Comparing Word Frequencies and Lexical Diversity with the ZipfExplorer Tool

Steven Coats[0000-0002-7295-3893]

English Philology, University of Oulu, 90014 Oulu, Finland
steven.coats@oulu.fi

Abstract. The ZipfExplorer is a tool for the interactive comparison and visualization of shared word type frequencies for two texts or corpora. The tool can be used to give insight into similarities and differences in textual and discourse content in terms of individual keywords or groups of keywords, and also calculates several measures of lexical diversity for the shared types of the selected texts. A selection of texts and corpora can be analyzed, and users can upload their own files for interactive comparison.

Keywords: Word Frequencies, Visualization, Lexical Diversity, Zipf.

1 Introduction

The study of lexical type frequencies and their distributions in texts and corpora has been an important focus of research in linguistics and natural language processing in recent decades [1], a development that has been facilitated by the creation of libraries in popular programming languages such as R or Python [2, 3, 4], access to large data sets (e.g. from social media or digitization projects), the sharing of code and data on platforms such as CLARIN, GitHub, or the Center for Open Science, and by continual advancements in technologies related to data processing and storage. The characterization, modeling, and analysis of word frequency profiles and other heavy-tailed distributions is an active field of research [5, 6, 7, 8], but from a broader perspective, word frequencies not only exemplify probability distributions, but can also shed light on differences in discourse in a range of medial, geographical, and social contexts. The comparison of word frequencies between texts or corpora is a fundamental procedure in corpus linguistics, but visualizations of data have not always been a focus in research or pedagogy. Interactive visualizations can be useful for exploratory and heuristic analysis during the research process, can complement textual reports in the presentation of results, and can be utilized in pedagogy, especially in data-based, empirical sciences [9, 10].

The ZipfExplorer[1] is a tool for the interactive visualization of the frequencies of shared lexical types in texts or corpora, built using the Bokeh module in Python [11], and named after Zipf's Law [12], the well-known observation that for any text, the

[1] https://zipfexplorer.herokuapp.com

frequency of a given word is approximately inversely proportional to its rank in the table of word frequencies for that text. By visualizing frequency information for two texts or corpora in an interactive form, the ZipfExplorer interactively demonstrates the concept of "keyness" [13, 14], or the extent to which a lexical item occurs more often than would be expected, and thus provides an immediately interpretable overview of differences in the discourse of the two texts. In addition, the tool provides measures of lexical diversity, and can be used to explore the relationship between lexical overlap and type diversity, which may be of theoretical interest. The code for the creation of the tool, as well as the texts used by the tool, are publicly available.[2]

2 Use of the Tool

The default view of the tool (Figure 1) shows word rank on the x-axis and relative frequency (per 10,000 tokens) on the y-axis. Above the graphs, the values for four lexical diversity measures for the shared vocabulary types are shown: the type-token ratio, the Gini coefficient, the power-law alpha exponent, and the Shannon entropy. Circles on the plots represent individual lexemes, and hovering over a word shows the word itself, its rank, frequency and relative frequency, as well as the log-likelihood measure [15, 16] and associated p-value for the type frequency compared to the other text in the shared word types. The plots can be manipulated with zoom, selection, and movement tools.

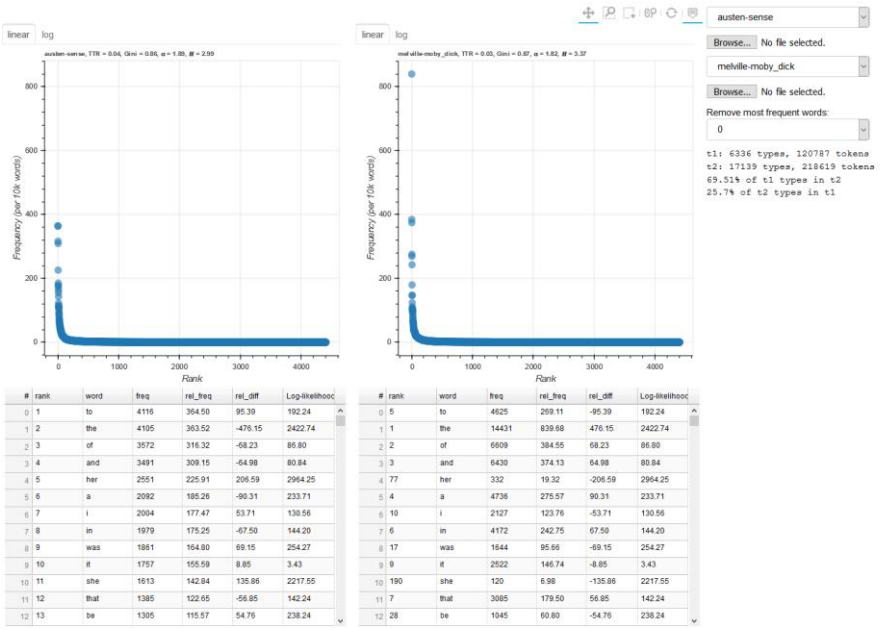


Fig. 1. Default tool view.

[2] https://github.com/stcoats/zipf_explorer

Hovering over a word on the plot will highlight the word type in both graphs, as will clicking on a word in the sortable tables below the plots. For each of the shared lexical types, the tables show the word's frequency rank and relative frequency in the shared types, its relative frequency, the difference in relative frequency compared to the other text, and the log-likelihood value for the comparison. Words with positive relative difference values are more frequent in the selected plot; those with negative values in the other plot. To the right of the plots, texts for comparison can be selected with drop-down lists, and two input buttons allow users to upload their own files (in .txt format). Uploaded files are automatically tokenized and converted to frequency tables. A 'Remove most frequent words' drop-down list removes 0, 10, 20, 50, 100, or 200 of the most frequent words in English, based on the Project Gutenberg English Corpus from Sketch Engine [17]. As many of the most frequent words are determiners, prepositions, conjunctions, or other function words that bear relatively little semantic information, removing frequent words can help to highlight content and discourse differences between the texts. Below the remove words drop-down list, the total number of types and tokens in each text is shown, along with the lexical overlap of the two texts, expressed as a percentage.[3]

The drop-down lists with the source texts comprise several literary texts and a corpus of inaugural addresses of U.S. presidents from NLTK [4], additional texts scraped from Project Gutenberg, the Brown Corpus and its subsections [18], and the Freiburg-Brown Corpus of American English [19]. A planned future tool feature is the automatic calculation of frequency information and lexical diversity measures from web pages.

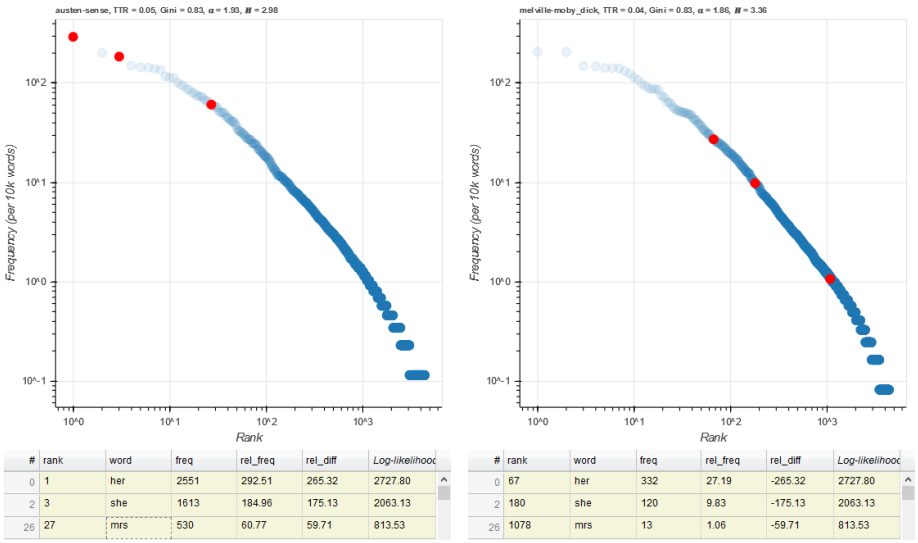
Figure 1 is the default linear-scale view for the shared vocabulary types in Jane Austen's *Sense and Sensibility* and Herman Melville's *Moby Dick*. The most frequent shared type is 'the', which occurs at a frequency of 840 per 10,000 tokens in the types shared by *Moby Dick*, but only 364 per 10,000 words in those shared by *Sense and Sensibility*, possibly indicating underlying stylistic differences in terms of the noun phrase structure of the texts. Although the two plots show heavy-tailed distributions, the degree to which they represent Zipf distributions is difficult to assess visually. Switching the visualization to double-logarithmic scale, however, by selecting the tab above the plots, results in the familiar shape of the Zipf distribution, and makes additional insights into the discourse of the two texts possible.

2.1 Sorting

Sorting word types by difference in relative frequency or by log-likelihood score gives access to the lexemes that are over- or underused in each of the texts, potentially shedding light on discourse differences. In Figure 2, 'her', 'she', and 'mrs' are shown to be much more frequent in *Sense and Sensibility*; the types 'ye', 'fish', 'sail', and 'black' are much more frequent in *Moby Dick* (Figure 3).

---

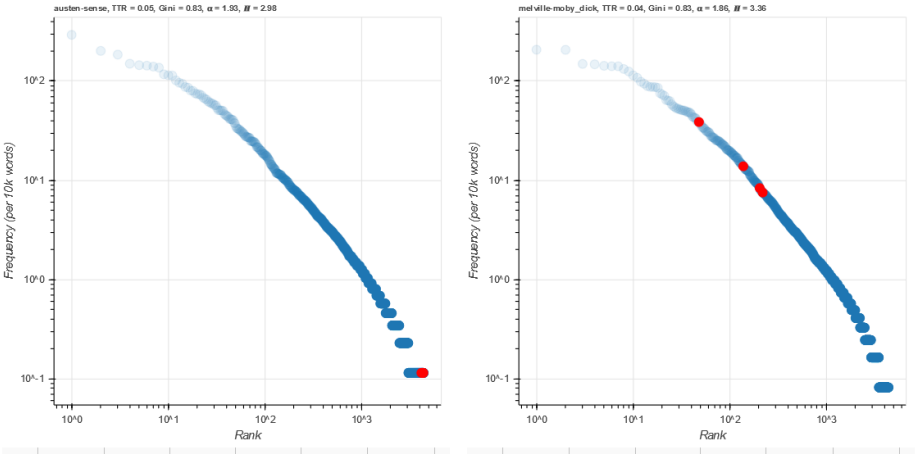
[3] Note that the TTR values shown above the plots are calculated on the basis of shared vocabulary types, whereas the numbers of types and tokens shown below the "remove words" drop-down consider all the types and tokens of each text.



#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
0	1	her	2551	292.51	265.32	2727.80
2	3	she	1613	184.96	175.13	2063.13
26	27	mrs	530	60.77	59.71	813.53

#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
0	67	her	332	27.19	-265.32	2727.80
2	180	she	120	9.83	-175.13	2063.13
26	1078	mrs	13	1.06	-59.71	813.53

Fig. 2. Double logarithmic scale view, *her*, *she*, and *mrs* highlighted.



#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
4320	4321	ye	1	0.11	-38.54	489.84
4363	4364	fish	1	0.11	-13.73	166.98
4158	4159	sail	1	0.11	-8.24	96.33
4239	4240	black	1	0.11	-7.42	85.87

#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
4320	48	ye	472	38.65	38.54	489.84
2348	621	yards	27	2.21	1.87	13.08
3364	869	yard	18	1.47	1.36	10.99

Fig. 3. Double logarithmic scale view, *ye*, *fish*, *sail*, and *black* highlighted.

2.2 *Hapax* Types

Hapax legomena are types that occur only once in a text or corpus. Using the tool to highlight all *hapax* types in *Moby Dick* shows how these words are distributed through the frequency ranks of *Sense and Sensibility*. Many are also *hapax* in the other text, or are found mainly in the tail of the frequency distribution for *Sense and Sensibility*, but some of the *hapax* highlight discourse differences: In addition to names of characters, the highest-ranked *Moby hapax* include types such as 'miss', 'park, or 'manners'.

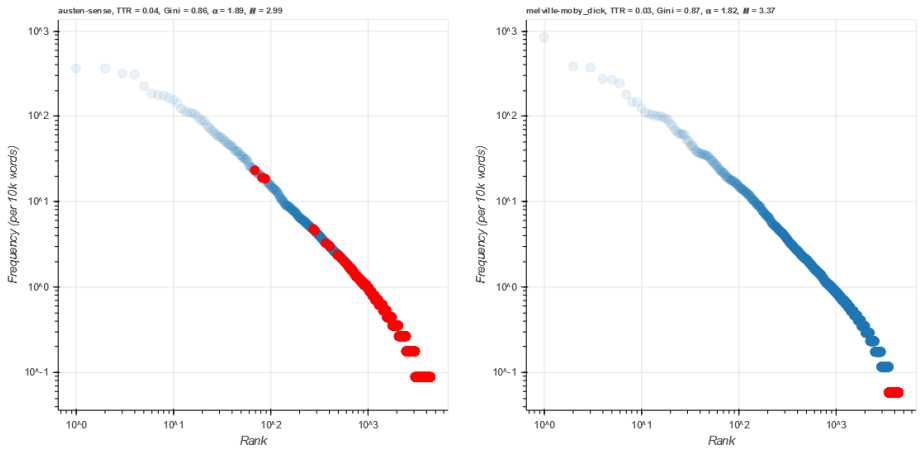


Fig. 4. Distribution of types that are *hapax* in *Moby Dick.*

3 Lexical Diversity

A number of different measures can be used to quantify lexical diversity [6, 20, 21]. The ZipfExplorer displays the Gini coefficient, a measure of diversity ranging from 0 (all words have the same frequency) to a theoretical maximum of 1 (all words have frequency zero except one word, n→∞), the type-token ratio, the exponent α for the best-fit function of the power-law distribution according to the equation $p(x) = Cx^{-\alpha}$, and the Shannon entropy ***H*** [22], which has a maximum theoretical value of $log_2(n)$ for data consisting of *n* shared word types.[4] The tool uses the *powerlaw* package in Python [7] to fit the empirical distribution and calculate the alpha exponent for the fitted function.

Many measures of lexical diversity are affected by sample size [1, 6]. When considering the shared vocabulary of two texts or corpora, typically the shorter text will exhibit lower Gini values and a higher type-token ratio, whereas the longer text will exhibit a smaller α exponent and a higher ***H*** value. Removing frequent words will often

[4] Note that the ZipfExplorer shows the Zipfian rank-frequency profile, not the frequency spectrum, or the number of times that a given frequency value occurs in a text. The Zipf profile is the complementary cumulative distribution of a frequency spectrum with the x- and y-axes reversed [23, 5].

increase values for the type-token ratio and the alpha parameter, and decrease the values for the Gini coefficient and the Shannon entropy. For texts or corpora that share a relatively large proportion of types, such as novels by the same author, a relatively high proportion of the shared types are content words such as nouns or adjectives, and so the lexical diversity measures may provide an indication of the extent to which diverse topical content has been addressed in the discourse of the text. For texts that share relatively few types, however, most of the shared vocabulary consists of function words, and the diversity measure values then show the extent to which the frequencies of these types are evenly distributed in the texts.

4 Summary and Future Outlook

The ZipfExplorer, a tool for the visualization and comparison of lexical frequency information in two different texts or corpora, allows the interactive exploration of word frequencies in a manner that may shed light on content and discourse differences. The measures of lexical diversity calculated by the tool can be interpreted in terms of text length and extent of shared vocabulary as well as diversity of informational content or use of function words.

The tool may be useful in educational contexts such as university courses: Giving students the chance to interactively visualize and compare word ranks and frequencies in different texts and corpora may complement course readings in which Zipf distributions and their mathematical properties are described. In addition, the tool can be used for "distant reading" approaches to textual data in literary, historical, or cultural studies, by providing evidence for discourse similarities or differences.

It is planned that future development of the tool will expand the range of document types that can be compared via user input, add additional diversity measures, and allow text-input-based selection of shared word types in order to (for example) allow comparison of words from specific semantic fields or grammatical classes. In addition, it is hoped that the Python scripts that underlie the tool, which are publicly available at GitHub, can be usefully adapted for other purposes by researchers or laypersons interested in interactive visualization of linguistic data.

References

1. Baayen, R. H.: Word frequency distributions. Kluwer, Dordrecht (2001).
2. Evert, S., Baroni, M.: zipfR: Word frequency distributions in R (R package version 0.6-10 of 2017-08-17). In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions, pp. 29–32, ACL, Stroudsburg, PA (2007).
3. Baayen, R. H., Shafaei-Bajestan, E.: languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics. (R package version 1.5.0). https://CRAN.R-project.org/package=languageR (2019).
4. Bird, S., Loper, E., Klein, E.: Natural language processing with Python. Newton, MA, O'Reilly (2009).

5. Newman, M. E. J.: Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46(5), pp. 323–351 (2005).
6. Clauset, A., Shalizi, C. R., Newman, M. E. J.: Power-Law distributions in empirical data. SIAM Review 51(4), 661–703 (2009).
7. Alstott, J., Bullmore, E., Plenz, D.: Powerlaw: a Python package for analysis of heavy-tailed distributions. PLoS ONE 9(1) (2014).
8. Gillespie, C. S.: Fitting heavy tailed distributions: The poweRlaw package. Journal of Statistical Software 64(2), pp. 1–16. http://www.jstatsoft.org/v64/i02/ (2015).
9. Cleveland, W. S.: Visualizing data. Hobart Press, Summit, NJ (1993).
10. Wilkinson, L.: The grammar of graphics, Springer, New York (2005).
11. Bokeh Development Team. Bokeh: Python library for interactive visualization. http://www.bokeh.pydata.org, last accessed 2019/09/30.
12. Zipf, G. K.: Human behavior and the principle of least effort. Addison-Wesley, Cambridge, MA (1949).
13. Scott, M., Tribble, C.: Textual patterns. John Benjamins, Amsterdam (2006).
14. Stubbs, M. Three concepts of keywords. In: Bondi, M., Scott, M. (eds.), Keyness in texts, pp. 21–42. John Benjamins, Amsterdam (2010).
15. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19, pp. 61–74 (1993).
16. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: WCC '00 proceedings of the workshop on comparing corpora, pp. 1–6. ACM, New York (2000).
17. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography 1, pp. 7–36 (2014).
18. Francis, W. N., Kučera, H.: A standard corpus of present-day edited American English, for use with digital computers. Brown University, Providence, RI (1979).
19. Hundt, M., Sand, A., Skandera, P.: Manual of information to accompany The Freiburg – Brown Corpus of American English (‘Frown’). Department of English, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany (1999).
20. Kunegis, J., Preusse, J.: Fairness on the web: Alternatives to the power law. In: Proceedings of WebSci 2012, June 22–24, 2012, pp. 175–184. ACM, New York (2012).
21. Bérubé, N., Sainte-Marie, M., Mongeon, P., Larivière, V.: Words by the tail: Assessing lexical diversity in scholarly titles using frequency-rank distribution tail fits. PLoS ONE 13(7) (2018).
22. Shannon, C. E.: A mathematical theory of communication. Bell System Technical Journal 27, 379–423; pp. 623–656 (1948).
23. Adamic, L.: Zipf, power-laws, and Pareto—a ranking tutorial. https://www.hpl.hp.com/research/idl/papers/ranking/ranking.html, last accessed 2019/09/30.