Zipf's Law for Word Frequencies

Christian Bentz[a,b]
[a]Chair of Digital Humanities, Saarland University, Saarbrücken, Germany
[b]Chair of Multilingual Computational Linguistics, University of Passau, Passau, Germany

Abstract

Zipf's law of word frequencies, often simply referred to as *Zipf's law*, is the most well-known quantitative pattern found in language data and beyond. This chapter gives a brief overview of the historical developments which led to the discovery of the law. This is followed by an overview of the methodological problems when fitting the law to language data. The mathematical formulations of the most common versions – also known as the *rank-frequency law*, *number-frequency law*, and *Zipf-Mandelbrot law* – are provided. Given this theoretical and methodological backdrop, the meaning of the law for linguistic analyses is discussed. The chapter closes with an outline of the ongoing debates about possible explanations for this ubiquitous statistical pattern.

Keywords: Word frequencies, Zipf's law of word frequencies, Zipf's rank-frequency law, Zipf-Mandelbrot law

Key points

- Brief historical overview for Zipf's law of word frequencies.
- Summary of methodological problems to be addressed in research on Zipf's law.
- Outline of the main mathematical formulations of the law.
- Discussion of possible explanations for Zipf's law.

1 Introduction

In any natural language text, words reoccur. Thus, for a set of word types $\mathcal{V} = \{w_1, w_2, \ldots, w_n\}$ we can assign a frequency f_i to each type w_i. For example, in the text chunk of the *Universal Declaration of Human Rights* (UHDR) below, the word type 'and' occurs four times ($f_{\text{and}} = 4$), the word types 'are' and 'in' occur two times each ($f_{\text{are}} = 2$, $f_{\text{in}} = 2$), and all the others ('All', 'human', 'beings', etc.) occur exactly once. Word types which only occur once are often referred to as *hapax legomena*. The frequencies above are *raw* or *absolute* frequencies. They are normalized to *relative* frequencies by dividing them with the overall number of tokens: $f_{\text{and}} = \frac{4}{30}$, $f_{\text{are}} = \frac{2}{30}$, $f_{\text{in}} = \frac{2}{30}$, etc.

Article 1
All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.
–Universal Declaration of Human Rights (UDHR), English

As the text size (number of tokens) increases, so will the counts of the word types. However, certain words (e.g. 'the', 'and', 'of') are much more likely to reoccur than others (e.g. 'conscience', 'endowed'). A relatively small subset of words accrue most of the token counts, while many others are left with just a few – if any. This is a general property of texts in English – and other languages.

To illustrate this, Table 1 gives orthographic word counts ranked from highest to lowest for three typologically very different languages (Abkhaz, English, and Hawaiian). By convention, the ranks are assigned numbers from 1 to n, where *increasing* rank values correspond to *decreasing* frequencies. The distributions of ranks vs. frequencies are visualised in Figure 1. Types with the highest frequencies typically belong to the categories of *function words* ('the', 'and', 'of', etc.), auxiliary verbs ('has', 'have'), and, in some cases, nouns ('right', 'азин'). While the former are generally highly frequent, the occurrences of the latter depend more on the topic of the respective text.

In the early 20th century, George Kingsley Zipf proposed a hyperbolic relationship between the ranks and frequencies of word types – which he argued to be (relatively) invariant across texts and languages. This has become known as *Zipf's law for word frequencies* or *rank-frequency law*. Often it is referred to simply as *Zipf's law*. In a few cases – to acknowledge Jean-Baptiste Estoup's earlier contribution – it is called *Zipf-Estoup law*.

2 Methodological preliminaries

When conducting research on Zipf's law of word frequencies, researchers need to deal with several methodological problems. The most salient problems are outlined in the following.

2.1 The tokenization problem

Word frequency distributions vary based on the tokenization of a text, namely, depending on the definition of "word". We can define words based on *orthography*: *word types are strings of alpha-numeric characters split on white spaces and punctuation*. This definition is adopted in the examples of this entry. However, it only applies to texts written in scripts which delimit tokens by white spaces. There are further "word" definitions based on linguistic criteria such as the *phonological word*, and the *morphosyntactic word*. There is currently no general consensus in linguistics on a coherent definition applicable to all languages (cf. Haspelmath, 2011; Wray, 2014).

Haspelmath (2023) proposes that the word definition should encompass: i) *free morphs*, ii) *clitics*, as well as iii) *roots* and *compounds* with potential affixes. For further definitions of these terms see Haspelmath (2023, p. 285-289). Examples are given in Table 2. Free morphs, clitics, and roots (with or without affixes) are typically delimited by white spaces and punctuation – at least in alphabetic and syllabic scripts. Compounds, on the other hand, are in some languages written with white spaces (e.g. *credit card*, and Italian *carta di credito*) and in some without (e. g. German *Kreditkarte*). However, it has been shown in Bentz et al. (2017) that compounds account for only around 1-2% variation in word frequency distributions of English and German. These results suggest that the definition by Haspelmath (2023) largely aligns with the *orthographic definition* given above when applied to corpus data – at least for languages written in scripts using white spaces.

Bentz, Christian (2026). Zipf's Law for Word Frequencies. In: Nesi, Hilary and Milin, Petar (Eds.) *International Encyclopedia of Language and Linguistics (3rd edition, Volume 14)*, Oxford: Elsevier. Author Accepted Manuscript, CC by 4.0.

Table 1 Frequencies and ranks for orthographic words in three languages. The counts are here taken from the first 1000 tokens of the respective UDHR texts. The five most frequent words, as well as the tail of five *hapax legomena* (word types with frequency of one) are displayed. Note that the *hapax legomena* are ordered alphabetically. This is why we find words starting with 'w' at the long tail of the distributions for English and Hawaiian. The Hawaiian words are translated using the dictionary at https://wehewehe.org (last accessed 05.08.2025). The Abkhaz words are translated using the parallel text in English as well as the grammar by Chirikba (2003). Misinterpretations are my own.

Hawaiian			English			Abkhaz		
Word	Freq.	Rank	Word	Freq.	Rank	Word	Freq.	Rank
ka 'the'	82	1	the	65	1	ауаҩы 'human'	39	1
i 'at, in, on, etc.'	75	2	and	57	2	дарбанзаалак 'every'	34	2
a 'of, by'	51	3	to	57	3	азин 'right'	30	3
ke 'the'	42	4	of	46	4	иара 'he'	23	4
nā 'the (plural)'	42	5	right	28	5	имоуп 'has'	23	5
...	...	...	...	...	...	...	...	...
waena 'middle'	1	169	women	1	339	шьаҭа 'basis'	1	618
wahi 'place'	1	170	working	1	340	шьаҭас 'based on'	1	619
wahine 'woman'	1	171	works	1	341	шьҭнахуазароуп 'it should promote'	1	620
wāhine 'women'	1	172	worship	1	342	ыказароуп 'it should be'	1	621
wai 'water'	1	173	worthy	1	343	ыкамзароуп 'it should not be'	1	622

Source: Universal Declaration of Human Rights (UDHR). These frequency-rank distributions are published as part of the supplementary data in Bentz (2018), see http://www.christianbentz.de/book.html.

Table 2 English examples of different categories subsumed under the "word" definition by Haspelmath (2023).

Category	Examples
free morph	*nice, work, now, ouch*
clitic	*the, to, 's*
root (plus affixes)	*tree, nice-r, go-ing, re-place-ment-s*
compound (plus affixes)	*flower-pot, wind-shield, flower-pot-s*

Source: Adapted from (Haspelmath, 2023, p. 285).

2.2 The problem of representation

A "language" is only indirectly and incompletely represented by a given amount of textual material. In this entry, the UDHR is used as a sample for illustration purposes. This is a tiny fraction of all the texts written in the respective languages. In order to prove the validity of linguistic laws more generally, further registers, genres, and modalities (spoken, written, signed) have to be included in a given analysis.

2.3 The problem of language diversity

There are more than 8000 languages spoken and signed in the world today according to Glottolog (Hammarström, 2025). Currently available text samples only cover a fraction of these. Standardized codes for languages (ISO 639-3) and scripts (ISO 15924) help to pinpoint the cross-linguistic breadth of a sample, and to built balanced samples including different language families, areas, and scripts.

2.4 The problem of model fitting

It is not trivial to define what it means for a law to "hold" for a given text and language from a modelling point of view. E. G. Altmann & Gerlach (2016) discuss in detail the issues surrounding model fitting for Zipfian and other quantitative laws. Five steps are necessary in this context:

- Firstly, a functional relationship needs to be defined. For example, Zipf's original proposal for the rank-frequency law features one parameter (α, see Equation 1 below). Mandelbrot adjusted the functional relationship providing another parameter (β, see Equation 2 below).
- Secondly, the parameters of a given formulation (i.e. α, β) need to be estimated by choosing a method, e.g. maximum likelihood (ML), least squares (LS), etc.
- Thirdly, the estimated parameter values have to be shown to significantly diverge from a proposed (baseline) value.
- Fourthly, the goodness of fit of the functional relationship has to be assessed.
- Fifthly, it needs to be assessed to what extend the estimated parameter values are influenced by *data availability*, i.e. text sizes.

2.5 Intermediate summary

Researchers working on Zipf's law face methodological problems relating to *tokenization*, *representativeness*, *language diversity*, and *model fitting*. These problems were here briefly sketched as a backdrop for the historical, mathematical, and linguistic sections on Zipf's law of word frequencies. Book-length treatises about working with word frequency distributions include Baayen (2001) and Popescu (2009). Quantitative analyses of diverse languages and writing systems are discussed, for instance, in G. Altmann & Fengxiang (2008) and Bentz (2018).

3 History

In the context of developing an optimal system for *shorthand writing* (stenography), an extensive collection of word frequencies for German was published in the late 19th century (Kaeding, 1898). In the discussion section of this volume, further collection efforts for Chinese words and characters are mentioned. In this context, the following statement is found:[1]

> Hence, a relatively small portion (not even one-seventh) of the entire vocabulary is of practical relevance [i.e. for learning and using the written language], and relatively few characters are frequent, while most are exceedingly rare: a result which will be found rather naturally in any frequency analysis of a given language.
>
> ***–Kaeding (1898, p. 670)***

[1] Translated from German.

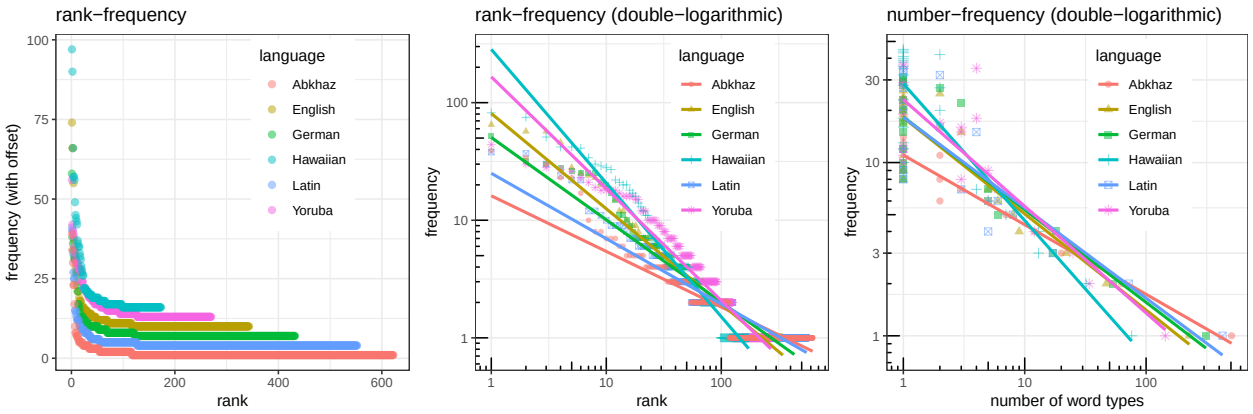

Figure 1 Rank-frequency and number-frequency distributions for orthographic words of the UDHR in six languages. Abkhaz (abk) is an Abkhaz-Adyghe language of the Northwestern Caucasus, Hawaiian (haw) is an Austronesian language of Hawaii, Yoruba (yor) is an Atlantic-Congo language of Nigeria and Benin. Left panel: Ranks (x-axis) versus frequencies (y-axis). Offsets on the y-axis are added (+4 per language) to prevent overplotting of the low-frequency tails. Middle panel: Ranks and frequencies on double-logarithmic scale (base 10) as is common for visualizations of Zipf's law of word frequencies. Lines of best fit are plotted according to the linear least squares method. Right panel: Number of word types n_f (x-axis) with a given frequency (y-axis), also on double-logarithmic scale (base 10). Lines of best fit are plotted according to the linear least squares method. The frequency distributions underlying this plot are based on the first 1000 tokens of the UDHR, published as supplementary data of Bentz (2018), see http://www.christianbentz.de/book.html.

The French stenographer Jean-Baptiste Estoup (1916) formulated more precisely that in consecutive chunks of French text there is a hyperbolic relationship between frequencies of previously unseen words and those already seen (see historical overview in Petruszewycz, 1973). Zipf (1949) cited Estoup as the first person to have uncovered such a relationship (cf. Petruszewycz, 1973, p. 50), and subsequently provided mathematical formulations for Estoup's observations. While Estoup focused on French, Zipf analyzed a considerable variety of languages (e.g. English, Old English, Norwegian, Nootka, Plains Cree, Gothic, Mandarin Chinese, etc.).

Extending on Zipf's proposal, Mandelbrot (1953, 1965) pointed out that especially high frequency types (low ranks) systematically diverge from the expected functional relationship proposed by Zipf. For Mandelbrot's adjustments to the mathematical formulas see Section 4.

In the time when Kaeding, Estoup, Zipf, and Mandelbrot established their analyses of the rank-frequency relationship, large corpora and automated processing tools were not yet available. The respective data on word frequencies would have to be laboriously collected from books. For example, Zipf (1949, p. 23-24) uses counts based on the *Ulysses* by James Joyce which encompasses 260,430 word tokens and 29,899 word types (i.e. ranks). Mandelbrot (1953, p. 491) gives a plot with a maximum of 50,000 tokens and approx. 1000 ranks (i.e. types). With the advent of digitization and computational tools, larger corpora of texts – also deriving from different authors – became available. This has led to further reassessments of the law of word frequencies. For example, Ferrer-i-Cancho & Solé (2001) use approx. 90 million tokens of the British National Corpus (BNC). Their results indicate a faster decay of frequencies than expected from the Zipfian formulation especially for higher ranks beyond 10,000.

4 Mathematical formulations for different flavors of Zipf's law

Zipf and other researchers have provided *different mathematical formulations* for the distributional properties of word frequencies in texts (see also Baayen, 2001; Popescu, 2009).

In the following, the mathematical formulas for the so-called *rank-frequency law*, the *number-frequency law*, and the *Zipf-Mandelbrot (ZM) law* are provided. These can be seen as different "flavors" of the law of word frequencies. Here the formulas are provided which are typically found in the recent literature. The actual historical formulas from the primary literature are given in the Appendix at the end of the chapter.

4.1 Zipf's rank-frequency law

In recent publications (cf. Ferrer-i-Cancho & Elvevåg, 2010; Piantadosi, 2014; E. G. Altmann & Gerlach, 2016), the *rank-frequency law* is typically formulated as

$$f \propto r^{-\alpha}, \qquad [1]$$

where r is the rank assigned to a word of frequency f – in a ranking from highest to lowest frequency. For example, the word with highest frequency in the UDHR example from above is 'and' ($r = 1$), followed by 'are' and 'in' ($r = 2$ and $r = 3$), and then all the hapax legomena ($r = 4$, etc.).

The exponent α represents the slope of a straight line on double-logarithmic scale, in this case with *ranks* on the x-axis and frequencies on the y-axis (see middle panel of Figure 1, and Figure 2). While Zipf (1949, p. 25) showed that this exponent is $\alpha \approx 1$ for English novels (e.g. the *Ulysses*), he later finds (Zipf, 1949, p. 84) that it is $\alpha \approx 0.4$ for texts in Nootka and Plains Cree.

Equation 1 is typically referred to when *Zipf's law of word frequencies* (or simply *Zipf's law*) is mentioned in quantitative analyses.

4.2 Zipf's number-frequency law

One of the earliest formulations of Zipf's law of word frequencies is found in the treatise on *The psychobiology of language* (Zipf, 1965, first published in 1935). In this original work, Zipf considered the *number of word types (n) with a certain frequency (f)*, here denoted as n_f. For example, in Article 1 of the UDHR above, 'and' is the only word type occurring four times (n_4 = 1), while 'are' and 'in' both occur twice ($n_2 = 2$), and all other word types only occur once ($n_1 = 22$).

Tabulating frequencies of word types for English, Chinese, and Latin, Zipf (1965, p. 40-44) proposed an inverse power law relationship between the observed frequencies (f), and the number of word types with a given frequency (n_f), i.e.

$$n_f = \frac{C}{f^\theta} = Cf^{-\theta},$$

where C is a proportionality constant. In parallel to α in Equation 1 above, θ represents the slope of a straight line fitted through the number-frequency points on double-logarithmic scale (see right panel of Figure 1). The proportionality constant C is often dropped from the formula, and the proportionality sign ($\propto$) is used instead, replacing the equality sign ($=$), thus yielding

$$n_f \propto f^{-\theta}.$$

Beware that this formulation with *numbers of word types* and *frequencies* is in later work often referred to as *number-frequency law* – as opposed to the more common *rank-frequency* law in Equation 1. The *number-frequency* and *rank-frequency* formulations, while given different names, are often treated as "two sides of the same coin" (cf. Semple et al., 2022, Table 1) since their exponents are connected via

$$\theta = \frac{1}{\alpha} + 1,$$

for $\theta > 1$ (cf. Mandelbrot, 1961; Ferrer-i-Cancho & Hernández-Fernández, 2008). Some researchers argue that the number-frequency formulation is preferable when investigating Zipf's law of word frequencies – at least from the perspective of model fitting and statistical hypothesis testing (Moreno-Sánchez et al., 2016; Corral et al., 2020).

4.3 Zipf-Mandelbrot law

To capture significant deviations from linearity (compare Figure 1, middle panel) of the rank-frequency points per word type, Mandelbrot (1953, 1965) introduced further parameters to the formula. In later publications (cf. Piantadosi, 2014), his adjusted formula is typically represented as

$$f \propto (r + \beta)^{-\alpha}. \quad [2]$$

Hence, the main addition to Zipf's law as formulated in Equation 1 is the β-parameter (cf. Manin, 2009, p. 275). This adjustment is relevant mostly for the highest frequency items at low ranks ($r < 100$), i.e. function words (Montemurro, 2001, p. 568). See also Figure 2 for the effect of parameters on the shape of the theoretical curve. The formula in Equation 2 is typically referred to as *Zipf-Mandelbrot law*.

4.4 Further adjustments: exponent regimes

Ferrer-i-Cancho & Solé (2001) argue that $\alpha \approx 1$ does not fit the rank frequency distribution well even for English – especially for large corpora such as the *British National Corpus* (BNC) with 90 million tokens. Namely, for higher ranks (i.e. low frequency words) there is a significant drop-off from the proposed line. They introduce two separate regimes, such that we have (Ferrer-i-Cancho & Solé, 2001, p. 168)

$$f(r) \propto \begin{cases} r^{-\alpha_1} & \text{if} \quad r < N \\ r^{-\alpha_2} & \text{otherwise,} \end{cases}$$

where $\alpha_1 \approx 1$ and $\alpha_2 \approx 2$, and N corresponds to the number of ranks which is roughly in the range $[10^3, 10^4]$. This result is confirmed in Montemurro (2001, p. 571) for a corpus of 2606 English books consisting of 183,403,300 word tokens and 448,359 word types. Here, it is found that $\alpha = 1.05$ in the first regime ($N < 6000$) and $\alpha = 2.3$ in the second regime ($N > 6000$). Moreover, the two regimes have been confirmed across seven languages (Chinese, English, French, German, Hebrew, Russian, Spanish) for the *Google Books Ngram* corpus (Petersen et al., 2012).

Potential causes for these two regimes are discussed in Williams et al. (2015a) and Williams et al. (2015b). Firstly, Williams et al. (2015a) argue that text mixing is a mechanism involved in the sharper drop of frequencies characterizing the second regime. Secondly, Williams et al. (2015b) illustrate that this drop is found in different orders of magnitude depending on whether letters, words, phrases, or clauses are used as types.

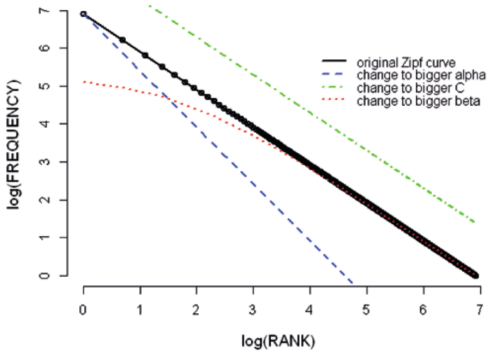

Figure 2 The effect of different parameter values for the Zipf-Mandelbrot law. α and β are the parameters, C is the normalizing constant. Reprinted with modifications from Bentz et al. (2014). Note: the natural logarithm of rank and frequency values is taken, and the resulting numbers are plotted on the axes. In contrast, in Figure 1 middle and right panel, the axis scales are transformed instead.

5 The linguistic meaning of Zipf's law

The parameter values in Zipf's law of word frequencies have linguistic relevance. In the examples of Table 1, Hawaiian (haw) has function words such as articles (e.g. *ka* 'the') and prepositions (*i* 'at, in, on, etc.') in the first ranks. These have high relative frequencies: $f^{\text{haw}}_{\text{ka}} = \frac{82}{1000}$ and $f^{\text{haw}}_{\text{i}} = \frac{75}{1000}$. In comparison, Abkhaz (abk) has more content words in the first ranks (e.g. ауаҩы 'human', азин 'right') which have lower relative frequencies: $f^{\text{abk}}_{\text{ауаҩы}} = \frac{39}{1000}$ and $f^{\text{abk}}_{\text{азин}} = \frac{30}{1000}$.

Moreover, Hawaiian has a relatively small number of word types – equivalent to the maximum number of ranks: $\max(r^{\text{haw}}) = 173$.

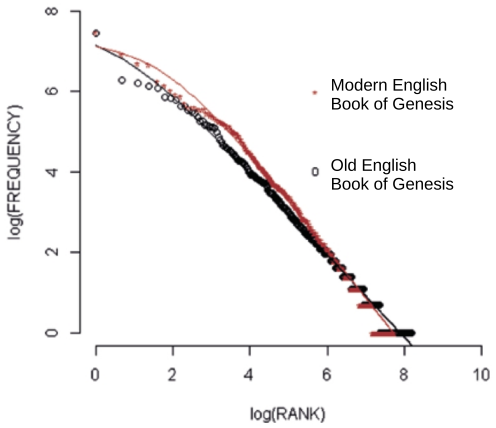

Figure 3 Zipf-Mandelbrot law fitted to the Modern English and Old English Book of Genesis via maximum likelihood estimation of the parameters. Reprinted with modifications from Bentz et al. (2014).

In contrast, Abkhaz has many more word types: $\max(r^{\text{abk}}) = 622$. Hawaiian is a highly *analytic* language with a short-tailed word frequency distribution with higher α and β values, while Abkhaz is a highly *synthetic* language with a long-tailed word frequency distribution with lower α and β values. English is located in between. Thus, fundamental typological features of languages are reflected in the parameters of the Zipf-Mandelbrot law.

Against this backdrop, Bentz et al. (2014) and Koplenig (2018) elaborate how the α and β parameters of the Zipf-Mandelbrot law have systematically changed over historical time. They tease apart different language change phenomena responsible for increases/decreases in the parameters.

Bentz et al. (2014) use parallel Bible texts of different historical periods of English. The Zipf-Mandelbrot parameters have *increased* from Old English towards Middle English and finally Modern English (see Figure 3 and Table 3). Besides such factors as style of translation and changes in orthography, this historical change is most strongly related to the trade-off between *synthetic* and *analytic* grammatical structures. For example, in Old English, inflectional case-marking was still common: *þæs land**es** gold*. In Modern English, this is often replaced by constructions with prepositions and the definite article: ***the** gold **of** that land*. As a consequence, a wide range of case-marked forms is traded-off for a small range of grammatical function words, and the word frequency distributions shift from long-tailed to short-tailed. This is reflected in an *increase* of the parameters α and β.

Koplenig (2018), on the other hand, analyses historical changes from 1800 to 2000 in six language varieties (American English, British English, French, German, Italian, Spanish) represented by the *Google Books Ngram* corpus (Michel et al., 2011). For this large-scale corpus, a systematic *decrease* of the ZM-parameters is detected in all languages over the 200 year period. For example, α dropped from around 1.11 to 1.03 in German, and from 1.1 to 1.05 in Spanish (Koplenig, 2018, p. 20). Further analyses reveal that the variance in α and β values is most strongly related to vocabulary size and the noun-pronoun ratio, and mostly unrelated to the mean sentence length (cf. Figure 7 in Koplenig, 2018). Increases in vocabulary sizes and noun-pronoun ratios lead to longer tailed distributions, and hence lower α-values.

Table 3 Maximum-likelihood estimated Zipf-Mandelbrot parameters for different versions of the Book of Genesis.

Text	α	β	C	R^2
Modern English (lemmatized)	1.29	6.16	16502	0.96
Modern English	1.22	5.29	11951	0.95
Old English	1.03	1.49	3329	0.92

Source: Adapted from Bentz et al. (2014).

This effect is witnessed also when children learn languages such as English, Dutch, German, and Swedish (Baixeries et al., 2013). The α-exponent of their language production *decreases* with age as new words and word-forms are learned and used – and the vocabulary expands. Similarly, aphasics speech (both fluent and non-fluent) yields significantly higher α-exponents than comparable speech of non-affected individuals in English, Greek, and Hungarian (cf. Van Egmond, 2018, p. 162).

Besides variation of ZM parameters within languages, the variation of values across languages of the world has also been assessed. Currently, the range of the α-exponent in the Zipf-Mandelbrot law is estimated to $0.4 < \alpha < 2.7$ for close to one thousand languages represented by different parallel corpora (Bentz et al., 2015), and to $0.77 < \alpha < 1.44$ for Bible translations of 100 languages (Mehri & Jamaati, 2017).[2]

6 Possible explanations for Zipf's law

Zipf (1949) himself argued that the statistical laws he formulated for natural languages are related to the *principle of least effort*. Hence, the frequency distributions of words are a reflection of the trade-off between "speaker's economy" and "auditor's economy" (Zipf, 1949, p. 20-21). From the perspective of the speaker, it is most convenient to just utter a single word, i.e. with maximum frequency. However, for the hearer this would make the task of mapping the uttered word to the intended meaning extremely difficult. According to Zipf, the actual frequency distribution of uttered words emerges as a compromise between the two.

Zipf's ideas were quickly challenged by analyses illustrating that power law distributions can be the outcome of random, non-communicative processes of sequence generation, sometimes referred to as "monkey typing" (Miller, 1957). In reply to this, several authors have pointed out that random typing is not a useful baseline when trying to causally *explain* the existence of Zipf's law in natural language production (Howes, 1968; Manin, 2009; Ferrer-i-Cancho & Elvevåg, 2010; Piantadosi, 2014). For random typing to be a sufficient causal explanation, we would have to assume that humans (or monkeys) indeed produce letters/words at random.

Moreover, it is worth noting that Miller (1957, p. 313-314) himself states: "Our monkeys are doing the best possible job of encoding information word-by-word [...]". Indeed, Ferrer-i-Cancho et al. (2022) prove that, counter-intuitively, random typing is an *optimal* coding process from the perspective of standard information theory. Hence, finding that Zipfian laws hold for random typing does not strictly contradict Zipf's original proposal that they relate to communicative efficiency.

[2]Note: this work uses the formulation in Equation 1, not the Zipf-Mandelbrot formulation in Equation 2. The single parameter (α) is called ζ there.

Zipf's original view is supported by information-theoretic models which reproduce Zipf's law for word frequencies under a critical balance between speaker's and auditor's needs (Ferrer-i-Cancho, 2005). In this context, experimental research has shown that Zipfian distributions increase learnability of linguistic input (Lavi-Rotbain & Arnon, 2022, 2023; Wolters et al., 2023), and that there is a bias towards these distributions in artificial language learning (Shufaniya & Arnon, 2022; Arnon & Kirby, 2024). Notably, the law does not seem to be restricted to the spoken and written domain, as it has been also detected in three sign languages (Kimchi et al., 2023).

In summary, Zipf's law of word frequencies is a statistical pattern which emerges in texts of diverse languages, and under different approaches of tokenization (characters, orthographic words, phrases, etc.). However, the exact parameter values depend on the typological features of a given language. One future direction of research lies in the delineation of the universal aspects versus the language specific aspects of the law.

Appendix. Historical notes.

This Appendix gives the original versions of the *number-frequency law*, the *rank-frequency law*, and the *Zipf-Mandelbrot law* as found in the primary literature. The derivations for the more recent formulas are also provided.

Zipf's number-frequency law

The original equation by Zipf (1965, p. 41) reads

$$ab^2 = k,$$

with b representing the frequencies ("occurrences") of words (e.g. *hapax legomena* would have frequency $b = 1$), and a being the number of words with a given frequency b (e.g. in Article 1 of the UDHR in English there are overall 22 *hapax legomena*). The law here states that a multiplied by b^2 will give a constant value k across different texts and languages.

Translated to the notation in the main text ($a \equiv n_f, b \equiv f, k \equiv C$) we have

$$n_f f^2 = C.$$

This can be rearranged to

$$n_f = \frac{C}{f^2} = Cf^{-2},$$

which is the specific case, i.e. for $\theta = 2$, of the more general formula

$$n_f \propto f^{-\theta},$$

Zipf's rank-frequency law

As mentioned in the main text in Section 4, the original formulation of the rank-frequency relationship is (Zipf, 1949, p. 24)

$$rf = C, \qquad [3]$$

which is equivalent to

$$f = Cr^{-1}.$$

The implicit assumption here as that the exponent has to be $\alpha = 1$, corresponding to the slope of a straight line through the rank-frequency points on double-logarithmic scale. Since Zipf himself noticed deviations from this exact slope in diverse languages, he introduced a parameter for the exponent (Zipf, 1949, p. 130), which originally is denoted as p (α in Equation 1), such that we have

$$rf^{1/p} = C.$$

Dividing by r and raising to the power of p yields

$$f = \frac{C^p}{r^p} = C^p r^{-p}.$$

Alternatively, the proportionality sign can be used without further specifying C, such that

$$f \propto r^{-p}.$$

Zipf-Mandelbrot law

In the original publication (Mandelbrot, 1965, p. 556), the formula reads

$$i(r, k) = Pk(r + V)^{-B},$$

where i is the frequency of a word, here seen as a function of the rank r and the number of word tokens k in a respective text, while P, V, and B are parameters to be estimated for a given empirical sample of word types. Note that $i(r, k)$ is the absolute frequency (i.e. raw counts), and $\frac{i(r,k)}{k}$ is the *relative* frequency of a word type i, namely normalized by the number of word tokens k. If we drop k, and use the more common f for frequencies, then we get

$$f(r) = P(r + V)^{-B}.$$

In fact, P corresponds to the constant C in Equation 3 of Section 4, and B corresponds to the parameter α. V is then the only new parameter, which is here represented as β. We then have

$$f(r) = C(r + \beta)^{-\alpha}.$$

Using the proportionality sign – replacing the proportionality constant C – this gives

$$f(r) \propto (r + \beta)^{-\alpha}.$$

Acknowledgements

I would like to thank the editor Ramon Ferrer-i-Cancho as well as two anonymous reviewers for detailed comments on this chapter. All remaining errors are my own. This work is funded by the European Union (ERC, EVINE, 101117111). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

See Also: Zipf, George Kingsley (1902–1950), Models of Zipf's law for word frequencies

Altmann, E. G., & Gerlach, M. (2016). Statistical laws in linguistics. In M. Degli Esposti, E. G. Altmann, & F. Pachet (Eds.), *Creativity and universality in language* (pp. 7–26). Springer.

Altmann, G., & Fengxiang, F. (2008). *Analyses of script: Properties of characters and writing systems*. Berlin/New York: Mouton de Gruyter.

Arnon, I., & Kirby, S. (2024). Cultural evolution creates the statistical structure of language. *Scientific Reports*, *14*(1), 5255.

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Dordrecht/Boston/London: Kluwer Academic Publishers.

Baixeries, J., Elvevåg, B., & Ferrer-i Cancho, R. (2013). The evolution of the exponent of Zipf's law in language ontogeny. *PLoS ONE*, *8*(3), e53227.

Bentz, C. (2018). *Adaptive languages: An information-theoretic account of linguistic diversity* (Vol. 316). Berlin/Boston: Walter de Gruyter GmbH & Co KG.

Bentz, C., Alikaniotis, D., Samardžić, T., & Buttery, P. (2017). Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. *Journal of Quantitative Linguistics*, *24*(2-3), 128–162.

Bentz, C., Kiela, D., Hill, F., & Buttery, P. (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, *10*(2), 175–211.

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE*, *10*(6), e0128254.

Chirikba, V. A. (2003). *Abkhaz*. Munich: Lincom Europa.

Corral, Á., Serra, I., & Ferrer-i Cancho, R. (2020). Distinct flavors of Zipf's law and its maximum likelihood fitting: Rank-size and size-distribution representations. *Physical Review E*, *102*(5), 052113.

Estoup, J. B. (1916). *Gammes sténographiques*. Paris: Institut Stenographique de France.

Ferrer-i-Cancho, R. (2005). Zipf's law from a communicative phase transition. *The European Physical Journal B-Condensed Matter and Complex Systems*, *47*(), 449–457.

Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2022). Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, *29*(2), 165–194.

Ferrer-i-Cancho, R., & Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE*, *5*(3), e9411.

Ferrer-i-Cancho, R., & Hernández-Fernández, A. (2008). Power laws and the golden number. In G. Altmann, I. Zadorozhna, & Y. Matskulyak (Eds.), *Problems of general, Germanic and Slavic linguistics* (pp. 518–523). Chernivtsi: Books - XXI.

Ferrer-i-Cancho, R., & Solé, R. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, *8*(3), 165–173.

Hammarström, R. . H. M. . B. S., Harald Forkel. (2025). *Glottolog 5.2*. https://doi.org/10.5281/zenodo.15525265. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, *45*(1), 31–80.

Haspelmath, M. (2023). Defining the word. *Word*, *69*(3), 283–297.

Howes, D. (1968). Zipf's law and Miller's random-monkey model. *The American Journal of Psychology*, *81*(2), 269–272.

Kaeding, F. W. (Ed.). (1898). *Häufigkeitswörterbuch der Deutschen Sprache: Festgestellt durch einen Arbeitsausschuß der deutschen Stenographiesysteme*. Steglitz bei Berlin: Mittler & Sohn.

Kimchi, I., Wolters, L., Stamp, R., & Arnon, I. (2023). Evidence of Zipfian distributions in three sign languages. *Gesture*, *22*(2), 154–188.

Koplenig, A. (2018). Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes–a large-scale corpus analysis. *Corpus Linguistics and Linguistic Theory*, *14*(1), 1–34.

Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of Zipfian distributions in language. *Cognition*, *223*(), 105038.

Lavi-Rotbain, O., & Arnon, I. (2023). Zipfian distributions in child-directed speech. *Open Mind*, *7*(), 1–30.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In W. Jackson (Ed.), *Communication theory* (pp. 486–502). Princeton: Academic Press.

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. In R. Jacobson (Ed.), *Structure of Language and its Mathematical Aspects* (Vol. 12, pp. 190–219). American Mathematical Society.

Mandelbrot, B. (1965). Information theory and psycholinguistics. In B. B. Wolman (Ed.), *Scientific Psychology*. Basic Books Publishing.

Manin, D. Y. (2009). Mandelbrot's model for Zipf's law: Can Mandelbrot's model explain Zipf's law for language? *Journal of Quantitative Linguistics*, *16*(3), 274–285.

Mehri, A., & Jamaati, M. (2017). Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. *Physics Letters A*, *381*(31), 2470–2477.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Lieberman Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*(6014), 176–182.

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, *70*(2), 311–314.

Montemurro, M. A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, *300*(3-4), 567–578.

Moreno-Sánchez, I., Font-Clos, F., & Corral, Á. (2016). Large-scale analysis of Zipf's law in English texts. *PLoS ONE*, *11*(1), e0147073.

Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., & Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports*, *2*(1), 943.

Petruszewycz, M. (1973). L'histoire de la loi d'Estoup-Zipf: documents. *Mathématiques et sciences humaines*, *44*(), 41–56.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(), 1112–1130.

Popescu, I.-I. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

Semple, S., Ferrer-i Cancho, R., & Gustison, M. L. (2022). Linguistic laws in biology. *Trends in Ecology & Evolution*, *37*(1), 53–66.

Shufaniya, A., & Arnon, I. (2022). A cognitive bias for Zipfian distributions? Uniform distributions become more skewed via cultural transmission. *Journal of Language Evolution*, *7*(1), 59–80.

Van Egmond, M. (2018). *Zipf's law in aphasic speech: An investigation of word frequency distributions*. Utrecht: LOT.

Williams, J. R., Bagrow, J. P., Danforth, C. M., & Dodds, P. S. (2015a). Text mixing shapes the anatomy of rank-frequency distributions. *Physical Review E*, *91*(5), 052811.

Williams, J. R., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M., & Sheridan Dodds, P. (2015b). Zipf's law holds for phrases, not words. *Scientific Reports*, *5*(1), 12209.

Wolters, L., Lavi-Rotbain, O., & Arnon, I. (2023). Zipfian distributions facilitate learning novel word-referent mappings. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Wray, A. (2014). Why are we so sure we know what a word is? In J. R. Taylor (Ed.), *The Oxford Handbook of the Word* (p. 725–750). Oxford: Oxford University Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, Massachusetts: Addison-Wesley Press.

Zipf, G. K. (1965). *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, Massachusetts: The M.I.T Press.