Linguistic Universals, Zipfian

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

Language universals attract considerable attention in the language and cognitive sciences. They hold promise to unveil the fundamental properties of human cognition and communication in comparison to other species. A particular set of universals are those investigated in the field of quantitative linguistics since the incipient research by George Kingsley Zipf in the early 20th century. This chapter briefly outlines some of the well-known and less well-known linguistic laws derived from Zipf's original work. It further discusses the methodological challenges when testing Zipfian laws for a given sample of language data. Against this theoretical backdrop, the chapter provides a review of the efforts to establish Zipfian laws – in particular *Zipf's law of word frequencies*, *Zipf's law of abbreviation*, and *Zipf's law of meaning distribution* – as language universals. The chapter concludes by discussing the wider relevance of Zipfian laws in comparative studies on animal communication. Finally, ongoing debates about "meaningful" explanations of quantitative laws are outlined.

Keywords: Zipfian laws, law of abbreviation, law of word frequencies, law of meaning distribution, language universals.

Key Points

- Outline of Zipfian laws, especially those relating to *word frequency*, *number-frequency*, *abbreviation*, *meaning-frequency*, *word returns*, and *word age*.
- Discussion of the methodological preliminaries for testing Zipfian laws empirically.
- Overview of research into the universality of Zipfian laws across diverse languages and text types.
- General discussion of statistical versus absolute universals in human language and other communication systems.

1 Introduction

Linguistic universals can be divided into *absolute* and *statistical* (Bickel, 2011). In this nomenclature, absolute universals hold without exception across all languages. This would include those necessarily true by virtue of deduction from basic assumptions of language structure. For example, if we assume that all languages have words, and that words are, in turn, composed of morphemes, then all languages necessarily have morphemes. Of course, even basic assumptions (e.g. that all languages have words) are disputed among linguists. Likewise, claims for absolute universality are generally controversial (Evans & Levinson, 2009). *Statistical universals*, on the other hand, are those tested on a sample of texts/languages. While a certain structural feature might not surface in particular languages, it can still display a strong cross-linguistic tendency.

In fact, strictly establishing absolute universality with a finite sample of languages is difficult. The vast majority of possible languages – those which emerged and vanished over the deep past of human evolution and those which will emerge in the future – are not represented. Some authors even argue that this makes it virtually impossible to establish the universality of a linguistic pattern given the limited breadth of available language data (Piantadosi & Gibson, 2013).

Another (though related) sense of *statistical universal* is referring to quantitative laws which arise in text statistics (cf. Altmann & Gerlach, 2016; Tanaka-Ishii, 2021). Several of these laws originated in the early 20th century – especially in the work of George Kingsley Zipf (Zipf, 1949). The most prominent is *Zipf's law of word frequencies*, also referred to simply as *Zipf's law*. Further well-known laws include *Zipf's number-frequency law*, *Zipf's law of abbreviation*, and *Zipf's meaning-frequency law*. Less well-known laws are *Zipf's law of word returns*, and *Zipfian laws of word age*. All of these are here subsumed under *Zipfian laws*. Zipf's law of abbreviation, for instance, is found without exception across thousands of languages and texts. It is hence a promising candidate for an absolute universal of human languages.

2 Methodological Challenges

In order to establish Zipfian laws as language universals, several methodological questions need to be addressed in a given study:

- How to define and count the *units of measurement*: "word", "length", "meanings"? For example, "length" can be measured in characters, strokes, phonemes, syllables, morphemes, or even durations in a speech corpus.
- How to represent a "language" by textual (or other) material? In other words, variation *within* languages has to be accounted for by considering different authors, genres, registers, and styles.
- How to account for the variation *across* languages? There are more than 8000 languages spoken, written, and signed in the world today (Hammarström, 2025). Language samples typically only represent a small fraction of these. Standardized language codes (ISO 639-3) and script codes (ISO 15924) help to build samples balanced for language families, areas, and writing systems.
- What does it mean to assert that a given law "holds" for a given text and language from a modelling point of view? Altmann & Gerlach (2016) discuss in detail the issues surrounding model fitting for Zipfian and other quantitative laws. Five major steps are necessary in this context:
  - Firstly, a functional relationship needs to be defined. For example, Zipf's original proposal for the law of word frequencies is a power law with one parameter (α, see Equation 1 below). Mandelbrot adjusted the functional relationship providing another parameter (β, see Equation 2).
  - Secondly, the parameters of a given formulation (i.e. α, β, θ, δ, γ, ω given in the equations below) need to be estimated by choosing a method, e.g. maximum likelihood (ML), least squares (LS), etc.

Bentz, Christian (2026). Linguistic Universals, Zipfian. 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 Table with word types and basic statistics for the UDHR example passage. There are overall 25 word types. The word types with equal frequencies are by convention ranked alphabetically.

Word type (w_i)	Rank (r_i)	Frequency (f_i)	Length (l_i)
and	1	4	3
are	2	2	3
in	3	2	2
a	4	1	1
...	...	...	...
spirit	22	1	6
they	23	1	4
towards	24	1	7
with	25	1	4

- Given a functional relationship and estimated parameter values, how to decide whether these values are significantly different from a baseline value – for example calculated by drawing randomly from a uniform distribution? To this end, Altmann & Gerlach (2016) give guidelines for statistical hypothesis testing.
- Finally, to what extend are estimated parameter values influenced by *data availability*, i.e. text sizes?

Against this methodological backdrop, past and current efforts to find the best fitting functional relationships and parameters for Zipfian laws are briefly reviewed in the following.

3 Zipfian Laws

Assume that a text can be split into segments, for instance, orthographic words. We thus have a set of word types, i.e. a vocabulary $\mathcal{V} = \{w_1, w_2, \ldots, w_n\}$, with n being the cardinality (size) of the vocabulary $n = |\mathcal{V}|$. Two fundamental statistics of a word type w_i are its frequency (f_i), and its length (l_i).

For example, in the UDHR passage below, the word type 'and' has a frequency of four, and a length of three – in terms of UTF-8 characters. The word type 'brotherhood', on the other hand, has a frequency of one, and a length of 11. These 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.

Furthermore, each word type can be assigned a rank (r_i) according to its frequency. By convention, the most frequent word type is assigned rank one, the second most frequent rank two, etc. For an overview of these basic statistics given the example text see Table 1.

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

3.1 Zipf's Law of Word Frequencies

Zipf (1949, p. 24) noticed that in such a ranking there is a hyperbolic relationship between the rank of a word type and its frequency. This can be captured by the equation

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

with α being a parameter to be estimated for a sample of word types and their respective frequencies. Several researchers before Zipf made similar observations, for example, Kaeding (1898) and Estoup (1916). However, today the law is typically referred to as *Zipf's law of word frequencies* or *rank-frequency law*. Over the years, there have been various adjustments to this law based on further empirical investigations. One of the most well-known being the *Zipf-Mandelbrot law*, which introduces an extra parameter (Mandelbrot, 1965, p. 556), such that we have

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

where β captures deviations from linearity (on double-logarithmic scale) especially for high-frequency words.

Zipf (1949, p. 25) provided a tabulation of word frequencies in the novel *Ulysses* by James Joyce, and showed that in this case $\alpha \approx 1$. Notably, however, he did not restrict his analyses to English, but extended them to a variety of languages such as Old English, Norwegian, Nootka, Plains Cree, Dakota, Gothic, and Mandarin Chinese, among others. He noticed that for languages such as Nootka and Plains Cree (Zipf, 1949, p. 84) the exponent changes to $\alpha \approx 0.4$. He asserted that while there are such differences regarding "statistical minutiae", the overall picture emerging is that there is an "obvious fundamental linearity in the various curves" (Zipf, 1949, p. 85). The linearity here refers to the double-logarithmic scale. Note that the power law in Equation 1 is represented by a straight line with slope $-\alpha$ when logarithms are taken for both f and r (hence *double-logarithmic* scale). See also Figure 1 for visualizations of the UDHR data.

However, Altmann & Gerlach (2016) illustrate that different fitting methods can lead to divergent parameter estimations – even for the same textual material. For instance, they compare maximum likelihood (ML) fitting of Zipf's law as given in Equation 1 (MLr), to a version of the number-frequency law given below in Equation 3 (MLf), as well as the least squares (LS) method applied to ranks and frequencies on double-logarithmic scale. For the *Ulysses* they find $\hat{\alpha}^{\text{MLr}} = 1.18$, $\hat{\alpha}^{\text{MLf}} = 1.15$, and $\hat{\alpha}^{\text{LS}} = 1.03$, such that there is a maximal divergence of 0.15 in these α-estimations. On the other hand, the values across different English texts for the same estimation method are more narrowly distributed, i.e. ranging between 1.18 and 1.22 for 10 different English novels. Remarkably, they find a mere 0.01 difference between the *Ulysses* as originally used by Zipf, and the entire English Wikipedia ($\hat{\alpha}^{\text{MLr}}_{\text{Ulysses}} = 1.18$ vs. $\hat{\alpha}^{\text{MLr}}_{\text{Wiki}} = 1.17$).

In conclusion, Altmann & Gerlach (2016) caution that whether the law "holds" in a statistical sense for a particular language depends considerably on i) the mathematical formulation of the law, ii) the estimation method, and (iii) on the text/corpus – though to a lesser extent.

Systematic variation in the law of word frequencies has been assessed for a variety of text genres such as children's and their caregivers' language production (Baixeries et al., 2013; Lavi-Rotbain & Arnon, 2022, 2023), schizophrenics and military combat texts (Ferrer-i-Cancho, 2005), and aphasics (Van Egmond, 2018). Such investigations of variation given genres have primarily been carried out on texts of Indo-European languages.

With regards to cross-linguistic variation, Bentz et al. (2015) estimate the α-parameter (in the Zipf-Mandelbrot law) with the ML

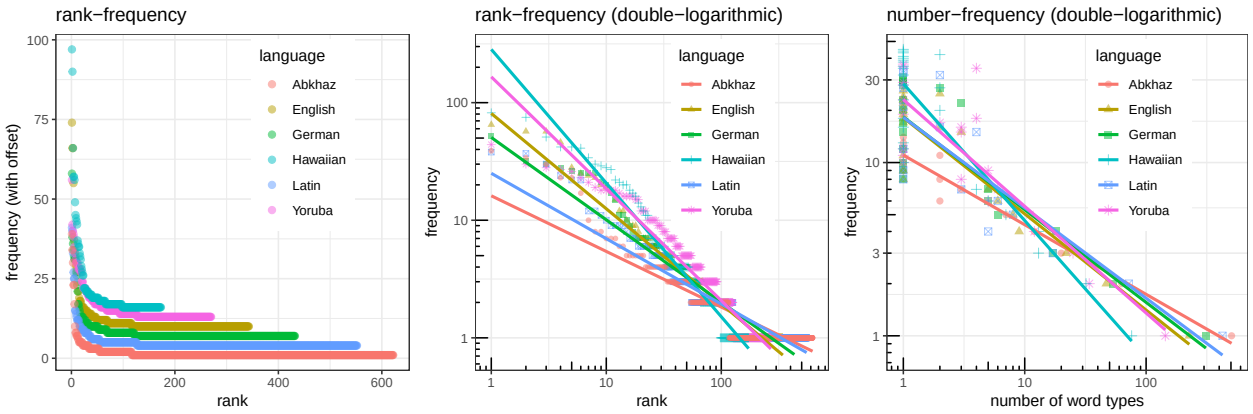

Figure 1 Rank-frequency and number-frequency distributions for orthographic words of the UDHR in six languages. Abkhaz is an Abkhaz-Adyghe language of the Northwestern Caucasus, Hawaiian is an Austronesian language of Hawaii, Yoruba is an Atlantic-Congo language of Nigeria and Benin. Left panel: Ranks (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 1000 tokens of the UDHR, published as supplementary data of Bentz (2018), see http://www.christianbentz.de/book.html.

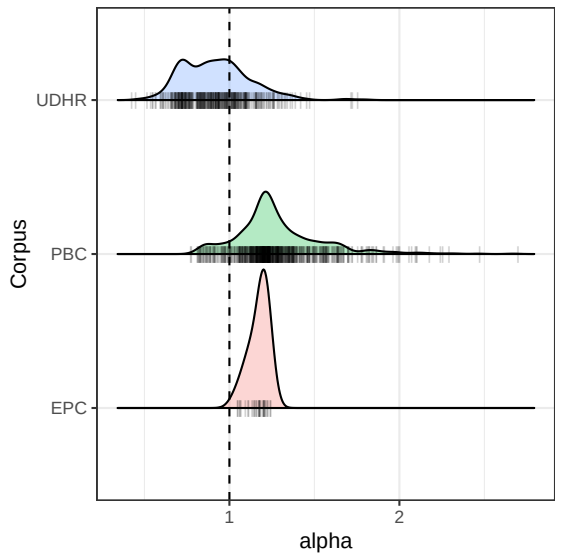

Figure 2 Zipf's law of word frequencies illustrated across typologically diverse languages. Density distributions of $\hat{\alpha}$-parameter values of the Zipf-Mandelbrot law are given for three corpora separately: Universal Declaration of Human Rights (UDHR), European Parliament Corpus (EPC), and Parallel Bible Corpus (PBC). This sample contains 984 texts, written in 764 languages of 84 language families. This data is given in the supplementary material of Bentz et al. (2015).

method for 984 parallel texts written in 764 languages (see Figure 2). They find that the values range from $\hat{\alpha} = 0.47$ in Eastern Canadian Inuktitut to $\hat{\alpha} = 2.67$ in Terêna (Arawakan language of South America). The average estimated value is $\overline{\alpha} = 0.92$ for the UDHR, $\overline{\alpha} = 1.17$ for the European Parliament Corpus (Koehn, 2005), and $\overline{\alpha} = 1.28$ for the Parallel Bible Corpus (Mayer & Cysouw, 2014).

A smaller range of α-parameter estimations is found in Mehri & Jamaati (2017). Across 100 Bible translations they identify a minimum value of $\hat{\alpha} = 0.77$ for Thai, and a maximum value of $\hat{\alpha} = 1.44$ for Korean.[1]

These results suggest that the law of word frequencies does not hold without exception across diverse languages – in the *strict* sense of $\alpha \approx 1$. It can still hold, however, in the *wider* sense of a hyperbolic relationship with $\alpha_{\text{min}} \approx 0.4$ and $\alpha_{\text{max}} \approx 2.7$.

3.2 Zipf's Number-Frequency Law

Rather than ranking words according to their frequencies, we can also count how often a certain frequency occurs. Let this be denoted as n_f, such that n_1 is the number of word types for which $f = 1$ (so-called *hapax legomena*), n_2 the number of word types for which $n = 2$, etc. Zipf (1949, p. 32)[2] formulated a law connecting f and n_f such that (Semple et al., 2022, p. 57)

$$n_f \propto f^{-\theta}, \qquad [3]$$

where θ is a parameter to be estimated like α and β in the laws above. The *rank-frequency* and *number-frequency* law are rarely discussed separately as they are often considered "two sides of the same coin" (cf. Semple et al., 2022, p. 57). However, some authors prefer the number-frequency over the rank-frequency formulation – especially in the context of model fitting and statistical hypothesis testing (Moreno-Sánchez et al., 2016; Corral et al., 2020).

3.3 Zipf's Law of Abbreviation

Besides the relationship between word frequencies and their ranks/numbers, Zipf (1965, p. 23) also pointed out that there is an inverse relationship between the "magnitude" of a word type (the

[1] 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.

[2] The equation in Zipf's original work reads $N(f^2 - \frac{1}{4}) = C$, with $N = n_f$ and C being a proportionality constant.

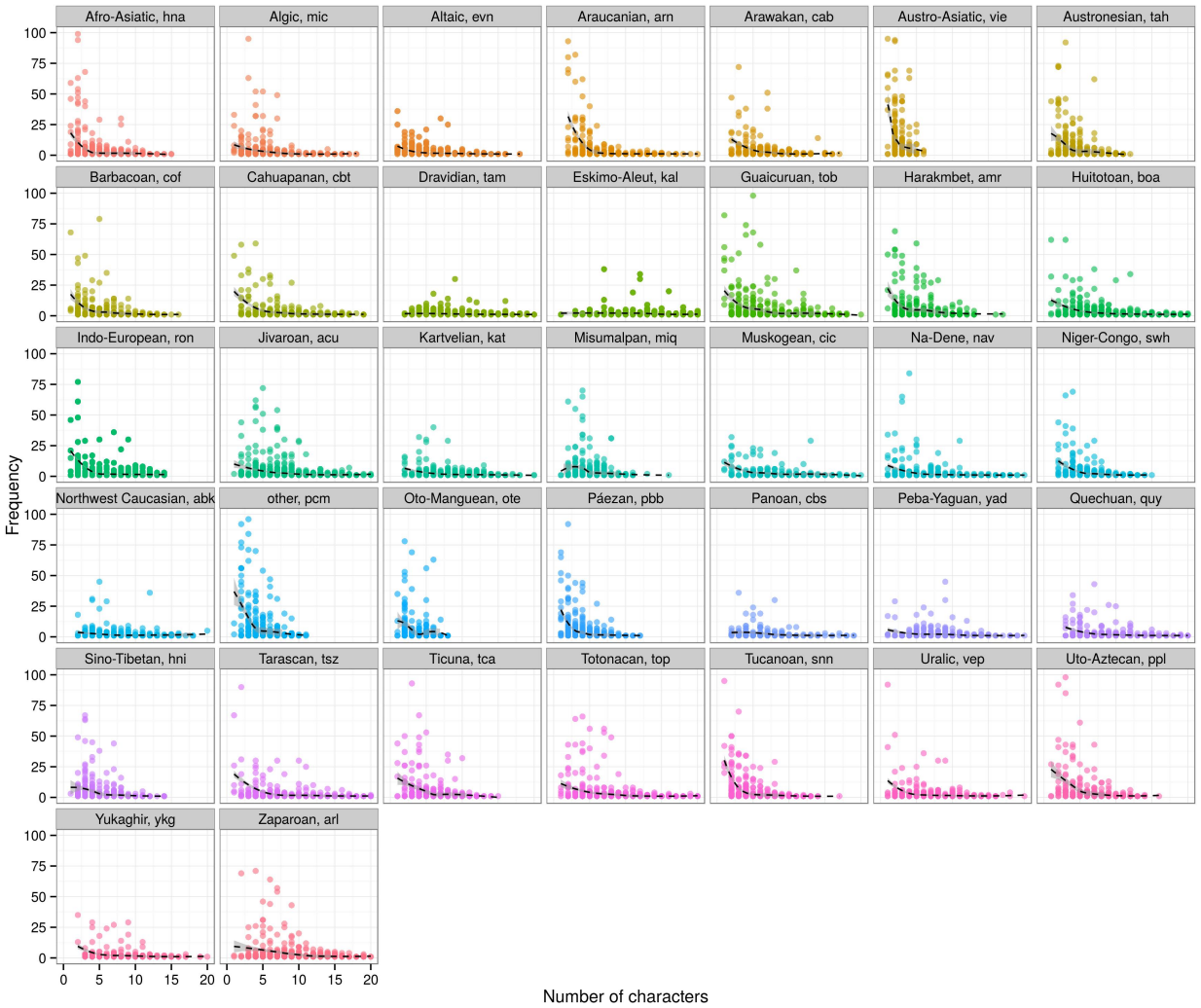

Figure 3 Zipf's law of abbreviation illustrated across 37 typologically diverse languages from different language families. The language families are chosen to represent different macro areas of the world. The language family names are given above the panels, followed by the ISO 639-3 codes identifying the specific languages. Length in UTF-8 characters are given on the x-axes, and raw frequencies on the y-axes. Reprinted from Bentz & Ferrer-i-Cancho (2016).

length (l_i), and its frequency of occurrence (f_i). *Frequent words tend to be shorter*. This is referred to as *Zipf's law of abbreviation*. Mathematically, this is typically formulated with the estimated probabilities ($\hat{p}_i$) of words (in the simplest case a *normalized* or *relative* frequency) standing in an inverse correlation with their lengths (l_i).

The law is then confirmed in empirical research either by fitting mathematical models (see, for example, Grzybek, 2007; Popescu et al., 2013) or by applying standard correlation coefficients, and one- or two-sided statistical tests for significance (see theoretical discussions in Ferrer-i-Cancho et al., 2022; Petrini et al., 2023).

Against this backdrop, Bentz & Ferrer-i-Cancho (2016) assess the Kendall rank correlation for 1262 texts (PBC and UDHR) written in 986 languages. Across all of these texts the Kendall τ coefficient is negative without exception. See Figure 3 for visual illustrations of the length/frequency relationship across diverse languages. The negative correlations are significant for all PBC texts, but non-significant for a few UDHR texts. This is attributed to the small text sizes of some UDHR texts.

While Bentz & Ferrer-i-Cancho (2016) use UTF-8 characters as length measurement, Petrini et al. (2023) also harness durations of word types in a speech corpus for 46 languages of 14 families. Both Pearson and Kendall correlations are negative without exception for this data. All correlation coefficients, except one, are significant. Again, this non-significance is attributed to an exceptionally small number of tokens for the respective language.

Along similar lines, Levshina (2022) uses large web corpora for seven typologically diverse languages (Arabic, Finnish, Hungarian, Indonesian, Russian, Spanish, and Turkish), and finds consistent negative Spearman correlations between length in UTF-8 characters (and phonemes in some cases) and frequency of occurrence.

Moreover, the law of abbreviation is shown to extend to individual characters in a wide variety of writing systems (overall 27) by Koshevoy et al. (2023).

3.4 Zipfian Laws of Meanings

While the laws above refer to frequencies of words regardless of their meaning, Zipf (1945, 1949) also formulated two (closely related) laws linking frequencies as well as ranks to the *number of meanings per word*. Namely, given the number of countable meanings μ per word, Zipf (1949, p. 28) proposes the relationship

$$\mu = \sqrt{f} = f^{\frac{1}{2}},$$

which can be generalized to

$$\mu = f^{\delta},$$

where δ is a parameter to be estimated from data, rather than assumed to be $\delta = \frac{1}{2}$ *a priori*. To estimate μ Zipf worked with meanings given per entry in a lexicon of English (cf. Zipf, 1949, p. 29). Zipf (1945) termed this the *meaning-frequency relationship*, and it is hence referred to as *Zipf's meaning-frequency law* in later work (cf. Ferrer-i-Cancho, 2016; Ferrer-i-Cancho & Vitevitch, 2018).[3]

A variant of this law uses ranks instead of frequencies, yielding

$$\mu \propto r^{-\gamma}. \quad [4]$$

Again, γ is a parameter to be estimated from data samples. Notice the *negation of the exponent*. This is due to the common definition that rank numbers are in an *inverse* relationship with frequencies: rank numbers *increase* with *decreasing* frequencies. Zipf (1949, p. 28) termed this latter formulation the *law of meaning distribution*, hence being referred to as *Zipf's law of meaning distribution* (cf. Ferrer-i-Cancho, 2016; Ferrer-i-Cancho & Vitevitch, 2018) in later publications.

In plain words, these equations express the general observation that more *frequent words tend to have more meanings*. Note though that Zipf excluded the 500 most frequent English words – i.e. largely function words – in his analysis.

To approximate the γ-parameter of the law of meaning distribution as given in Equation 4, Zipf (1945, p. 251) used a list of the 20,000 most frequent words in English – as well as their number of meanings separately listed in the *Thorndike-Century Senior Dictionary*. He plotted the frequency-ranks of word types versus their average number of meanings (by successive frequency bins of 1000, but excluding the 500 most frequent words) on double-logarithmic scale, and estimated the slope between them (with the least squares method) to the range of $\hat{\gamma} = 0.41$ to $\hat{\gamma} = 0.49$ (Zipf, 1945, p. 253).

In a recent study, Casas et al. (2019) replicate Zipf's γ-estimations for English, and provide estimations also for Dutch and Spanish. They employ a database of child directed speech (CHILDES) as well as a Wikipedia corpus for frequency counts, and derive meaning counts (number of senses) from multilingual variants of *WordNet*. They confirm that $\hat{\gamma} \approx 0.5$ for English data when bins of 1000 ranks are used (as originally proposed by Zipf). This holds even for different corpora, and different age groups (children versus adults). However, smaller bins (500 and 100 ranks respectively) systematically influence the γ-estimations, i.e. $\hat{\gamma}_{500} \approx 0.38$ and $\hat{\gamma}_{100} \approx 0.3$.

In a similar study, Bond et al. (2019) revisit Zipf's γ-estimations also using *WordNet* for meaning counts, and further online corpora for frequency counts across overall eight languages (Chinese, English, French, Indonesian, Japanese, Polish, Portuguese, and Spanish). Similar to Casas et al. (2019) they also find a somewhat lower parameter range for English ($\hat{\gamma} = 0.4$ to $\hat{\gamma} = 0.42$) with frequency bins of 500 compared to Zipf's original estimations with bins of 1000. The cross-linguistic range, on the other hand, reaches from $\hat{\gamma} = 0.21$ in Chinese to $\hat{\gamma} = 0.51$ in French (cf. Table 4 in Bond et al., 2019).

Furthermore, they illustrate that the goodness of fit of the law (measured by R^2) drops from high values for frequency bins of 500 (0.86 to 0.98) to low values (0.04 to 0.22) for individual word types. A similar effect is observed in Casas et al. (2019) for English, Dutch, and Spanish. In other words, the law of meaning distribution holds for average values of meaning counts when word types are binned, but displays a poor fit to predict the number of meanings from frequencies (or the other way around) for individual word types.

Nagata & Tanaka-Ishii (2025) take a different approach to the meaning-frequency law. Instead of estimating the number of meanings m from dictionaries directly, they turn to distributional vector semantics inherent to current language models. They measure the variation in vector directions of the trained word vector for a given word type, denoted as v. Following the rationale of vector semantics, the breadth of meanings is here conceptualized as the "contextual diversity" of a word vector. This gives

$$v \propto f^{\delta},$$

as an alternative formulation of the meaning-frequency law.[4] Given this new formulation, Nagata & Tanaka-Ishii (2025) estimate the δ-parameter for two large-scale English (BNC and CCOHA) and Japanese (Aozorabunko and BCCWJ) corpora, as well as 27 translations of the Bible in 24 languages. The (sub)word vectors necessary to calculate v then come from the final layer of BERT language models. Their estimated δ-values range from $\hat{\delta}^{\text{hun}}_{\text{Bible}} = 0.013$ for Hungarian to $\hat{\delta}^{\text{cmn}}_{\text{Bible}} = 0.072$ for Mandarin Chinese.

In summary, the results of these empirical studies (Casas et al., 2019; Bond et al., 2019; Nagata & Tanaka-Ishii, 2025) once more exemplify that different factors are involved when testing whether Zipfian laws "hold" for a given language sample. In the case of parameter estimations for Zipfian laws of meaning, *intra-language* variation, i.e. different text types representing the same language, seems to play a minor role compared to methodological choices (bin sizes) as well as *inter-language* variation, i.e. differences across languages.

3.5 Zipf's Law of Word Returns

The laws discussed above deal with word types and their respective statistics (frequencies, lengths, meanings) derived from texts or dictionaries. However, in language usage, word types are normally not produced in isolation, but as part of a sequence of tokens. A question arising in this context is whether the occurrences of word types in token sequences are also governed by general laws. For example, when exactly is a *word type expected to return* in a text?

In Zipf (1949, p. 40), manual counts are reported for the number of pages (not tokens) intervening between the occurrences of word types with certain frequencies. Zipf (1949, p. 41) derives an equation

[3] Beware notational confusion: Zipf (1949, p. 28) uses upper case F_i for the frequency of occurrence of a word of i-th rank in a rank-frequency profile, and with m_i meanings; while f_i represents the average frequency of occurrence of the m_i meanings in his notation. We here use f instead to represent frequency of occurrence per word in line with the earlier definitions in this chapter, and the common usage in the recent literature.

[4] Nagata & Tanaka-Ishii (2025) have α instead of δ in their formula.

for so-called *word returns*, which can be captured as

$$P(\tau_w) \propto \tau_w^{-\omega}, \quad [5]$$

where $P(\tau_w)$ denotes a relative frequency distribution over the sequence of interval lengths τ_w for given word types w, and the exponent ω is a parameter (see Appendix for the original formulation of the law). For example, the word type *and* in the UDHR passage above occurs in token positions $(7, 11, 18, 20)$. If we count how many tokens intervene between each occurrence of this word type, then we get $\tau_{\text{and}} = (4, 7, 2)$ (cf. Altmann et al., 2009). Zipf's original proposal is that the interval lengths in τ are related to their number of occurrences via a power law. This might be called *Zipf's law of word returns*.

Whether or not such a power law relationship holds for the given word types (or other structural elements) in natural language data is a matter of empirical investigation. This and further functional relationships have been proposed and simulated in Zörnig (2010), and tested for a range of word types in English and Portuguese by Altmann et al. (2009). Altmann et al. (2009) come to the conclusion that a stretched exponential distribution fits the interval lengths for the respective words better than a Poisson process.

3.6 Zipfian Laws of Word Age

Zipfian laws are typically based on statistical properties of words derivable from *synchronic* textual material (their frequencies, lengths, return patterns, etc.). However, Zipf was aware of the importance of *diachrony*, putting the lens also to historical properties of words – such as their *age*. He had a student of his (Dr. Nai-Tung Ting) collect the dates and origins (source language) of adoption for several thousand English word roots (Zipf, 1949, p. 110). Tabulating and visualising this data (see e.g. a reprint in Figure 4) he made the following two general observations:

- The ages of words (time since first adoption) correlate with their frequencies: *old words tend to be the most frequent*.
- There is an inverse relationship between ages and lengths of words: *old words tend to be shorter*.

Of course, the latter is connected to the former via the law of abbreviation. Unfortunately, Zipf (1949) did not explicitly formalise these observations beyond the verbal description of the figures.

4 Zipfian Laws beyond Human Languages

The quantitative linguistics research on Zipfian laws is embedded into research on animal communication systems, which, in turn, is embedded into the wider research on information carrying systems in general, including those investigated in molecular biology (e.g. DNA, RNA, Proteins) and ecology (e.g. species distributions). Zipfian laws also emerge in music (cf. Zipf, 1949, p. 337) and – beyond communicative systems – in population sizes for metropolitan districts (Zipf, 1949, p. 375) and income distributions of countries (Zipf, 1949, p. 485).

Semple et al. (2022) provide an overview of the research on linguistic laws beyond human languages.[5] For example, Zipf's law of abbreviation has been investigated for the duration-frequency relationship in Formosan macaques' vocal communication (Semple et al., 2010), male gibbon morning songs (Huang et al., 2020), songs of lemurs (Valente et al., 2021), vocalizations in bats (Luo et al., 2013), as well as songs of various whale species (Arnon et al., 2025; Youngblood, 2025).

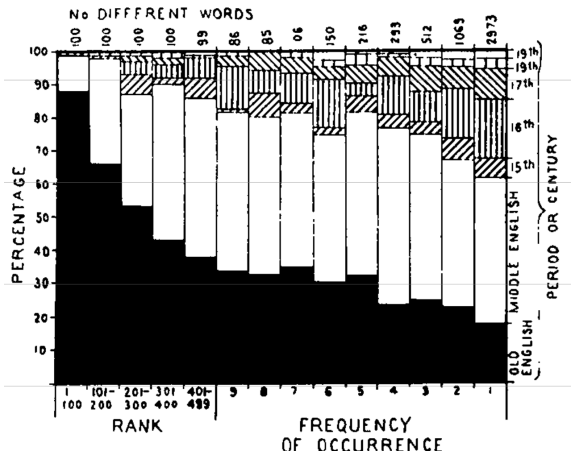

Figure 4 Visualisation of "chronological strata" for English words. The x-axis gives bins of ranks (1-100, 101-200, etc.) up to ranks 401-499. For higher ranks the word types are binned by their frequency (9 down to 1). The numbers of different word types per bin are given on top of the figure. Words are further sub-binned and coloured by their period or century of origin (Old English: black, Middle English: white, 15th century: oblique hatching, 16th century: vertical hatching, etc.). The percentages of words originating in each period/century are given on the y-axis. Reprinted from Zipf (1949, p. 111) Figure 3-10 panel A.

Eventually, insights from comparative studies across species and biological domains will help to empirically delineate statistical properties which are part of information carrying sequences in general, from those which are specific to communication systems – human languages as a special case.

5 Possible Explanations for Zipfian Laws

Zipf (1942, 1949) proposed early on that the laws he discovered are related to universal natural principles of *unity* and *diversity* – which also surface in efficient communication. For example, he discusses the trade-off between having just a single word for all meanings, as opposed to having unique words for each meaning (Zipf, 1942, p. 58). The former communication system would save "mental effort" when producing a word, but lack precision for understanding the "mental reference"; while the latter communication system would suffer from the inverse problem. Actual word meaning distributions are then a balance between these forces of *unification* and *diversification*.

These ideas have quickly come under criticism since (some) Zipfian laws are argued to occur also in random typing (Miller, 1957). Some of the issues with the random typing argument are discussed in Ferrer-i-Cancho (2007) and Ferrer-i-Cancho et al. (2022). One issue is that, counter-intuitively, random typing is indeed an example of *optimal* coding (Ferrer-i-Cancho et al., 2022). Another issue is that random typing, regardless of its information-theoretic status, is not a plausible baseline for human cognitive processes (cf. Ferrer-i-

[5] See also the bibliography compiled at `https://cqllab.upc.edu/biblio/laws/`.

Cancho & Elvevåg, 2010; Piantadosi, 2014). To overcome the latter issue, Petrini et al. (2023) establish a range of more plausible baselines for the law of abbreviation, and show that natural languages systematically deviate from these.

More generally, the criticisms of Zipf's work are levelled at the laws which are most prominent and directly measurable on sequences parsed into tokens – the law of word frequencies and the law of abbreviation. However, it should not be forgotten that Zipf has proposed a whole range of *interconnected* laws and potential explanations for these. As an example, the law of meaning distribution and the laws of word age cannot simply be attributed to stochastic processes – since randomly generated sequences at a given point in time have neither meanings nor ages.

At a minimum, interactive components are needed in any model explaining the emergence of Zipfian laws over time in communication systems (cf. Kanwal et al., 2017). It is exactly the *interconnectedness* of laws which has given rise to the framework of *synergetic linguistics* (cf. Köhler, 2014) as a branch of traditional quantitative linguistics.

5.1 Conclusion

In summary, the law of word frequencies is the most widely recognized law carrying Zipf's name – often simply referred to as *Zipf's law*. Taken together, thousands of languages and texts have been investigated for its presence. However, whether it is truly without exception depends on the definition of what it means for a statistical law "to hold". The strong version of the law with $\alpha \approx 1$ cannot be upheld, as many languages and texts systematically diverge from this value.

The *law of abbreviation* is of similar prominence. It has been established to hold in thousands of languages and texts without notable exceptions. It thus emerges as a strong candidate for a statistical – potentially even absolute – universal. The *law of meaning distribution* is somewhat less prominent. It has currently been investigated across less than one hundred languages. Finally, the *law of word returns* and the *laws of word age* have gone largely unnoticed, and are rarely – if ever – mentioned in the context of Zipfian laws.

Turning back to the discussion of *absolute* versus *statistical* universals: Zipfian laws constitute *absolute* universals in Bickel (2011)'s sense *if* if it holds true that they inevitably derive from our a priori assumptions regarding fundamental properties of sequence generation, such as the usage of "intermittend silences" (Miller, 1957). Of course, this would render them less interesting for the language sciences in particular.

Current research suggests that the laws derive from general principles of information transfer (Ferrer-i-Cancho et al., 2022; Semple et al., 2022), but they do not necessarily hold without exception. The task for linguistic and animal communication research is then to delineate the situations in which they surface from those in which they do not. This research paradigm holds promise to provide a general typology of information carrying systems, and, in particular, to establish statistical commonalities and differences between animal and human communication.

Appendix. Historical notes.

This Appendix gives the version of the *law of word returns* as found in the original publication by Zipf (1949, p. 41), and derives the formulation in the main text of this chapter from it.

The original formulation is

$$N^p \times I_f = \text{a constant (approximate)},$$

where N is the "number of intervals of a like I size", and f "refers to the frequency of occurrence of the different words of like frequency whose intervals are being measured". The exponent p "represents graphically the absolute slope of the line fitted to the points when the data are plotted on doubly logarithmic paper with N on the abscissa and I_f on the ordinate". If we take C to represent the constant, then this can be rearranged to

$$N^p = C \times I_f^{-1},$$

and via taking the p-th root of both sides to

$$N = \sqrt[p]{C \times I_f^{-1}} = \sqrt[p]{C} \times I_f^{-\frac{1}{p}}.$$

If we use the proportionality sign (abstracting way from the concrete constant) we have

$$N \propto I_f^{-\frac{1}{p}}.$$

Note that $P(\tau_w)$ is a normalized version of the count N, such that $P(\tau_w) = \frac{N}{f-1}$ for interval lengths of a given word type. Further note that I represents the interval lengths for a given word type and hence corresponds to τ_w in Equation 5. While Zipf restricted his original analyses to word types of particular frequencies f, he later proposes that the power law relationship holds for word types in general. With these correspondences we get to

$$P(\tau_w) \propto \tau_w^{-\omega},$$

with $\omega = \frac{1}{p}$.

Acknowledgements

I would like to thank the editor Ramon Ferrer-i-Cancho as well as two anonymous reviewers for detailed comments on this entry. 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, Zipf's law of word frequencies, Zipf's law of abbreviation, Zipfian laws of meaning, Word Returns

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, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE*, *4*(11), e7678.

Arnon, I., Kirby, S., Allen, J. A., Garrigue, C., Carroll, E. L., & Garland, E. C. (2025). Whale song shows language-like statistical structure. *Science*, *387*(6734), 649–653.

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. doi: 10.1371/journal.pone.0053227.

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

Gruyter GmbH & Co KG.
Bentz, C., & Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics.* Universität Tübingen.
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.
Bickel, B. (2011). Absolute and statistical universals. In P. C. Hogan (Ed.), *The Cambridge Encyclopedia of the Language Sciences* (p. 77-79). Cambridge: Cambridge University Press.
Bond, F., Janz, A., Maziarz, M., & Rudnicka, E. (2019). Testing Zipf's meaning-frequency law with wordnets as sense inventories. In P. Vossen & C. Fellbaum (Eds.), *Proceedings of the 10th Global Wordnet Conference* (pp. 342–352). Wroclaw, Poland: Global Wordnet Association.
Casas, B., Hernández-Fernández, A., Català, N., Ferrer-i Cancho, R., & Baixeries, J. (2019). Polysemy and brevity versus frequency in language. *Computer Speech & Language*, *58*(1), 19–50.
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.
Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain sciences*, *32*(5), 429–448.
Ferrer-i-Cancho, R. (2005). The variation of Zipf's law in human language. *European Physical Journal B*, *44*(1), 249-257. doi: 10.1140/epjb/e2005-00121-8.
Ferrer-i-Cancho, R. (2007). On the universality of Zipf's law for word frequencies. In P. Grzybek & R. Köhler (Eds.), *Exact methods in the study of language and text* (p. 131-140). Berlin: Gruyter. doi: 10.1515/9783110894219.131.
Ferrer-i-Cancho, R. (2016). The meaning-frequency law in Zipfian optimization models of communication. *Glottometrics*, *35*(1), 28-37.
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., & Vitevitch, M. (2018). The origins of Zipf's meaning-frequency law. *Journal of the American Association for Information Science and Technology*, *69*(11), 1369–1379.
Grzybek, P. (2007). History and methodology of word length studies. In P. Grzybek (Ed.), *Contributions to the science of language* (p. 15-90). Boston/Dordrecht/London: Kluwer Academic Publishers.
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.
Huang, M., Ma, H., Ma, C., Garber, P. A., & Fan, P. (2020). Male gibbon loud morning calls conform to Zipf's law of brevity and Menzerath's law: insights into the origin of human language. *Animal Behaviour*, *160*(1), 145–155.
Kaeding, F. W. (Ed.). (1898). *Häufigkeitswörterbuch der deutschen sprache: Festgestellt durch einen arbeitsausschuß der deutschen stenographiesysteme*. Steglitz bei Berlin: Mittler & Sohn.
Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*(1), 45–52.
Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: Papers* (pp. 79–86). Phuket, Thailand.
Köhler, R. (2014). Laws of language and text in quantitative and synergetic linguistics. In B. Szmrecsanyi & B. Wälchli (Eds.), *Aggregating dialectology, typology, and register analysis. linguistic variation in text and speech* (pp. 426–450). Berlin/Boston: Walter de Gruyter.
Koshevoy, A., Miton, H., & Morin, O. (2023). Zipf's law of abbreviation holds for individual characters across a broad range of writing systems. *Cognition*, *238*(), 105527.
Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of Zipfian distributions in language. *Cognition*, *223*(1), 105038.
Lavi-Rotbain, O., & Arnon, I. (2023). Zipfian distributions in child-directed speech. *Open Mind*, *7*(), 1–30.
Levshina, N. (2022). Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy*, *24*(2), 280.
Luo, B., Jiang, T., Liu, Y., Wang, J., Lin, A., Wei, X., & Feng, J. (2013). Brevity is prevalent in bat short-range communication. *Journal of Comparative Physiology A*, *199*(1), 325–333.
Mandelbrot, B. (1965). Information theory and psycholinguistics. In B. B. Wolman (Ed.), *Scientific psychology*. Basic Books Publishing.
Mayer, T., & Cysouw, M. (2014). Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014* (pp. 3158–3163).
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.
Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, *70*(2), 311–314.
Moreno-Sánchez, I., Font-Clos, F., & Corral, Á. (2016). Large-scale analysis of Zipf's law in English texts. *PLoS ONE*, *11*(1), e0147073.
Nagata, R., & Tanaka-Ishii, K. (2025, July). A new formulation of Zipf's meaning-frequency law through contextual diversity. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 15323–15335). Vienna, Austria: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.acl-long.744/ doi: 10.18653/v1/2025.acl-long.744.
Petrini, S., Casas-i Muñoz, A., Cluet-i Martinell, J., Wang, M., Bentz, C., & Ferrer-i Cancho, R. (2023). Direct and indirect evidence of compression of word lengths. Zipf's law of abbreviation revisited. *Glottometrics*, *54*(1), 58-87.
Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(), 1112–1130.
Piantadosi, S. T., & Gibson, E. (2013). Quantitative standards for absolute linguistic universals. *Cognitive Science*, *38*(4), 736–756.
Popescu, I.-I., Naumann, S., Kelih, E., Rovenchak, A., Overbeck, A., Sanada, H., … et al. (2013). Word length: aspects and languages. In *Issues in quantitative linguistics* (Vol. 3, pp. 224–281). Lüdenscheid: RAM Verlag.
Semple, S., Ferrer-i Cancho, R., & Gustison, M. L. (2022). Linguistic laws in biology. *Trends in Ecology & Evolution*, *37*(1), 53–66.
Semple, S., Hsu, M. J., & Agoramoorthy, G. (2010). Efficiency of coding in macaque vocal communication. *Biology letters*, *6*(4), 469–471.
Tanaka-Ishii, K. (2021). *Statistical universals of language: Mathematical chance vs. human choice*. Cham, Switzerland: Springer.
Valente, D., De Gregorio, C., Favaro, L., Friard, O., Miaretsoa, L., Raimondi, T., … et al. (2021). Linguistic laws of brevity: conformity in Indri indri. *Animal cognition*, *24*(4), 897–906.
Van Egmond, M. (2018). *Zipf's law in aphasic speech: An investigation of word frequency distributions*. Utrecht: LOT.
Youngblood, M. (2025). Language-like efficiency in whale communication. *Science Advances*, *11*(6), eads6014.
Zipf, G. K. (1942). The unity of nature, least-action, and natural social science. *Sociometry*, *5*(1), 48–62.
Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology*, *33*(2), 251–256.
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.
Zörnig, P. (2010). Statistical simulation and the distribution of distances between identical elements in a random sequence. *Computational statistics & data analysis*, *54*(10), 2317–2327.