus than the one visualized. This is a necessity when the corpus visualized is too small to be used for training a word2vec model. However, our user study [SAKL24] showed that the use of an external corpus for creating the semantic axis could lead to confusion. When possible, it is therefore probably better to use a word2vec model trained on the actual corpus visualized, also at the expense of a poorer semantic representation resulting from the corpus used being small.

4. Example word rain configurations

We here provide examples of different types of word rain configurations that might be relevant for studying longitudinal changes in a corpus. Whether a particular configuration is relevant or not depends on the research questions and/or what lens is relevant to apply on the corpus. We have experimented with different configurations that use different word frequency cut-offs and/or filters to include or exclude words. In order to make it easier to remember and discuss them, we have given names to the configurations we currently find most interesting.

Time-typical could arguably be claimed to be a default configuration for exploring tf-idf-based longitudinal change in a corpus. With this configuration, the most over-represented words for each time period are shown. To what extent a word is over-represented in this time era is measured by ranking words according to tf-idf. It might be relevant to provide a stop word list or a maximum document frequency cut-off, since although the potential stop words are likely to be fairly stable over time, the idf value might not be high

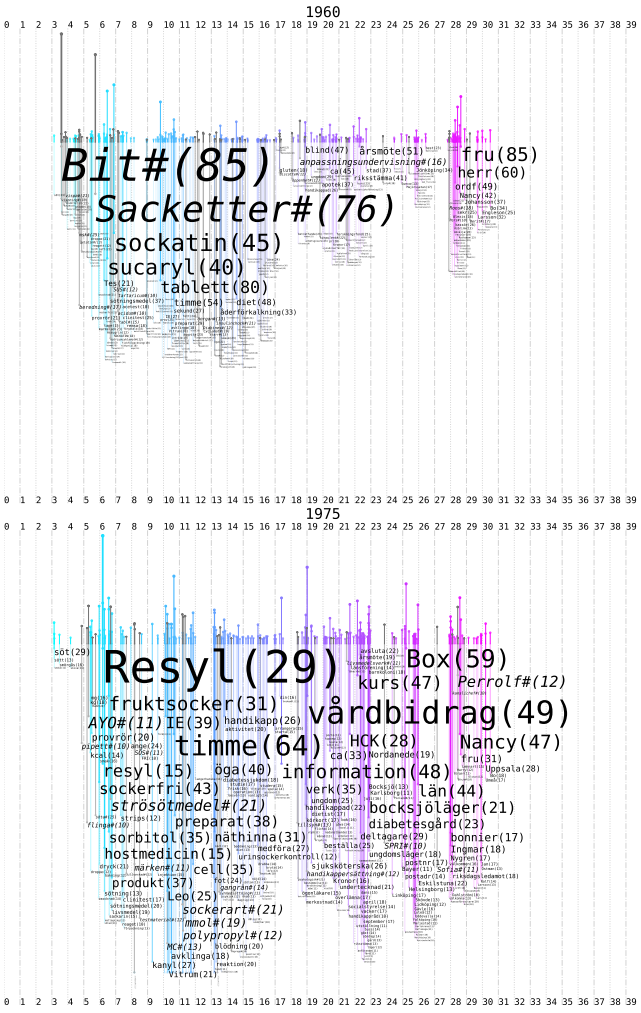

Figure 2: *Examples from two years from the series "Time-typical", where the top 300 most prominent words are visualized.*

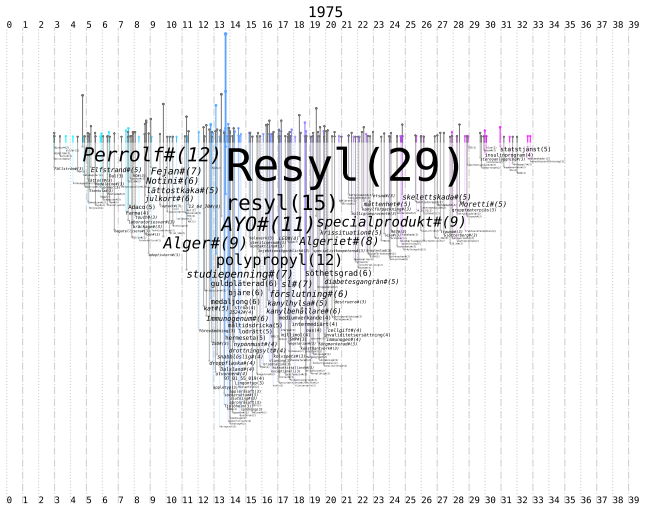

Figure 3: *Example from one year from the series "Mayfly phenomena", i.e., where only words that occur in periodicals from a maximum of five years are included. The top 300 most prominent words are visualized.*

enough to position the potential stop words at a lower ranking than the more "interesting" words. For instance, words that occur in all documents could be removed.

Mayfly phenomena are words that occur only a brief period of time in the whole time period studied. To capture these words, the maximum document frequency parameter can be used. That is, only words that occur in a maximum of x documents (= time periods) are included.

Rather innovative and semi-sticky. This configuration is similar to the mayfly phenomena configuration. But it also aims to exclude words that occur *only* in a few documents. So both the minimum document frequency cut-off and the maximum document frequency cut-off is used, to extract words that do not occur the entire time period studied but that also are not mayflies.

Appears in many time periods is a form of opposite to "Time-typical", i.e. a study of words that occur (more or less) faithfully during the entire time period studied. For instance, the minimum document frequency could be set to half or 75% of the documents in the corpus.

Focus on new words. This configuration focuses on new words and how they develop. We have implemented this configuration by compiling a list consisting of words that occur during the first x years of the time series, and these are then treated as stop words, i.e. words not to include in the visualization. That is, those words are considered to be "old words", to make it possible to focus on the development of the new words that emerge over time in the corpus visualized. These "new words" could, for instance, be studied from the mayfly perspective (with a low maximum document frequency), or it is possible to instead focus on the (semi-)sticky words.

5. Applying some of the configurations on a corpus

We will illustrate the configuration types we find most useful for the current Word Rain implementation by three concrete examples. We showcase the settings using one of the patient organization periodicals studied within the ActDisease project [Act24], namely periodicals from the Swedish Diabetes Association, published from the middle of the 20th century until 1990 [Swe] and processed as described by Aangenendt et al. [ASS24]. Periodicals refer to publications that are issued at regular intervals, and in this case there are typically six issues per year.

We pre-processed the corpus of diabetes periodicals by lemmatising it using efselab [Öst18]. As the model to use for creating the semantic x-axes, we then trained a word2vec model on the corpus using the Gensim library [RS11], with the CBOW algorithm, a window size of 3, and a vector size of 50. Words had to occur at least 10 times in the corpus to be included in the model.

We divided the corpus into years, which meant that all texts published during a year were combined into one document, and treated

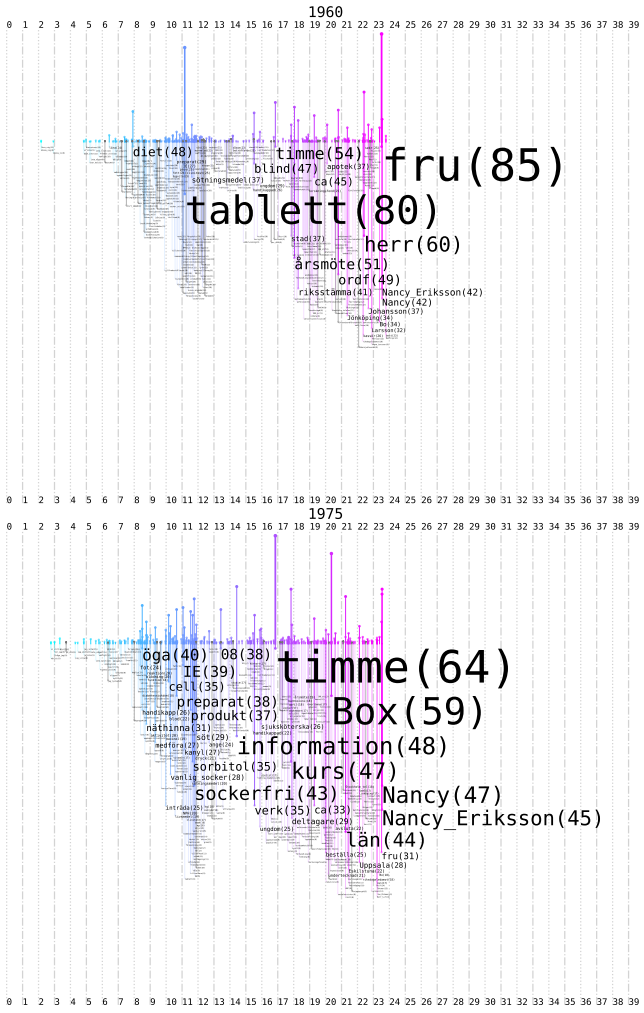

Figure 4: *Two years from "Appears in many time periods". A word has to occur in at least 75% of the years studied to be included.*

as one document. For each configuration option explored, we thus created a visualization series consisting of one word rain visualization per year. Thereby, a cut-off using for example "minimum number of occurrences in a document" is here equivalent to the minimum number of times a word occurs in the periodicals during that year.

For all word rains generated, we applied a cut-off of only visualizing words that occur at least 10 times in the entire corpus. We further configured all word rains to emphasis new words – i.e. words not included in visualizations for previous years – with a gray bar color and "#" following the word. The number within parentheses next to the words shows (the easily interpretable) raw word frequency for the year visualized, regardless if term frequency or tf-idf was used for extracting prominent words. The vertical throughlines and their numbering have been added (as a result of the aforementioned user-study) to facilitate the comparison of word rains. Table 1 summarizes the configuration parameters used.

Time-typical We generated two versions of this configuration. For the first version, we included the *top 300* words in the visualization and applied an inclusion cut-off of at least 10 occurrences during the year (*min occ. in document* = 10). For the second one, we included the *top 1000 words*, and required a word to occur at least 5 times during a year to be included (*min occ. in document* = 5). To remove stop words, we applied a *maximum document frequency* cut-off that excludes words that occur during all years investigated. In addition, we created a stop word list containing numbers and also manually added words referring to the diabetes association itself and to "sockersjuk(a)" (suger-sickness, which was the word previously used for referring to diabetes), as these words took up a lot of space in the visualization without providing any new information. Visualizations for two of the years in the time series are shown in Figure 2 (top 300), and for one year in Figure 1 (top 1000, where the most prominent words have been translated into English). Note that two different semantic t-SNE projections are used for the top 300 visualization and the top 1000 visualization. That is, the horizontal positions for the two visualizations in Figure 2 (which use the same t-SNE projection) can be compared, whereas the visualization in Figure 1 uses another projection.

Figure 1 exemplifies how a lot of information can be encoded into the word rain. The words displayed with a large font give an indication to the semantic content of the region. E.g, (at the vertical lines 4 and 5) there are names of persons – and a few cities – as well as titles/position such as Member of Parliament, (at 7-8) more cities, (at 12) a place for diabetes retreats called Nordande and mentions of different camps/retreats, (at 20-21) body parts and symptoms/diseases, (at 27-28) sweeteners and medicines. Figure 2 exemplifies how two word rains can be compared. For instance, words related to Nordande are there in the 1975 graph (at 23), but are not present in the graph from 1960. Sweetener/medication words (at position 6-7) and food related words (at position 3-4), on the other hand, exist for both years, but the words prominent in 1960 are not necessarily the same ones as in 1975. This shows how the Word Rain technique provides the possibility to compare the content of different years on a level higher than the word level, i.e. on a semantic level.

Mayfly phenomena To achieve a mayfly phenomena visualization, we set the *minimum document frequency* threshold to 1 and the *maximum document frequency threshold* to 5. That is, words that occurred during a maximum of five years were included in the visualization. We also required a word to occur at least twice in the texts for a year for it to be included in the visualization (*min occ. in document*=2). Note that we here still kept the cut-off requiring a word to occur at least 10 times in the corpus as a whole, but to achieve even more focus on the mayfly words, this criterion could be dropped. The result of this configuration for one year can be seen in Figure 3. It can, for instance, be seen (around position 14) that there are many words related to specific beverages this year.

Appears in many time periods Here, we still removed words that occur every year, to remove very generic words, such as "and". But we also used the *minimum document frequency* threshold to require words to occur in at least 75% of the years studied to be included in the visualization. In contrast to the other time-series, we used raw document frequency (*tf*) for ranking the words. Fig-

		# top words	prominence metric	max document frequency	min document frequency	min occ. in document
Figure 1	Time-typical-top-300	300	tf-idf	#documents - 1	1	10
Figure 2	Time-typical-top-1000	1000	tf-idf	#documents - 1	1	5
Figure 3	Mayfly phenomena	300	tf-idf	5	1	2
Figure 4	Appears in many time periods	300	tf	#documents - 1	30	2

Table 1: *Overview of the (cut-off) parameters used for the visualizations. All texts published during a year are treated as a single document. Thereby, e.g. "max document frequency" is equivalent to the maximum number of years in which a word is allowed to be present in the periodicals for it to be included in the visualization.*

ure 4 shows two years with the *top 300* most prominent words. For this configuration, we also showcase the possibility to visualize n-grams. Hardly any n-grams were included among the top 300 words though, with a few exceptions such as "vanlig socker" (regular sugar), "utan socker" (without sugar) and "Nancy Eriksson" (chairperson of the Swedish Diabetes Association between 1956 and 1978 [Söd24]). When comparing the two graphs, we can, for instance, see that "Fru" (Mrs.) and "Herr" (Mr.) are more often used in 1960 than 1975.

6. Discussion and future work

We have here structured and exemplified some of the configuration options implemented for the Word Rain algorithm, which we envision might be useful for studying longitudinal changes in corpora. We have shown that by varying the configuration used, it is possible to highlight different aspects of the corpus content and different types of longitudinal changes.

The set of configurations offered by the current Word Rain implementation is, however, not complete. We have identified a number of other configuration possibilities that might be useful for exploring a corpus that spans over a longer time period. One such extension could consist of providing additional measures than tf-idf for extracting over-represented words for a time period. Likelihood ratio [Dun93] is one such measure, which has previously been used for extracting and visualizing over-represented words (using a classic word cloud) in patient organization periodicals similar to those studied here [SBL19]. Another possible extension consists of providing more advanced functionality for measuring innovation and stickiness. For instance to provide a rolling window for when a word is still to be counted as new, and how many years onward it must be retained in the corpus to be considered sticky (i.e., similar to a previously used innovation measure [KPST21], but applied on word level, rather than on text level). Finally, it could be interesting to construct word rains where discrepancy between two time periods are shown in just a single plot. For instance, it might be relevant to use a prominence measure in the form of the *increase* (or decrease) in word frequency compared to the frequency of the previous time period in the series studied. E.g., the word frequency increase compared to the previous year would be used as the word prominence measure.

There are also possible extensions of the current Word Rain implementation that go beyond additional configuration options for word prominence calculations. One of the conclusions in the aforementioned study on patient organization periodicals [SBL19] was that word frequency statistics is not enough in itself – it was also necessary to read the actual texts to understand the reason why some words are over-represented. To enable the researcher to easily switch between the distant corpus reading provided by the word rain, and a close reading of individual texts could, therefore, be one such future direction. One way of implementing this could be to allow the user to select a subset of visualized words and examine the texts in which they appear. Also for the layout, there are possible improvements. For instance, there is currently a configuration parameter governing how much the font size is decreased for less prominent words. This parameter was here set to optimize the layout for the year 1975 in the examples, but e.g. for the year 1960 in Figure 2, there is space not used in the graph. Either automatic methods that make sure the entire space is used could be investigated, or more granular configuration options for governing the font size decrease could be provided. Another potential useful feature addition would be to also allow word rains with different configuration parameters to share the same semantic x-axis. For instance, it might be practical to let the Figures 2 and 3 share the same semantic x-axis, to facilitate comparisons between the mayfly words and the time-typical ones.

The aim here has been to showcase the possibilities of using the Word Rain technique to visualize longitudinal changes in a corpus, and to suggest different types of configurations for explorations using different lenses. The next step consists of carrying out such explorations, and to investigate whether the configurations suggested here are useful. That is, to explore to what extent the Word Rain visualization can help us understand how the content of patient organization periodicals has evolved over time.

Acknowledgments

We would like to thank Ylva Söderfeldt, as well as the reviewers, for their valuable input on the work presented here.

The work was 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 Word Rain is funded by the Swedish Research Council: Huminfra (2021-00176), InfraVis (2021-00181) and Swe-CLARIN/The National Language Bank of Sweden (2017-00626).

References

[Act24] ACTDISEASE: Actdisease. https://www.actdisease.org, 2024. 4

[AHES24] AHLTORP M., HESSEL J., ERIKSSON G., SKEPPSTEDT M.: 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)* (2024). 1

[ASS24] AANGENENDT G., SKEPPSTEDT M., SÖDERFELDT Y.: Curating a historical source corpus of 20th century patient organization periodicals. In *Proceedings of the Huminfra Conference (HiC 2024)* (2024), pp. 76–82. URL: https://ecp.ep.liu.se/index.php/hic/article/view/895, doi:10.3384/ecp205011. 4

[BKP14] BARTH L., KOBOUROV S. G., PUPYREV S.: Experimental comparison of semantic word clouds. In *Experimental Algorithms* (2014), Gudmundsson J., Katajainen J., (Eds.), Springer International Publishing, pp. 247–258. doi:10.1007/978-3-319-07959-2_21. 1

[CDH24] CDHU, CENTRE FOR DIGITAL HUMANITIES AND SOCIAL SCIENCES, UPPSALA: Word rain. https://github.com/CDHUppsala/word-rain, 2024. 2

[CG23] CUBA GYLLENSTEN A.: *Quantifying Meaning*. PhD thesis, KTH Royal Institute of Technology, 2023. 2

[CVW09] COLLINS C., VIÉGAS F. B., WATTENBERG M.: Parallel tag clouds to explore and analyze faceted text corpora. In *Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology* (2009), VAST '09, IEEE, pp. 91–98. doi:10.1109/VAST.2009.5333443. 2

[DES*15] DIAKOPOULOS N., ELGESEM D., SALWAY A., ZHANG A., HOFLAND K.: Compare Clouds: Visualizing text corpora to compare media frames. In *Proceedings of the 2015 IUI Workshop on Visual Text Analytics* (2015). 2

[Dun93] DUNNING T.: Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*, 1 (1993), 61–74. URL: https://aclanthology.org/J93-1003. 6

[HPP*19] HEARST M. A., PEDERSEN E., PATIL L., LEE E., LASKOWSKI P., FRANCONERI S.: An evaluation of semantically grouped word cloud designs. *IEEE transactions on visualization and computer graphics 26*, 9 (2019), 2748– 2761. 2

[JBR*18] JÄNICKE S., BLUMENSTEIN J., RÜCKER M., ZECKZER D., SCHEUERMANN G.: Tagpies: Comparative visualization of textual data. In *Visigrapp (3: Ivapp)* (2018), pp. 40–51. 2

[JFSC15] JÄNICKE S., FRANZINI G., SCHEUERMANN G., CHEEMA M.: On close and distant reading in digital humanities: A survey and future challenges. a state-of-the-art (star) report. In *Eurographics Conference on Visualization (EuroVis)* (05 2015).

[JS16] JÄNICKE S., SCHEUERMANN G.: Tagspheres: Visualizing hierarchical relations in tag clouds. In *International Conference on Information Visualization Theory and Applications* (2016), vol. 3, SCITEPRESS, pp. 15–26. 2

[KPST21] KELLY B., PAPANIKOLAOU D., SERU A., TADDY M.: Measuring technological innovation over the long run. *American Economic Review: Insights 3*, 3 (September 2021), 303–20. URL: https://www.aeaweb.org/articles?id=10.1257/aeri.20190499, doi:10.1257/aeri.20190499. 6

[LBSW12] LOHMANN S., BURCH M., SCHMAUDER H., WEISKOPF D.: 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* (2012), AVI '12, Association for Computing Machinery, pp. 753–756. doi:10.1145/2254556.2254701. 2

[LRKC10] LEE B., RICHE N. H., KARLSON A. K., CARPENDALE S.: SparkClouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (Nov.–Dec. 2010), 1182–1189. doi:10.1109/TVCG.2010.194. 2

[Öst18] ÖSTLING R.: Part of speech tagging: Shallow or deep learning? *North. Eur. J. Lang. Technol.* (2018). 4

[PVG*11] PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M., DUCHESNAY E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830. 2

[RS11] REHUREK R., SOJKA P.: Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3*, 2 (2011). 4

[SAKL24] SKEPPSTEDT M., AHLTORP M., KUCHER K., LINDSTRÖM M.: From word clouds to word rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization* (2024), 14738716241236188. URL: https://doi.org/10.1177/14738716241236188, arXiv:https://doi.org/10.1177/14738716241236188, doi:10.1177/14738716241236188. 1, 2, 3

[SBL19] SÖDERFELDT Y., BERGLUND K., LINDSTRÖM M.: Towards mining the history of the active patient.: A mixed-methods discourse analysis of the journal allergia, 1957–1990, 2019. 6

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (1996), VL '96, pp. 336–343. doi:10.1109/VL.1996.545307. 1

[Söd24] SÖDERFELDT Y.: Joint efforts in the Swedish model: The Swedish Diabetes Association under Nancy Eriksson (1956-1978) . Department of History of Science and Ideas, Uppsala University, 2024. Manuscript submitted for publication. 6

[Ste20] STEFANOWITSCH A.: *Corpus linguistics: A guide to the methodology*. Language Science Press, 2020. 2

[Swe] SWEDISH DIABETES ASSOCIATION: Diabetes : de sockersjukas tidskrift. https://gupea.ub.gu.se/handle/2077/64597. 4

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605. 2

[VW08] VIÉGAS F. B., WATTENBERG M.: Tag clouds and the case for vernacular visualization. *Interactions 15*, 4 (July 2008), 49–52. doi:10.1145/1374489.1374501. 1

[WDN13] WANG J., DENT K. D., NORTH C. L.: Fisheye word cloud for temporal sentiment exploration. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (2013), CHI EA '13, Association for Computing Machinery, pp. 1767–1772. doi:10.1145/2468356.2468673. 2