

Multilingual Word Rains: Text Visualisation Built on Cross-Language Word Similarities

Magnus Ahltop

Nakajima Koen Research Institute

Stockholm, Sweden

`magnus@nakajimakoen.org`

Maria Skeppstedt

Department of Linguistics

Stockholm University, Sweden

`maria.skeppstedt@ling.su.se`

Abstract

In this study, the text visualisation technique Word Rain is extended with functionality for generating multilingual word rains. Using multilingual pre-trained word embeddings provided by the MUSE library, we show that cross-language studies on lexical frequency with Word Rain are possible in the same way as has previously been possible for monolingual corpora. We make the programming code for generating multilingual word rains available, and we also provide a web page where visualisations can be produced by simply uploading text files that the user wants to visualise. To illustrate the technique, a tentative study of common word groups and cross-language frequency similarities and differences is performed on debates from three parliaments included in the CLARIN ParlaMint corpus.

1 Introduction

The standard word cloud can provide a quick overview of a text or corpus, in the form of a graph that displays the most frequent words in a font size that corresponds to their frequency (Cao & Cui, 2016, pp. 57–59). However, the random (or sometimes alphabetical) word positioning of the standard word clouds makes them unpractical for a deeper exploration of frequent words, and even more so for a comparison of word frequencies between different word cloud graphs.

Despite its limitations, the standard word cloud is often used for analytical purposes, e.g., when studying and comparing texts within digital humanities (Hicke et al., 2022). We have therefore previously developed the Word Rain visualisation technique (Skeppstedt et al., 2024), within the CLARIN and Dariah infrastructures. The aim of the development was to create a visualisation technique that would retain the simplicity of the standard word cloud, while also making the visualisation suitable for exploring word frequencies and for comparing frequencies between texts. Similar to the classic word cloud, the Word Rain technique extracts the most frequent words from a text/corpus, and displays them with a font size corresponding to their frequency. In contrast to a word cloud, the Word Rain technique automatically orders the words along the x-axis in correspondence to their meaning. This causes words with a similar meaning to be positioned horizontally close to each other in the word rain graph.

The semantically motivated word positioning of the Word Rain technique helps the user to identify semantic categories among the frequent words, as well as to locate areas in the graph that might be relevant to explore more closely. The semantic positioning also simplifies comparison between two texts, since several word rain graphs can be generated using the same semantic word positioning along the horizontal axis. Word rains can either be generated by employing the Python programming code in the Word Rain code repository¹, or by uploading texts to the Word Rain Web Service at `wordrain.org` (Ahltop & Skeppstedt, 2025). When generating word rains using the programming code, there is a wider range of configuration possibilities available, such as the possibility to change colour scheme or to employ user-defined stop word lists². The Word Rain Web Service, on the other hand, does not require the effort of downloading (and knowledge of how to use) Python code.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹<https://github.com/CDHUppsala/word-rain>

²We refer to Skeppstedt et al. (2025) for a detailed description of the customizability of the Word Rain programming code.

We have, previously, aimed at providing solutions to use cases that include the exploration of monolingual corpora. However, for many text analysis tasks, it is also relevant to be able to visualise, study and compare the content of corpora containing texts written in several different languages. The main aim here is therefore to provide and describe our resources for making it possible to create multilingual word rains that can be used in such studies. More specifically, we have extended the Word Rain Web Service, to make it possible to upload corpora containing texts in several languages, and also provided the Web Service with a few pre-defined language options for such a multilingual text exploration and comparison. In addition, we are making a new version of the Python code for generating word rains available. It is now able to handle multilingual corpora for all combinations of languages, as long as aligned pre-trained language models are available for the language combinations. Finally, we have applied the updated code in a proof-of-concept study, where the Word Rain technique is used for visualising lexical similarities and differences in the CLARIN ParlaMint corpora from the United Kingdom, France and Austria.

2 Background

Visualisation of corpora across languages can be approached in several different ways. One method is to create visualisations and conduct the analysis for texts in each language separately and then compare the analyses. This approach was, for instance, taken for analysing Hungarian and Slovenian parliamentary corpora (Sárossy & Ligeti-Nagy, 2024). The authors divided each corpus into two subcorpora (coalition MPs and opposition MPs), and then compared frequent keywords, as well as graphs with collocations containing these keywords. A similar approach was taken by Söderfeldt et al. (2025), who created, analysed and compared four separate topic modelling-based visualisations of patient organisation publications written in English, French, German and Swedish.

When it comes to approaches for including texts written in different languages within the same visualisation, there are a number of solutions for how to visualise and compare an original text with its translation into another language (Yousef & Janicke, 2021). However, for creating visualisations of text collections containing non-parallel texts in several different languages, less work has been carried out.

A possible method for handling visualisations of multilingual text collections is to first automatically translate all texts into a single language, before creating a visualisation. However, such an approach would risk hiding differences between languages, i.e., differences that might be one of the aims of the analysis to find. For instance, there might be closely related concepts where each one of them is described by a separate term in one of the languages, whereas a translation into a language with less linguistic specificity for these concepts could hide the variation in the original language. We, therefore, aimed for a solution to a multilingual Word Rain visualisation, where the words would be retained in their original language, but where the word rain graphs would still be easily comparable across languages.

The semantic word positioning along the x-axis in a word rain graph is based on word embeddings. Multi-dimensional word embeddings, which have been trained on word co-occurrence patterns in a monolingual text corpus, are projected down to one dimension, to determine a word's x position in the graph. Thereby, words that occur in similar contexts in the training corpora are positioned close to each other on the x-axis in a word rain graph.

A commonly used method for handling multilinguality when working with word embedding models is to use aligned models. Then, one embedding model per language is still used, but the models are aligned with each other, so that similar words across languages have similar embeddings. For instance, in two standard word embedding models for German and French, the words *Präsident* in German and *président* in French (both meaning 'president') have word embeddings that are totally unrelated. In contrast, if the two embedding models are properly aligned, they have similar embeddings. Therefore, if using the embedding in the German model for the word *Präsident* when querying for similar words in the French model, the word *président* should be among the retrieved ones.

Word embedding models can be aligned between languages in a supervised fashion, by using content-aligned multilingual text corpora (such as Wikipedia) (Søgaard et al., 2015) or by using multilingual dictionaries (Lample et al., 2018) as data for training. It is also possible to achieve aligned word embeddings in a fully unsupervised fashion (Lample et al., 2018).

3 Method

Most of the updates required for handling multilinguality was carried out in the code for generating word rains. Some updates were also needed for the Word Rain Web Service code repository.

3.1 Updating the functionality for creating word rain visualisations

The previous implementation of the Word Rain visualisation technique needed to be extended in two main ways to handle multiple languages. First, we added the ability to use different stop word lists for each sub-corpus, instead of just using one stop word list for the entire corpus. Second, we added the ability to use different word embedding models for each sub-corpus. These changes enable us to use embedding models that match the language of each sub-corpus included in the visualisation. For instance, for the proof-of-concept study described below, we could use English embeddings for the UK data, French embeddings for the French data, and German embeddings for the Austrian data³.

In addition, we added a name format requirement. For generating word rains, the user needs to provide a file path to a main folder, containing the corpus that is to be visualised. This main corpus-folder should contain subfolders, which, in turn, contain the actual texts. The subfolders represent a sub-division of the main corpus into sub-corpora, and one word rain graph is generated for each such sub-corpus. For instance, when we previously have generated word rains for analysing changes over time in a corpus, we have created subfolders based on the publication year of the texts, and thus generated one word rain graph per year (Skeppstedt et al., 2025). As a means to indicate the language of the texts in a subfolder, we added the requirement that each such subfolder must be prefixed by a language code, followed by a hyphen. Name of subfolders could, thus, for instance be `en-house_of_lords` and `en-house_of_commons`.

3.2 Updating the Word Rain Web Service

A number of updates were also required for the Word Rain Web Service, in addition to connecting it to the new version of the code for generating word rains.

As described above, to create graphs with a word positioning aligned between languages, we need word embeddings that are themselves aligned. We therefore added aligned word embeddings provided by the MUSE library (Lample et al., 2018). These pre-trained embeddings have been aligned in a supervised fashion, using multilingual dictionaries. As a first step, we added three aligned embedding models to the web service, for French, English, and German.

We created a separate user interface for multilingual aligned word rains, and made it available at: <https://wordrain.org/multilingual> (see Figure 1). As for the monolingual word rains, the user selects texts to visualise by uploading one or several texts to the web page, (i.e., not folders, as when using the code directly). In contrast to the monolingual service, there is, however, no drop-down list for selecting language. Instead, all uploaded files must be prefixed with a language code matching one of the languages of the aligned embedding models, followed by a hyphen. That is, currently with either `de-`, `en-` or `fr-`.

Similar to the monolingual service, the user can either upload plain text files, or OOXML (.docx) files. After the text files have been uploaded and processed, the user can adjust two parameters in the graphical presentation: i) the height ratio between the part of the graph containing the words and the part containing the vertical bars (see section 4.1), and ii) the rate with which the font size should decrease with decreased word frequency. (More specifically, the decrease rate is governed by a power function, for which the exponent is configured by the user.) Finally, a word rain graph is generated for each uploaded file, with an aligned semantic axis between the graphs.

4 Proof-of-concept study

To illustrate the functionality of the multilingual Word Rain technique we performed a proof-of-concept study.

³The extended code can be found at <https://codeberg.org/ahltorp/wordrain/src/branch/multilingual>

Word Rain

Multilingual aligned Word Rain

This page will take multiple text files in different languages and generate a Word Rain visualization where the words are aligned between the languages.

The currently supported languages are:

- English (en)
- French (fr)
- German (de)

The names of the uploaded files have to start with the language code of the language in the file (en, fr, de) and a hyphen "-".

Examples:

- en-house-of-commons.txt
- en-house-of-lords.txt
- de-austrian-national-council.txt
- fr-french-national-assembly.txt

Choose number of words to plot:

300

Upload either:

- text files (plain text, UTF-8)
- Word .docx files

 Upload files



中島公園研究所
Nakajima Koen
Research Institute

The Word Rain service is published by Nakajima Koen Research Institute.

Contact us at:
info@wordrain.org

Source code for Word Rain is available at <https://github.com/CDHUppsala/word-rain>.

Source code for the Word Rain service is available at <https://github.com/sprakradet/wordrain-service>.

The development of the Word Rain service has been partly funded by [CLARIN](#).

Figure 1: The multilingual aligned Word Rain page at wordrain.org/multilingual.

For this study, we used the ParlaMint corpus. ParlaMint is a corpus collection of debates from 29 parliaments across Europe. We used the ParlaMint (Erjavec et al., 2024) corpora ParlaMint-GB, ParlaMint-FR and ParlaMint-AT from the resource repository CLARIN.SI. We extracted the data for the 2017 parliamentary debates in .txt format. For 2017, ParlaMint-AT had 1 699 623 tokens, ParlaMint-FR had 4 680 332 tokens, and ParlaMint-GB had 14 573 789 tokens when processed through Word Rain’s default tokeniser. We treated all texts from each of the languages as one separate sub-corpus, and thus generated one word rain graph for each language.

The graphs were generated using the code in the Word Rain Python code repository. We configured the graph generation to transform all words into lower case before computing word frequencies. The number of words per word rain was set to the 500 most frequent words, and we used the standard NLTK stop word lists for each of the three languages. The functionality to automatically identify frequent n-grams was not enabled.

Font size (and thereby also bar height, see below) was configured to be normalised between graphs so that the font size for the most frequent word was the same in all three graphs. Frequencies, as indicated by font size/bar height, are therefore comparable within graphs but not across graphs, while the frequency relations between words are comparable across graphs. As mentioned above, the font size/bar height decreases with decreased word frequency at a rate governed by a configurable power function. Here, it was configured to decrease linearly.

4.1 Resulting Word Rain graphs

The resulting word rains can be seen in Figure 2. In the same way as in a traditional word cloud, the words have a larger font size if they are more frequent. Since the words are semantically ordered along the horizontal axis according to their positions in the word embedding models, words with similar meaning are placed together horizontally with a high probability. With the modifications carried out in this study, this is now also the case for semantically related words in different languages.

The bar emanating from the faint circle in the upper left corner of each word forms an additional indication of relative word frequency, as the height of the non-transparent part of a bar directly corresponds to the font size of the word. However, since it might be difficult to determine exactly which bar corresponds to which word in areas with many words, the main purpose of the bars is to provide an overview of the frequencies present in a given area, rather than to indicate the frequency of each individual word. Finally, the colours illustrate the horizontal position of the words, and are chosen to make it easier to compare horizontal word positions between graphs.

As an example of similar words being placed closely, the words *Präsident* ‘president’ in German and *président* in French are very close to each other on the horizontal axis (to the left, in orange). The same is true for *wissen* in German, *savoir* in French and *know* in English (to the right, in magenta). This indicates that the word embeddings are indeed fairly aligned.

When exploring and comparing the three word rain graphs, we can first turn our attention to the green part, where there is an area with a high frequency in all graphs around German *Herr*, French *monsieur* and English *hon*. The words in that area do not mean the same thing, but if we study the common words more closely, we can see a pattern. In Austria, they use several forms of *Herr* ‘Mr’, *Frau* ‘Ms’, *Kollege* ‘colleague’ and *geehrte* ‘honourable’. In France, they use *monsieur* ‘Mr’, *Mme* ‘Ms’, *madame* ‘Ms’ and *collègue* ‘colleague’. In the UK we find *hon*, *noble*, *lord*, *lady*, *gentleman*, *Mr*, *baroness*, *colleagues* and *friend*. These are ways of addressing other members of parliament and other persons, for example government ministers, and we can see the influences of the House of Lords in the UK, where the titles baroness and lord are used.

In the orange part of the graphs, there is another cluster of words denoting persons. These are not persons being addressed, but instead roles in the parliament, such as German *Präsident/Präsidentin* ‘President, masculine and feminine form’, *Abgeordneter* ‘Member of Parliament’, French *ministre* ‘minister’, *rapporteur* ‘manager of a bill’, and English *minister*, *secretary*, *chancellor*.

In the part of the graph where green turns into yellow, there is yet a cluster denoting persons. However, this cluster seems to denote persons in general (not specific to the parliament). For all three languages,

[illegible][illegible][illegible]

there is a frequently occurring word meaning ‘persons in general’, i.e. *Menschen*, *personnes* and *people*. These frequent words are then accompanied with closely positioned, less frequent ones, that denote more specific categories of persons. Some of these categories occur among the top 500 most frequent words for all three languages, such as words for ‘children’ (*Kinder*, *enfants*, *children*) and ‘young’ (*jungen*, *jeunes*, *young*). Other examples of words in this cluster include: German *Bürger/Bürgerinnen* ‘citizen’, *Bevölkerung* ‘population/people’, *Frauen* ‘women’, French *concitoyens* ‘fellow-citizen’, *étudiants* ‘students’, and English *citizens*, *families*, *students*, and *women*.

All three clusters denoting persons provide examples of the advantages of keeping the original languages instead of using an automatic translations. Examples of nuances that might have been lost in the translation are the feminine and masculine forms of words, the exact English titles, as well as the exact terms used for the professional roles within the three parliaments.

We have so far focused on similarities in the form of semantic categories that are frequent in all three graphs, (and on differences within these semantic categories). There are, however, also differences between the three corpora, when it comes to which words and semantic categories that are frequent. For instance, the very high frequency French word *parole* ‘speech’ is less frequent in the other languages. The word has corresponding words on the same position on the horizontal axis in English (*speech*) and in German (*reden/Reden* ‘speak, speeches’), but they are not nearly as frequent.

We can also observe a larger difference, if we instead look at the far left of the word rain graphs, where there is a high frequency area (in orange) that is only present in the Austrian data, with one of the most prominent words being *Österreich* ‘Austria’. This word, and the various forms of it, is – perhaps naturally – not among the 500 most frequent words in either France or the UK. There are also names of political parties in Austria: *ÖVP*, *FPÖ*, *SPÖ*, *Grünen* and *NEOS*.

A difference where the words are present in the other languages but with very different frequencies is *cette* ‘this’ in French and *would* in English. The first is probably explained by it not being in the stop word list for French, even though it is the feminine form of the masculine *ce*, which is present in the stop word list. The second can perhaps be a real difference in language use, since constructions with *would* like *I would disagree* and *I would urge* seem to be very common, and maybe even used as something close to a discourse particle rather than having its usual syntactic function, but further studies would be needed to draw any solid conclusions.

As a last example, *amendement* in French has its counterpart in English *amendment* at the same horizontal position, but is used much more frequently in French. This is probably because *amendement* is a very important term in French lawmaking.

4.2 Discussion

A general problem when comparing the most frequent words between texts, is that more specific terms might be used for related concepts in one text, while another text might use a single term for all these related concepts. An incorrect conclusion might then be drawn that this concept is an important one in the second text (as it appears among the most frequent terms), whereas it is not important in the first. As mentioned in the background, this difference in linguistic specificity is also present across languages. A term used in one language might correspond to several terms in another language even when the terminological domain is the same. When using a word rain instead of a standard frequency list, however, vertical clusters containing related concepts are created. Entire clusters of related words can thereby be compared between texts, instead of comparing frequencies for individual words. For instance, in a standard word frequency list, it would look like the use of formal titles to address people would be more frequent in the German texts than in the English ones, since *Herr* ‘Mr’ is very common (and *Frau* ‘Mrs’ less, but still rather, common). However, when looking at the vertical title cluster in the English word rain (i.e., containing *hon*, *noble*, *lord*, *lady* ...), it is clear that titles are very common also in the English texts, but that there is a larger diversity among the titles used.

The standard NLTK stop word lists used were not ideal for the task, and should be amended to include common words that are not relevant to the analysis, such as *cette* ‘this’. One word of caution, however, is that some commonly used words, such as *would*, might be relevant for some types of analysis. The

resulting word rain could probably also have been improved by using specifically Austrian stop word lists and word embedding models instead of the generic German used in this study.

The pre-existing aligned word embedding models used for this study make no distinction between upper and lower case characters (Lample et al., 2018), forcing us to transform all words into lower case. This causes pairs of words having initial upper case (proper nouns) or lower case (other words) to be conflated in all of the three languages. This is, however, especially problematic for German, where all nouns use initial upper case in the standard orthography. An example of two conflated German words that appear in our visualisation is the verb *reden* ‘speak’ and the noun *Reden* ‘speeches’ (both forms have frequency > 20 in the corpus).

5 Conclusions and future work

We have been able to successfully extend Word Rain to produce multilingual word rains, to support comparison of word frequencies in corpora containing texts in different languages. For this functionality, we have also made programming code, as well as the Word Rain Web Service, available, which makes it possible for everyone to generate multilingual word rains. That is, we have provided a multilingual text visualisation tool both for those who need the flexibility of the programming code as well as for those who want to be able to upload texts to produce visualisations with pre-configured settings. Finally, we have used the CLARIN ParlaMint resource for a proof-of-concept study that tentatively investigates what words are commonly used in the parliaments of different countries based not only on raw frequency lists, but grouped according to cross-language semantic information.

In the proof-of-concept study, we only used three of the 29 corpora in ParlaMint, partly due to limited time and space, but also because it was necessary for the authors to understand the languages to be able to validate the quality of the resulting alignments. The study has also been limited by not having co-authors that are experts in parliamentary debates and political science. We therefore hope future studies using this technique can be made together with authors from various parts of Europe and from different scientific fields. We also plan to apply the Word Rain technique on multilingual corpora from other domains, and in collaboration with domain experts analyse differences between texts written in different languages.

Finally, for the monolingual use cases of the Word Rain technique, many word rains have been produced using a word2vec model trained on the corpus that is visualised, instead of using pre-trained word embedding models. The words are then not positioned along the x-axis based on distributional similarities in a large general corpus, but instead based on distributional similarities in the specific corpus visualised. Custom word embeddings would also enable other types of equivalence criteria for words, such as lemmatisation or distinguishing upper-case-initial words from their lower-case-initial counterparts. Future studies on custom word embedding models in multilingual Word Rain visualisations should also include quality evaluations of different kinds of aligned models.

While there are easy-to-use libraries for training standard word embedding models on a specific corpus, e.g. the Gensim library (Řehůřek & Sojka, 2010), we are not aware of any programming library that is equally easy to use for aligning word embedding models trained on different languages. To simplify the alignment process – and thereby also simplify the generation of multilingual word rains where the word positions reflect the specific distributional similarities in the corpora visualised – is therefore yet a project for future work.

Acknowledgments

Part of the work presented here was conducted within Swe-CLARIN, which is funded by the Swedish Research Council (dnr 2023-00161).

References

Ahlthrop, M., & Skeppstedt, M. (2025). Word Rain as a Service: Making semantically structured word clouds available to everyone. In V. Vandeghinste & T. Kontino (Eds.), *Selected papers from the CLARIN Annual Conference 2024*. <https://doi.org/https://doi.org/10.3384/ecp216.03>

- Cao, N., & Cui, W. (2016). *Introduction to text visualization* (Vol. 1). Atlantis Press. <https://doi.org/10.2991/978-94-6239-186-4>
- Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., Calzada Pérez, M., ... Fišer, D. (2024). Multilingual comparable corpora of parliamentary debates ParlaMint 4.1 [Slovenian language resource repository CLARIN.SI]. <http://hdl.handle.net/11356/1912>
- Hicke, R. M. M., Goenka, M., & Alexander, E. (2022). Word clouds in the wild. *Proceedings of the IEEE Workshop on Visualization for the Digital Humanities*, 43–48. <https://doi.org/10.1109/VIS4DH57440.2022.00015>
- Lample, G., Conneau, A., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. *International Conference on Learning Representations*. <https://openreview.net/forum?id=H196sainb>
- Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.
- Sárossy, B., & Ligeti-Nagy, N. (2024). How to talk the talk? A comparative overview of keyword usage in Hungarian and Slovenian parliamentary corpora. In V. Vandeghinste & T. Kontino. (Eds.), *Clarín annual conference proceedings 2024*.
- Skeppstedt, M., Ahltop, M., Kucher, K., Aangenendt, G., Lindström, M., & Söderfeldt, Y. (2025). The word rain visualisation technique applied to digital history: How to visualise, explore and compare texts using semantically structured word clouds. In G. Bouma, D. Dannélls, D. Kokkinakis, & E. Volodina (Eds.), *Huminfra handbook: Empowering digital and experimental humanities*. University of Tartu Library. <https://doi.org/https://doi.org/10.58009/aere-perennius0175>
- Skeppstedt, M., Ahltop, M., Kucher, K., & Lindström, M. (2024). From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization*, 23(3), 217–238. <https://doi.org/10.1177/14738716241236188>
- Söderfeldt, Y., Burchell, A., Reed, J., & Skeppstedt, M. (2025). Topic timelines for enabling close and distant reading of discursive shifts. a pilot case using periodicals of european diabetes organizations. *Journal of Open Humanities Data*. <https://doi.org/10.5334/johd.286>
- Søgaard, A., Agić, Ž., Martínez Alonso, H., Plank, B., Bohnet, B., & Johannsen, A. (2015, July). Inverted indexing for cross-lingual NLP. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1713–1722). Association for Computational Linguistics. <https://doi.org/10.3115/v1/P15-1165>
- Yousef, T., & Janicke, S. (2021). A survey of text alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 1149–1159. <https://doi.org/10.1109/TVCG.2020.3028975>