ZipfExplorer: A Tool for the Comparison of Shared Lexis

Steven Coats[0000-0002-7295-3893]

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

Abstract. Word frequency statistics and lexical diversity measures can provide insights into discourse differences between texts. The ZipfExplorer, a tool and online app for the interactive visualization and comparison of word frequencies in two texts, shows side-by-side rank-frequency profiles and interactive tables of shared lexis, enabling keyword analysis and shedding light on discourse differences. Four lexical diversity measures (type-token ratio, Gini coefficient, power-law alpha parameter, and Shannon entropy) are calculated for the shared word types. Word frequency information is provided for a selection of mainly literary texts, and users can upload their own files. This paper provides an overview of the visualization of word frequency distributions, describes the functionality of the ZipfExplorer tool and demonstrates some of its features, and briefly discusses the lexical diversity measures calculated by the tool.

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

1 Introduction

Word frequencies are a fundamental starting point for many analytical procedures in corpus-based linguistic, literary, or cultural analysis and for natural language processing tasks.[1] The study of word frequency distributions and their statistical properties continues to be an active topic of research in computational linguistics [2, 3, 4, 5, 6, 7,8], and in recent years, the analysis of word frequencies has been facilitated by the availability of large corpora or other data sets and open access to data via platforms such as CLARIN, GitHub, or the Center for Open Science as well as by dedicated libraries of scripting functions in popular programming languages such as R or Python [9, 10, 11, 12]. The representation of word frequencies in an interactive visualization format, however, has not generally been a primary focus, despite the fact that interactive visualizations can facilitate exploratory data analysis, enhance pedagogy, and complement textual presentation of research [13, 14].

In language and linguistics or literary or cultural studies, the comparison of word frequencies in two texts or between a selected text and a reference corpus is a primary method for gaining insight into differences in discourse content. The ZipfExplorer[2] is an online tool for the interactive visualization of word frequencies in texts or corpora,

[1] This paper, an expanded version of [1], includes a more detailed discussion of the Zipf distribution and the lexical diversity measures calculated by the ZipfExplorer. In addition, some code changes have been made to enhance the useability of the tool.

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

named after Zipf's Law [15, 16], the fact that for most longer natural language texts or corpora, the frequency of a given word type is approximately inversely proportional to its rank in a sorted list of the frequencies of word types for the text. The ZipfExplorer provides an interactive means to show the concept of "keyness" [17, 18], or the extent to which a lexical item occurs more often or less often than would be expected in comparison to a reference text. The tool shows word frequency distributions for the textual overlap of two texts, or the word types that they share, a text aspect that may also be of theoretical interest in terms of its relationship to the concept of textual entailment, or recognizing, given two text fragments, whether the meaning of one text can be inferred (entailed) from the other [19], as well as to word error rate and derived measures of textual similarity used in speech recognition [20]. In addition, the tool, built using the Bokeh module in Python [21], calculates several lexical diversity measures (type-token ratio, Gini coefficient, power-law alpha parameter, and Shannon entropy). The code for the tool is publicly available.[3]

2 Background

Among the first to systematically study lexical type frequencies was the early 20th-century American Germanist George Zipf, who noted that when the words of a text are ordered in decreasing frequency, the relationship between a the frequency and the rank for a word of rank *r* can be expressed as $f_r \approx Cr^{-1}$, where *C* represents a constant. A Zipfian rank-frequency profile, when plotted in double logarithmic space, is typically close to a straight line, but the shape of a frequency distribution for only those lexical types that are shared with a comparison text or reference corpus depends not only on the frequency information of the particular texts under consideration, but also on the degree of textual overlap between the two texts. Visualizations of shared lexis, in addition to highlighting discourse similarities and differences between texts through the examination of particular word types, can also give insight into the interplay between frequencies, derived lexical diversity measures, and the shape of discrete frequency distributions in general.

Following Zipf [15, pp. 45–48; 16, p. 25], word frequency distributions are typically displayed in double logarithmic space, with frequency on the y-axis and frequency rank on the x-axis, as in the top right quadrant of Figure 1, which shows four visualizations of the word frequency distribution for Charles Dickens' 1859 novel *A Tale of Two Cities*. Each circle on the plot corresponds to a distinct word type. The most frequent type, at the top left of the plot, is the word type "the", occurring 8,058 times in the text, followed by "and", "of" and other common words.

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

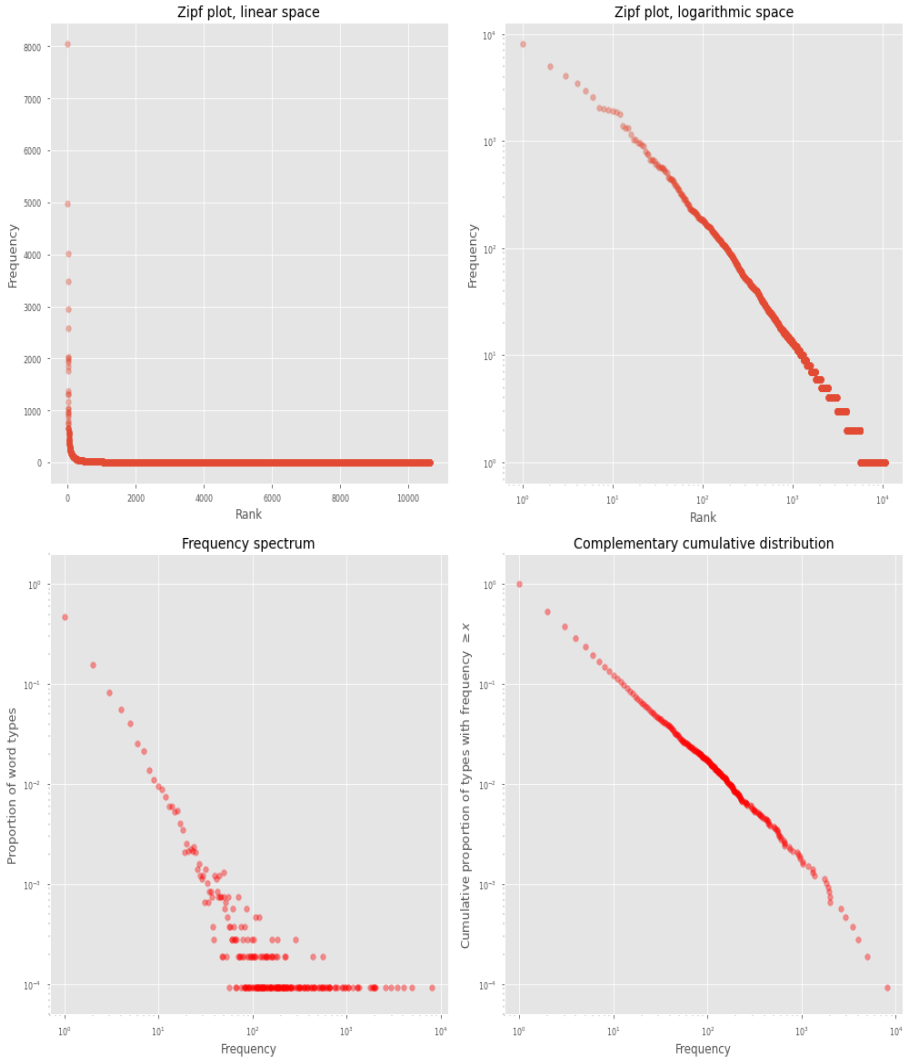


Fig. 1. Four representations of word frequency information for *A Tale of Two Cities*.

The plot in the top left represents the same information in linear space, whereas the lower left plot is the so-called degree distribution (sometimes also referred to as the frequency spectrum): Here, the word frequency counts themselves have been binned, so that the top left circle is the proportion of all word types that occur once in the novel (the *hapax legomena*). *Hapax* comprise 47% of the word types in the novel; words that occur twice (*dis legomena*) 16%, and so on. While the information contained in the Zipf rank-frequency plot and the degree distribution plot is equivalent, the latter plot is more difficult to interpret in terms of discourse, as points on the plot do not correspond to individual word types. In the bottom left of Figure 1, the complementary cumulative

distribution function is depicted: the cumulative proportion of types with a frequency equal to or greater than a given frequency. Thus, 100% of types in the novel occur at least once, 53% at least twice, 37% at least three times, and so on. The complimentary cumulative distribution visualization is the reflection of the Zipf double-logarithmic profile across a line extending from the bottom left to the top right of the subplot. Because the upper two plots in Figure 1 are intuitively easier to understand, the ZipfExplorer visualizes rank-frequency utilizes them, rather than the degree distribution or the complementary cumulative distribution.

3 Tool Functionality

In Figure 2 the default linear-scale view for the shared vocabulary types in Mary Shelley's *Frankenstein* and H. G. Wells' *War of the Worlds* is depicted: Each subplot shows the rank-frequency profile for the text selected via the dropdown menus to the right of the plots. Points on the plots show word relative frequency (per 10,000 words) on the y-axis and type rank in an ordered list of the frequencies of all words in the shared lexis on the x-axis. Values for the lexical diversity measures type-token ratio, Gini coefficient, alpha exponent of the best-fit power-law distribution, and Shannon entropy are shown above the plots. Hovering over a word type will show its rank, frequency, relative frequency, and the log-likelihood measure [22, 23] and associated p-value compared to the shared lexis of the comparison text.

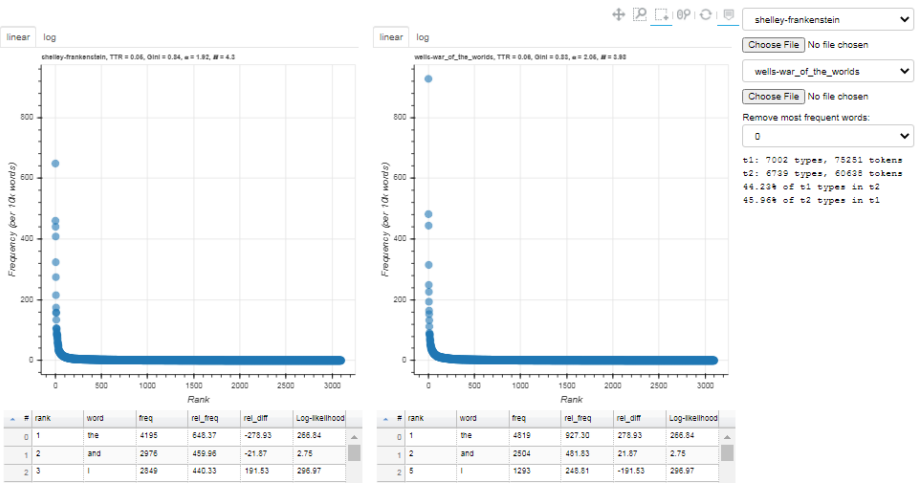


Fig. 2. Default tool view.

Words can be highlighted with a hover tool and selected with a box-drawing tool (in the toolset above the right-hand subplot). Selected words are highlighted in the sortable tables below the plots; clicking on a word in one of the tables highlights it in the plots.

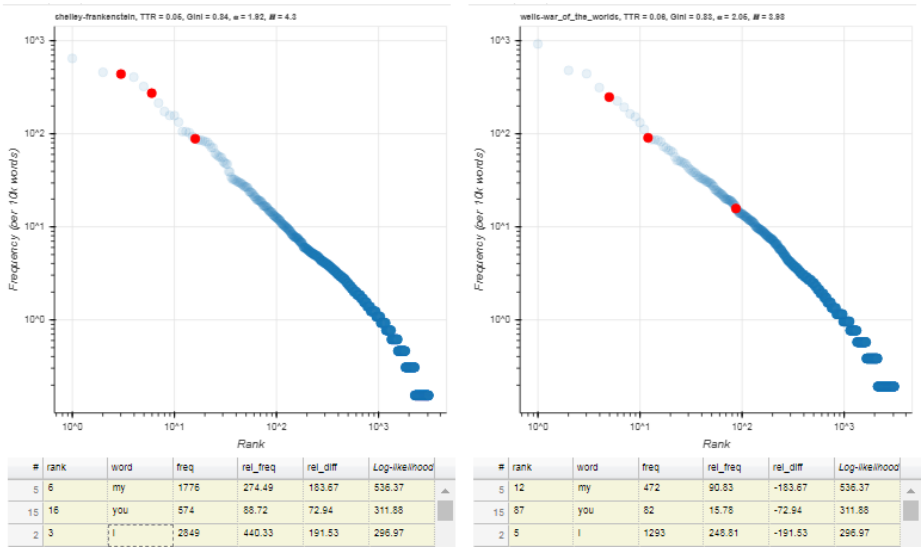
The tables show frequency rank, the word form, frequency, relative frequency, difference in relative frequency compared to the other text, and the log-likelihood value: higher log-likelihood values indicate are calculated for types with larger frequency differences.

The default texts available for comparison are selectable via a drop-down menu to the right of the plots. In addition, users can upload their own texts for comparison with the upload buttons. 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 [24]. 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 menu, the total number of types and tokens in the original texts is shown along with the percentage of types that are shared in the two texts. To examine the word frequency distribution of a single text, rather than the distributions of the shared lexis in two texts, the same text can be selected for both plot windows.

The source texts are a selection of mainly literary texts from Project Gutenberg, a corpus of inaugural addresses of U.S. presidents from NLTK [25], the Brown Corpus and its subsections [26], and the Freiburg-Brown Corpus of American English [27].

3.1 Sorting

The columns in the tables below the subplots can be sorted. They show original word order in the left-hand text, word form, rank in the frequency table, relative frequency, difference to the other text in relative frequency, and log-likelihood score. Sorting can show items that are much more relatively frequent in a text. In Figure 2, the personal pronouns 'my', 'you', and 'I' are more frequent in *Frankenstein*, a text with a first-person point of view, than in the third-person *War of the Worlds*.

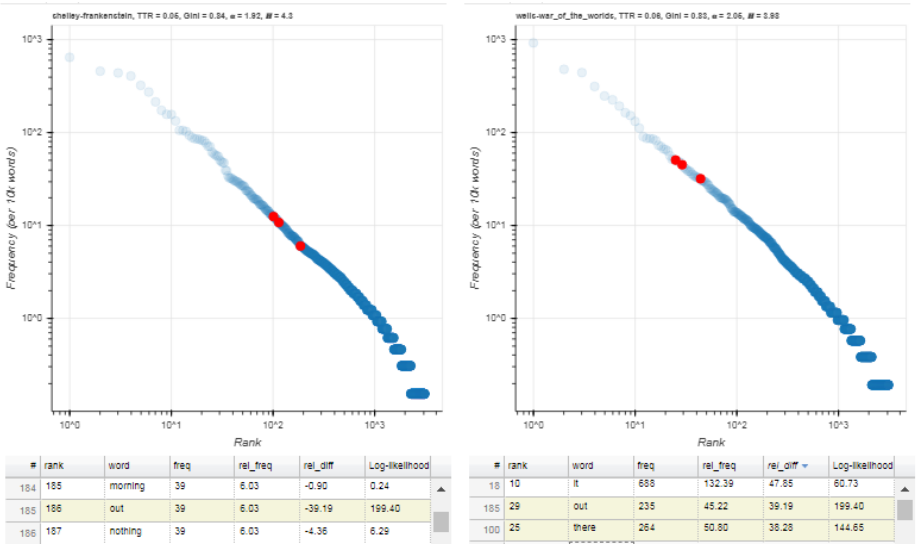


#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
5	6	my	1776	274.49	183.67	536.37
15	16	you	574	88.72	72.94	311.88
2	3	I	2849	440.33	191.53	296.97

#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
5	12	my	472	90.83	-183.67	536.37
15	87	you	82	15.78	-72.94	311.88
2	5	I	1293	248.81	-191.53	296.97

Fig. 3. *My*, *you*, and *I* in *Frankenstein* and *War of the Worlds*.

The types ‘up, ‘out’, and ‘there’ (Fig. 4) are more relatively frequent in *War of the Worlds*; when considered along with other prepositions, place adverbials and location names, it becomes clear that spatial organization plays a greater role as a narrative element in *War of the Worlds* than in *Frankenstein*.



#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
184	185	morning	39	6.03	-0.90	0.24
185	186	out	39	6.03	-39.19	199.40
186	187	nothing	39	6.03	-4.36	6.29

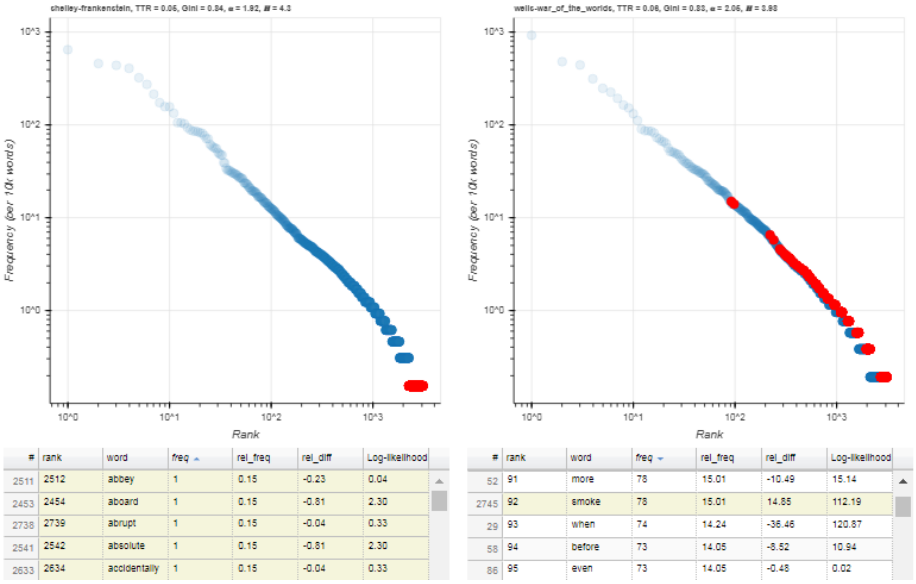
#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
18	10	it	688	132.39	47.85	60.73
185	29	out	235	45.22	39.19	199.40
100	25	there	264	50.80	38.28	144.65

Fig. 4. *Up*, *out*, and *there* in *Frankenstein* and *War of the Worlds*.

‘

3.2 *Hapax* Types

Hapax can also shed light on discourse differences. Highlighting the *hapax* types in *Frankenstein* (in the left-hand rank-frequency profile of Fig. 5) shows their ranks and relative frequencies in *War of the Worlds*: although many are also *hapax* in the other text, or are found mainly in the tail of the frequency distribution for *Frankenstein*, the types ‘smoke’ and ‘red’ are much higher in the profile in *War of the Worlds* – a frequency difference that reflects the discourse content of the latter novel.



#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
2511	2512	abbey	1	0.15	-0.23	0.04
2453	2454	aboard	1	0.15	-0.81	2.30
2738	2739	abrupt	1	0.15	-0.04	0.33
2541	2542	absolute	1	0.15	-0.81	2.30
2633	2634	accidentally	1	0.15	-0.04	0.33

#	rank	word	freq	rel_freq	rel_diff	Log-likelihood
52	91	more	78	15.01	-10.49	15.14
2745	92	smoke	78	15.01	14.85	112.19
29	93	when	74	14.24	-36.46	120.87
58	94	before	73	14.05	-8.52	10.94
86	95	even	73	14.05	-0.48	0.02

Fig. 5. Distribution of types that are *hapax* in *Frankenstein*.

3.3 Stopword Removal

Using the drop-down menu to the right of the subplots, 10, 20, 50, 100, or 200 of the most-frequent types in the Gutenberg corpus can be removed from the visualizations and tables. Because these types, which are mostly function words such as determiners, pronouns, and relativizers or common verbs, structure texts in important ways but contribute relatively little to discourse content, removing them may serve to highlight discourse differences between two texts.

In terms of the distribution shape and the derived lexical diversity statistics, the removal of common words has the effect of increasing the relative frequency of the remaining words. In effect, removing stopwords tends to change the shape of the Zipf profile in double-logarithmic space to a more curvilinear form – when function words are no longer considered, word frequencies deviate substantially from a power-law distribution. As can be expected, removal of common words tends to increase the lexical diversity of the texts for the remaining shared types, which tend to be more uniformly distributed in terms of their relative frequencies.

4 Lexical Diversity

The ZipfExplorer displays four lexical diversity measures: the type-token ratio, the Gini coefficient, the exponent α for the best fit of a power-law distribution, and the Shannon entropy ***H*** [3, 4, 28]. These measures, while related, can be used to highlight different aspects of lexical diversity. The type-token ratio, $\frac{number\ of\ distinct\ types}{number\ of\ tokens}$ has a range in the interval (0,1], with smaller values indicating less lexical diversity.

The Gini coefficient, which can be calculated with

$$G = \frac{2\sum_i^n ix_i}{n\sum_i^n x_i} - \frac{n+1}{n}$$

ranges from 0 (no diversity) to 1 (maximum diversity) for *n* word types with relative frequencies *x*.

The exponent α results from the best-fit line to the degree distribution function (lower left plot in Fig. 1) for the frequency information for the shared lexical types, calculated using the *powerlaw* package in Python [9] with the equation $f(n) \propto n^{-\alpha}$. The alpha parameter is related to the slope *z* of the Zipf rank-frequency profile by $\alpha = 1 + \frac{1}{z}$ ($z = \frac{1}{\alpha - 1}$) [29]. The parameter typically ranges in value between ~1.5 and 3, although higher or lower values are calculated for the shared vocabulary of extremely dissimilar or extremely short texts.[4]

Shannon entropy [30], calculated with

$$H = -\sum_i^n x_i \log_2 x_i$$

has a maximum theoretical value of $log_2(n)$, for data consisting of n unique types.

The diversity statistics calculated by the tool provide evidence for the sensitivity of lexical diversity measures to sample size [2, 4]. For the textual overlap between two texts or a text and a corpus, the shorter text will likely exhibit lower Gini values and a

[4] In these cases, however, word frequencies are unlikely to be distributed according to a power law, and thus the measure is not necessarily a good diversity indicator. See Clauset, Shalizi and Newman (2009).

higher type-token ratio, whereas the longer text will exhibit a smaller α exponent and a higher ***H*** value. Removing frequent words will often increase values for the type-token ratio and the alpha parameter, and decrease the values for the Gini coefficient and the Shannon entropy, although this depends on the texts in question, their original frequencies, and the degree of textual overlap. For texts with a relatively large proportion of shared types, such as two novels by the same author, and with the removal of frequent function words, the lexical diversity measures may give insight into topical diversity in terms of narrative development. For texts that share relatively few types, the relationship between the measure values and the properties of the underlying original texts is less straightforward.

5 Conclusion

The ZipfExplorer enables the interactive exploration of word frequencies in the shared lexis of two comparison texts or corpora, potentially shedding light on discourse similarities and differences and properties of frequency distributions. The lexical diversity measures type-token ratio, Gini coefficient, alpha parameter of the power-law function, and Shannon entropy, calculated by the tool, vary according to text length and textual overlap and are also affected by the removal of common function words.

In a pedagogical context, the ZipfExplorer provides a hands-on way to make frequency information concrete. Given the increasing importance of artificial intelligence models not only in linguistics and other sciences, but ultimately in many working-life and administrative domains and in the contexts of daily life, the tool can serve as a starting point for understanding how linguistic frequency distributions underlie the large data sets used train machine learning models.

The tool may also be useful for the comparison of various discrete distributions in computational studies of language or digital humanities, and for applied analysis in literary, historical, or cultural studies in which “distant reading” approaches are employed. Planned further development of the tool is to allow upload of different file formats, enable text extraction from URLs, and enable automatic annotation of part-of-speech tags or named entities whose frequency distributions may be of interest. It is also hoped that other researchers will use the code for the tool (or parts thereof), available at GitHub, in order to create new and exciting ways to visualize linguistic data such as word frequency information.

References

1. Coats, S.: Comparing word frequencies and lexical diversity with the ZipfExplorer tool. In Sanita Reinsone, Inguna Skadiņa, Anda Baklāne and Jānis Daugavieti (eds.), Proceedings of the 5th Digital Humanities in the Nordic Countries Conference, Riga, Latvia, October 21–23, 2020, pp. 219–225. CEUR, Aachen, Germany (2020).
2. Baayen, R. H.: Word frequency distributions. Kluwer, Dordrecht (2001).

3. 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).
4. Clauset, A., Shalizi, C. R., Newman, M. E. J.: Power-Law distributions in empirical data. SIAM Review 51(4), 661–703 (2009).
5. Lü, L, Zhang, Z.-K., Zhou, T. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. PLoS One 5(12) e14139 (2010). https://doi.org/10.1371/journal.pone.0014139
6. Montemurro, M. A.: Beyond the Zipf-Mandelbrot law in quantitative linguistics. Physica A: Statistical Mechanics and its Applications 300(3–4), 567–578 (2001).
7. Newman, M. E. J.: Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46(5), 323–351 (2005).
8. Piantadosi, S. T. Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review 21(5), 1112–1130 (2014).
9. Alstott, J., Bullmore, E., Plenz, D.: Powerlaw: A Python package for analysis of heavy-tailed distributions. PLoS ONE 9(1) (2014).
10. 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).
11. 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).
12. Gillespie, C. S.: Fitting heavy tailed distributions: The poweRlaw package. Journal of Statistical Software 64(2), 1–16. http://www.jstatsoft.org/v64/i02/ (2015).
13. Cleveland, W. S.: Visualizing data. Hobart Press, Summit, NJ (1993).
14. Wilkinson, L.: The grammar of graphics, Springer, New York (2005).
15. Zipf, G. K.: The psycho-biology of language. Routledge, London (1936).
16. Zipf, G. K.: Human behavior and the principle of least effort. Addison-Wesley, Cambridge, MA (1949).
17. Scott, M., Tribble, C.: Textual patterns. John Benjamins, Amsterdam (2006).
18. Stubbs, M. Three concepts of keywords. In: Bondi, M., Scott, M. (eds.), Keyness in texts, pp. 21–42. John Benjamins, Amsterdam (2010).
19. Androutsopoulos, I., Malakasiotis, P.: A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research 38, 135–187 (2010).
20. Morris, A., Maier, V., Green, P.: From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In: Proceedings of INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, pp. 2765–2768 (2004).
21. Bokeh Development Team. Bokeh: Python library for interactive visualization. http://www.bokeh.pydata.org, last accessed 2019/09/30.
22. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19, 61–74 (1993).
23. 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).
24. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography 1, 7-36 (2014).
25. Bird, S., Loper, E., Klein, E.: Natural language processing with Python updated for NLTK 3.0. Newton, MA, O'Reilly (2019).

26. Francis, W. N., Kučera, H.: A standard corpus of present-day edited American English, for use with digital computers. Brown University, Providence, RI (1979).
27. 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).
28. 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).
29. Adamic, L.: Zipf, power-laws, and Pareto—a ranking tutorial. https://www.hpl.hp.com/research/idl/papers/ranking/ranking.html, last accessed 2020/12/04.
30. Shannon, C. E.: A mathematical theory of communication. Bell System Technical Journal 27, 379–423; 623–656 (1948).