1 1/2 years of developing Word Rain

Magnus Ahltorp
Language Council of Sweden
Institute for Language and Folklore
magnus.ahltorp@isof.se

Maria Skeppstedt
CDHU, Department of ALM
Uppsala University
maria.skeppstedt@abm.uu.se

Abstract

The Word Rain visualisation technique is a development of the classic word cloud, providing more possibilities for analysis of the texts visualised. We here briefly describe the work carried out so far, the reasoning behind it, and what could lie ahead.

1 Introduction

The Word Rain text visualisation is a development of the classic word cloud. We will here provide a background to why such a development is needed, explain the new visualisation technique, and summarise the work carried out so far.

The classic word cloud is not only a frequently used method for visualising the most important content of a text (Brath and Banissi, 2016), it is also often used for analytical tasks, despite providing minimal support for such tasks (Hicke et al., 2022). For instance, the word cloud does not support categorisation of prominent words to any greater extent than a simple word frequency list, and comparing two word clouds is very difficult. Further on, a word cloud offers a space-efficient method for displaying the prominent words in a text, but the words which are displayed in a small font mainly function as decoration, since the random (or alphabetical) order in which the words are sorted does not support any meaningful exploration of the word cloud.

Examples of benefits of the classic word cloud include the representation of word prominence (i.e., typically word frequency) by font size, which is an intuitive visualisation that does not require any explanation. It is also a more visually appealing method for providing an overview of a text content than, e.g., a word frequency list or textual summary. In addition, there are several web services that let you produce a word cloud from a text, without requiring any coding, as well as several code packages for producing and customising your word cloud[1]. Finally, the static visualisation type makes the word cloud usable in many situations, e.g., as an article illustration or on a printed poster.

When developing the Word Rain visualisation, we aimed to retain the positive aspects of the word cloud as far as possible, while also finding solutions to the three problems mentioned above associated with the classic word cloud. That is, problems associated with the lack of support for (i) word categorisation, (ii) text comparison and (iii) locating semantically interesting areas when zooming into the visualisation.

2 Word Rain

By positioning the words in a random or alphabetical order, the classic word cloud misses the opportunity to use the word position to convey information. There are previous modifications of the classic word cloud that use position to convey the semantics of the words, by positioning words with a similar meaning close to each other (Xu et al., 2016). There are also previous approaches, where position is used to convey word prominence and which place more important words in the middle of the word cloud (Lohmann et al., 2009). The main innovation of the Word Rain technique is that we do both; we use the *position on the x-axis to convey the semantic information* and the *position on the y-axis to convey word prominence*.

The semantic positioning is currently implemented by using a word2vec-model and creating a one-dimensional t-SNE projection for the vectors corresponding to the words that are to be visualised. The words are then placed in the graph in their order of prominence. The most prominent (e.g., the most frequent) word is positioned first, at the x-coordinate given by the t-SNE projection, and with a y-position at the very top of the graph. Thereafter, the algorithm positions the

[1] E.g., pypi.org/project/wordcloud/, used here.

second most prominent word at its t-SNE-projected x-position. If there is still free space available at the very top of the graph for that x-position, the word is placed there. Otherwise, the word "rains downwards" in the graph, i.e., it keeps the same x-position, but the y-position is decreased until the word reaches an empty spot in the graph. Using this positioning algorithm, the words are continued to be positioned in a decreasing order of prominence. The y-position, therefore, does not strictly follow prominence, as low-prominent words can receive a high y-position if they do not have any more prominent words close-by that make them rain down in the graph. Prominence is, however, also indicated by font size and by the height of vertical bars attached to the words. These bars – which in addition function as markers for semantically important areas – can be configured to be left- or right-aligned (depending on the writing direction).

Figure 1 provides a concrete example of the difference between a classic word cloud and a word rain. The sample stems from a presentation at Språkrådsdagen 2024, where the Word Rain technique was presented (Ahltorp, 2024). The three problems identified above have thus been addressed as follows: (i) the semantic ordering along the x-axis supports the user in identifying word categories among the most important words, (ii) when several word rains are created with the same semantic x-axis, the position on the x-axis can be used for comparing two word rains, and (iii) it is meaningful to include many words (also those displayed with a font size that requires zooming to see), since the prominent words guide the user to which areas might be interesting to explore closer.

Figure 1: A word rain contrasted with a classic word cloud applied on the same data.

3 Development history

At a seminar on the topic of visual text analytics (Collins et al., 2022), we learned about the need for a text visualisation technique that (a) could be used in the same contexts as word clouds are used today, and (b) is as easy to produce and interpret as word clouds, but also (c) would be more useful for analytical tasks.

Our first attempt in producing such a visualisation focused on the task of comparing texts, and we initially created a visualisation consisting of graphs with words positioned through the more traditional use of t-SNE projections: projecting word2vec-vectors corresponding to prominent words onto a 2-dimensional space (Skeppstedt et al., 2022).

Although this visualisation functioned well for the task at hand, it did not meet our requirements for a more generally applicable development of the word cloud. In particular, it was difficult to achieve a word positioning that was both consistent between the different graphs, and that was also sufficiently space-efficient and made sure the words did not collide. For our next attempt, we therefore created what is now the Word Rain visualisation, where these problems were solved by instead using a one-dimensional projection. The word positioning could then be consistent between different graphs by always using the same x-coordinate for a word. By moving less prominent words downwards in the graph until free space is found, we avoided collisions, while also achieving a word prominence ordering along the y-axis.

We presented the first version of the Word Rain at the DHNB conference 2023. The theme of the conference was "Sustainability: Environment, Community, Data", and in our presentation we applied the Word Rain visualisation technique to IPCC reports (Skeppstedt and Ahltorp, 2023).

We then continued the work on the technique following three separate paths. The first was more theoretical and focused on fine-tuning the visualisation, improving code quality, and performing user evaluations of the technique. For this path, we collaborated with the iVis group at Linköping University. We here continued with the climate change theme, and applied Word Rain to three different

tasks related to texts on the topic of climate change (Skeppstedt et al., 2024).

The second path consisted of – together with other researchers – using the Word Rain visualisation to study longitudinal changes in corpora. This is partly carried out in the form of CDHU's[2] collaborative pilot projects which aim to increase interest in digital methods among researchers within digital humanities and social sciences (Skeppstedt and Rosenbaum, 2023; Skeppstedt and Ahltorp, 2024), and mainly within the ActDisease project (Skeppstedt and Aangenendt, 2024). The ActDisease project is an interdisciplinary ERC project, where mixed methods are used for studying patient organisations during the 20th century. In the project, digital humanities methods are applied to publications issued by patient organisations[3].

Finally, the third part consists of collaborating with developers of dictionaries, studying to what extent Word Rain can be used for assessing the coverage of a dictionary in relation to different text genres (Ahltorp et al., 2024). The three paths have led to small and iterative improvements of the Word Rain technique, for instance more parameters that make it possible for the user to customise the word rains. These include the rate with which the font size is to be decreased for less prominent words, the possibility to only include a set of pre-defined words in the visualisation, the possibility to slightly customise the t-SNE projection, and the possibility to let the user define their own word prominence measure function, in addition to the built-in measures term frequency and tf-idf.

We are currently in the process of writing a handbook chapter about the Word Rain technique, mainly focused on how it can be used in digital history (Skeppstedt et al., 2025). For this handbook chapter, we target three different types of users with an increasing level of programming skills and/or need to customise the visualisation. For the first type of user, with a limited need for customisation, we have created and published a web interface at `https://wordrain.isof.se/`, which makes it possible to simply upload texts to create word rains (Ahltorp and Skeppstedt, 2024). The code for the Word Rain web service is also available as open-source, to make it easy for anyone familiar with web services to set up their own word rain generation interface[4]. For a user who wants to be able to configure their word rains to suit their particular visualisation requirements, there is open-source Python code available on GitHub for generating word rains[5].

Figure 2 forms an example of a word rain generated by the Word Rain web service. We have here taken a partial transcript (Hoffman, 2024) of the 10 September 2024 US presidential election debate between Kamala Harris and Donald Trump, separated it into one text file per speaker, and uploaded it to the service. Since both files were uploaded together, they share the same x-axis, making it possible to compare the two graphs. For instance, to the very right – in orange – both graphs contain person identifiers, which makes it possible to compare which persons the two speakers mention most frequently. Similarly, moving a bit to the left into the area with bars in magenta, frequently mentioned countries can be compared. There are also examples of word clusters/categories frequently used by one of the speakers, but not by the other. For example, there are two light-blue clusters that are scarcely populated in the graph for Donald Trump, which contain frequent words used by Kamala Harris; words such as "*support/values/dreams/responsibility*", and "*young/woman/child/mother*".

4 Summary and future challenges

There are still many aspects in which the Word Rain visualisation could be improved and expanded, and there are still many open research questions connected to the technique. One open research question regards the relationship between word frequency-based text visualisations and the possibilities offered by the ever-growing language models. We hypothesise that the word frequency-based visualisations provide another, perhaps complimentary understanding of the text content, than for instance a short textual summarisation of a corpus. However, this is only a hypothesis that should be tested in future studies.

The other open question regards the usage of the word rains. The basic aim of the visualisation is fundamentally the same as that of a word frequency list, or a set of word frequency lists: to let the user know which are the most frequent words,

[2] Centre for Digital Humanities and Social Sciences, Uppsala University

[3] `https://www.actdisease.org`

[4] `https://github.com/sprakradet/wordrain-service`

[5] `https://github.com/CDHUppsala/word-rain`

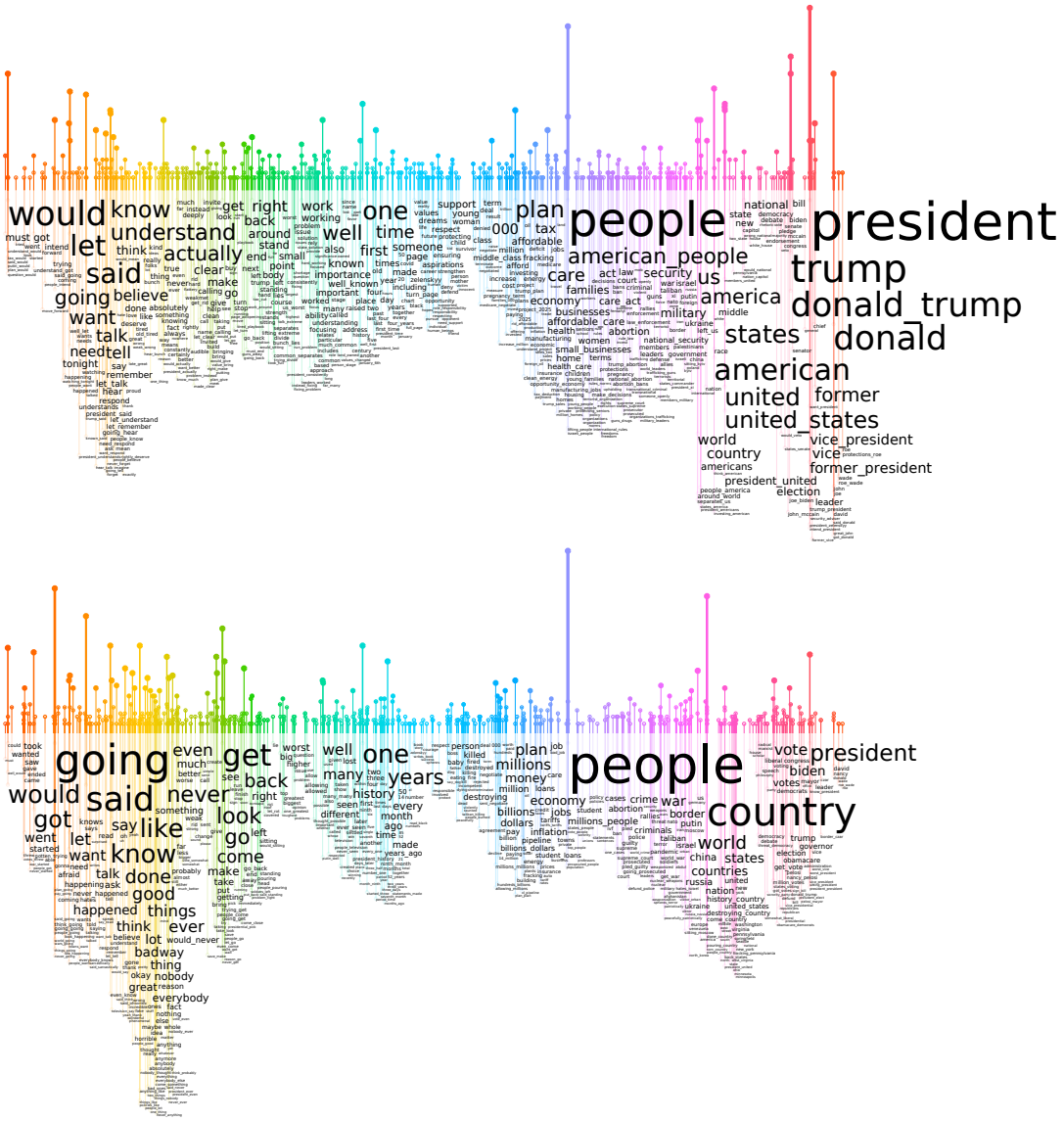

Figure 2: A word rain of the 10 September 2024 Harris–Trump US presidential election debate generated by the Word Rain web service. Kamala Harris is the upper word rain and Donald Trump is the lower.

and how frequencies differ between corpora belonging to different genres and/or how frequencies vary over time. The semantic word positioning is only meant to support and simplify these tasks. However, since the words are semantically ordered, it is also possible to use a word rain to study to what extent different words are paradigmatically similar in the corpus used for training the word2vec model. The problem then is that much information regarding the semantics is lost when a multi-dimensional space is projected down to only one dimension. Therefore, while it is possible to study semantic word similarity with the Word Rain visualisation, it is not an optimal visualisation for this task. A static visualisation optimised for word similarity is thus also needed.

However, despite there still being many ways in which the technique could be developed, we believe we have moved the niche of word frequency-based, static text visualisation a step forward during these 1 1/2 years when we have worked on the Word Rain technique. The Word Rain text visualisation might not be the one that replaces the classic word cloud as the more analytically powerful text visualisation alternative. However, we do believe that whoever creates this replacement will be able to find inspiration in our work on Word Rain.

Acknowledgments

The development of Word Rain as a digital humanities research infrastructure resource is funded by the Swedish Research Council: Huminfra (2021-00176), InfraVis (2021-00181) and Swe-CLARIN/The National Language Bank of Sweden (2017-00626).

The ActDisease project 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.

References

Magnus Ahltorp. 2024. Ordregn – visualisering av klimatprat. *Utbildningsradion*.

Magnus Ahltorp, Jean Hessel, Gunnar Eriksson, and Maria Skeppstedt. 2024. 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)*.

Magnus Ahltorp and Maria Skeppstedt. 2024. Word rain as a service. In *Proceedings from the CLARIN Annual Conference 2024*.

Richard Brath and 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.

Christopher Collins, Antske Fokkens, Andreas Kerren, Chris Weaver, and Angelos Chatzimparmpas. 2022. Visual Text Analytics (Dagstuhl Seminar 22191). *Dagstuhl Reports*, 12(5):37–91.

Rebecca M. M. Hicke, Maanya Goenka, and Eric Alexander. 2022. Word clouds in the wild. In *2022 IEEE 7th Workshop on Visualization for the Digital Humanities (VIS4DH)*, pages 43–48.

Riley Hoffman. 2024. Harris-Trump presidential debate transcript. https://abcnews.go.com/Politics/harris-trump-presidential-debate-transcript/story?id=113560542. ABC News. Retrieved 2024-09-11.

Steffen Lohmann, Jürgen Ziegler, and Lena Tetzlaff. 2009. Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Human-Computer Interaction – INTERACT 2009*, pages 392–404, Berlin, Heidelberg. Springer Berlin Heidelberg.

Maria Skeppstedt and Gijs Aangenendt. 2024. Using the Word Rain Technique to Visualize Longitudinal Changes in Periodicals from the Swedish Diabetes Association. In *Vis4NLP 2024 - Workshop on Visualization for Natural Language Processing*. The Eurographics Association.

Maria Skeppstedt and Magnus Ahltorp. 2023. The words of climate change: Tf-idf-based word clouds derived from climate change reports. In *DHNB2023 Book of Abstracts: Sustainability: Environment, Community, Data*.

Maria Skeppstedt and Magnus Ahltorp. 2024. Using topics2themes and word rain to visualise topics in Swedish news on climate change. In *Proceedings from the CLARIN Annual Conference 2024*.

Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, Gijs Aangenendt, Matts Lindström, and Ylva Söderfeldt. 2025. The word rain visualisation technique applied on digital history: How to visualise, explore and compare texts using semantically structured word clouds (abstract accepted). In *Huminfra Handbook*. Huminfra.

Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, and Matts Lindström. 2024. From word clouds to word rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization*, page 14738716241236188.

Maria Skeppstedt, Gunnar Eriksson, Magnus Ahltorp, and Rickard Domeij. 2022. Den ena texten och den andra: Visualisering av textpar med hjälp av ordmoln. In *LIVE and LEARN - Festschrift in honor of Lars Borin*.

Maria Skeppstedt and Paul Rosenbaum. 2023. Text mining on applications for cultural funding. https://www.clarin.eu/content/clarin-bazaar-2022. The CLARIN Bazaar 2023. https://www.clarin.eu/content/clarin-bazaar-2023.

Jin Xu, Yubo Tao, and Hai Lin. 2016. Semantic word cloud generation based on word embeddings. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pages 239–243.