Zipf's Law of Abbreviation

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

Words which occur more often tend to be shorter. This is the core proposition behind *Zipf's law of abbreviation*, also known as *the law of brevity*. This chapter gives a brief historical overview of the research on this law since its inception by George Kingsley Zipf. Methodological issues and caveats relating to estimating probabilities and measuring lengths of words are discussed. The chapter also draws an important distinction between a) studies on word length distributions, on one hand, and b) studies on the relationship between word frequencies and lengths, on the other. Furthermore, two lines of methodological approaches to the law of abbrevation are outlined: model fitting and correlation analyses. The chapter concludes by a discussion of possible explanations for the law of abbreviation, and by linking the respective research on natural languages with research on animal communication systems.

Keywords: Word lengths, Zipf's law of abbreviation, brevity law

Key Points

- Brief historical overview of research on the law of brevity.
- Outline of data preparation steps necessary for testing the law.
- Distinction between studies on word length distributions and studies on the relationship between frequency and length for particular word types.
- Explanation of correlation metrics and statistics.
- Brief discussion of animal communication research harnessing the law of abbreviation.
- Brief discussion of possible explanations for Zipf's law of abbrevation.

1 Introduction

To illustrate the relationship between word frequencies and word lengths take the two example sentences below, stemming from the English *Universal Declaration of Human Rights* (UDHR). The word types 'and', 'in', as well as 'are' have the highest frequencies. Namely, 'and' occurs four times ($f_{\text{and}} = 4$), while 'are' and 'in' occur two times respectively ($f_{\text{are}} = 2$, $f_{\text{in}} = 2$). The other word types in this example occur only once – sometimes referred to as *hapax legomena*. Note that "frequency" here refers to *absolute/raw frequency*, i.e. a count without normalization. These counts can be normalized by dividing them by the overall number of tokens. This yields *relative* frequencies: $f_{\text{and}} = \frac{4}{30}$, $f_{\text{are}} = \frac{2}{30}$, $f_{\text{in}} = \frac{2}{30}$, etc.

Moreover, these word types have lengths in characters of three ($l_{\text{and}} = 3$, $l_{\text{are}} = 3$) and two ($l_{\text{in}} = 2$) respectively. While some of the *hapax legomena* are here of similar lengths as the most frequent items (e.g. 'All', 'act', 'one'), most of them are longer (e.g. 'human', 'beings', 'conscience').

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

This negative association between frequency and length is not just a phenomenon of this particular text – or English writing for that matter. It rather extends to other languages and modalities as well. Table 1 gives the type frequencies and lengths for three typologically very different languages (Abkhaz, English, Hawaiian). The same data is visualized in Figure 1. Words of high frequency (horizontal axes) quite generally tend to be shorter (vertical axes) across the different languages. A negative association between the two measures is visible in the panels – though the strength of the effect might differ.

2 History

In the late 19th century, the stenographer Friedrich Wilhelm Kaeding led a project to collect frequencies as well as lengths (in syllables, consonants and vowels) for German words. The result was published in a 671 page volume entitled *Häufigkeitswörterbuch der Deutschen Sprache* (Kaeding, 1898). Based on this extensive data, Kaeding (1898, p. 650) noticed that the average length of function words in German is markedly lower (1.27 syllables/word) than for content words (2.95 syllables/word). Among a plethora of statistics, the table in Figure 2 is found in Kaeding's volume as well. In the leftmost column, it gives the length in syllables (monosyllabic, bisyllabic, trisyllabic, etc.). In the second column from the left, the cumulated frequencies of words with a given length in syllables are provided. In sum, more than 10 million word tokens were counted. This table is reprinted in Zipf's early work, followed by the statement:

> These figures indicate that in German there is a decided preference in usage for short words, and that the magnitude of words stands in an inverse (not necessarily proportionate) relationship to the number of occurrences [...]
> ***–Zipf (p. 23 1965, first published 1935)***

This constitutes one of the first formulations of the so-called *law of abbreviation* or *law of brevity*. Zipf subsequently cites further collection efforts, such as the *Six Thousand Common English Words* by Eldridge (1911), as well as his own earlier work with statistics for Classical Latin texts of Plautus, as well as Chinese as spoken in Beijing (Zipf, 1932). This serves him as evidence that the law of brevity holds across different text types and languages.

In the times of Kaeding, Eldridge, and Zipf, frequency and length statistics were collected by painstaking manual work. The project by Kaeding involved 110 collection offices employing close to 1000 workers across Germany. With the advent of the digital age, text collections and processing tools became available across a variety of languages. For example, Bentz & Ferrer-i Cancho (2016) have illustrated that a negative correlation between lengths (in characters) and frequencies of orthographic words (Table 1) holds without exception

Bentz, Christian (2026). Zipf's Law of Abbreviation. 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 lengths for orthographic word types 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.	Length	Word	Frequency	Length	Word	Freq.	Length
ka 'the'	82	2	the	65	3	ауаҩы 'human'	39	5
i 'at, in, on'	75	1	and	57	3	дарбанзаалак 'every'	34	12
a 'of, by'	51	1	to	57	2	азин 'right'	30	4
ke 'the'	42	2	of	46	2	иара 'he'	23	4
nā 'the (plural)'	42	2	right	28	5	имоуп 'has'	23	5
...	...	...	...	...	...	...	...	...
waena 'middle'	1	5	women	1	5	шьаҭа 'basis'	1	5
wahi 'place'	1	4	working	1	7	шьаҭас 'based on'	1	6
wahine 'woman'	1	6	works	1	5	шьҭнахуазароуп 'it should promote'	1	14
wāhine 'women'	1	6	worship	1	7	ыказароуп 'it should be'	1	9
wai 'water'	1	3	worthy	1	6	ыкамзароуп 'it should not be'	1	10

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

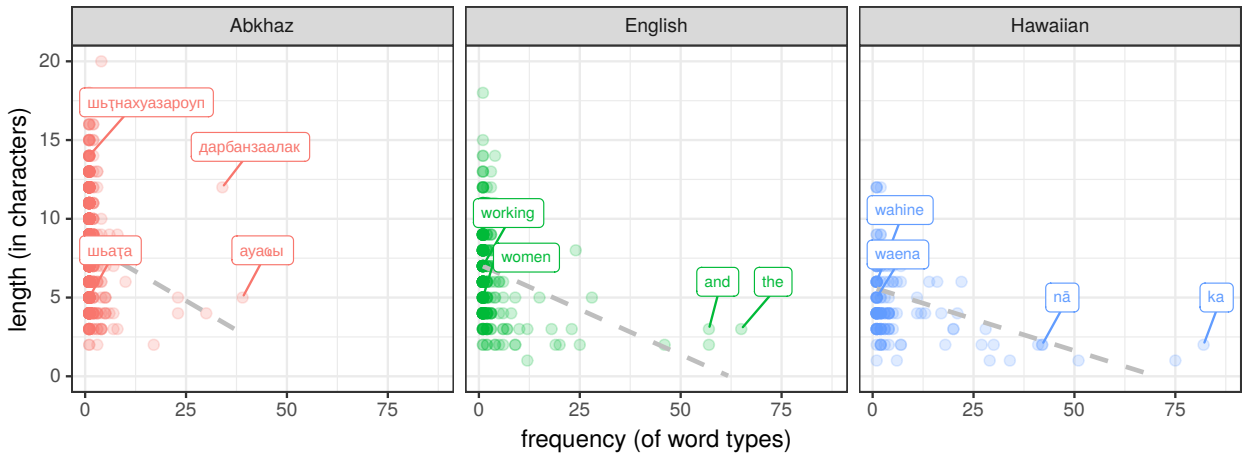

Figure 1 Type frequencies (x-axes) versus length in characters (y-axes) for orthographic words in Abkhaz, English, and Hawaiian. Each dot corresponds to a word type. The two most frequent words as well as two *hapax legomena* (of frequency one) are labelled for each language. A linear model fitted to the data is overplotted (gray dashed line). This data is taken from the UDHR, see also Table 1. Beware confusion: From an information-theoretic perspective, lengths are seen as a function of frequencies, i.e. "responding" to frequency changes, rather than the other way round. However, the direction of causality is still an open problem, and in some publications the axes are inverted.

across parallel texts featuring 986 languages. The law of abbreviation is hence one of the most robust candidates for a language universal.

3 Data Preparation

Zipf (1965, p. 15) himself was already aware that counting words and measuring their "magnitudes" hinges upon some fundamental assumptions. Firstly, the items have to be defined for which the frequency/probability is estimated. Secondly, a measure of the length/duration has to be applied to these items. For a broader discussion of the issues and caveats surrounding word length studies see also Grotjahn & Altmann (1993).

3.1 Estimating Frequencies/Probabilities

Many studies on the law of abbreviation use orthographic words as items of the vocabulary, such that we have a set $\mathcal{V} = \{w_1, w_2, \ldots, w_n\}$, where w_i is a given word type. Typically, their *relative frequencies* are used as an approximation of their *probability*, such that for each word type we have:

$$\hat{p}_i = \frac{f_i}{\sum_{i=1}^{n} f_i},$$

with f_i being the absolute frequency of a given word, and n the size of the vocabulary $n = |\mathcal{V}|$. There is a wide range of more complex estimators for word probabilities – applying different smoothing strategies (see Bentz, 2018, for a discussion in the context of entropy estimation). Also, rather than estimating probabilities for words as

	Wörter	Prozent	Silben
1 silbig	5 426 326	49,76	5 426 326
2 "	3 156 448	28,94	6 312 896
3 "	1 410 494	12,93	4 231 482
4 "	646 971	5,93	2 587 884
5 "	187 738	1,72	938 690
6 "	54 436	0,50	326 616
7 "	16 993		118 951
8 "	5 038		40 304
9 "	1 225		11 025
10 "	461		4 610
11 "	59	0,22	649
12 "	35		420
13 "	8		104
14 "	2		28
15 "	1		15
	10 906 235*)		20 000 000

Figure 2 Original table of word lengths in syllable numbers, i.e. monosyllabic, bisyllabic, trisyllabic, etc. (leftmost column), and cumulative word frequencies (second column to the left under "Wörter", i.e. words), reprinted from Kaeding (1898, p. 24).

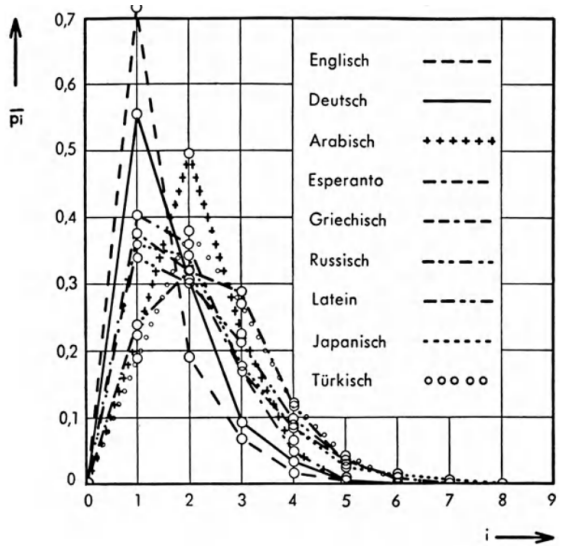

Figure 3 Word length distributions. Length is here measured in syllables (denoted i instead of l_i) on the horizontal axis. Probabilities (relative frequencies) per length are given on the vertical axis. This includes data from eight natural languages and one artificial language (Esperanto). Reprinted from Fucks (1955, p. 85). Note that these distributions have probability-maxima at $l_i = 1$ and $l_i = 2$ depending on the respective language. In English and German, monosyllabic words are most probable. In Arabic and Turkish, bisyllabic words are most probable.

independently and identically distributed (i.i.d.) variables, another approach is to estimate their probabilities in context, e.g. by applying n-gram models (Piantadosi et al., 2011; Kalimeri et al., 2015; Meylan & Griffiths, 2021; Levshina, 2022), or newer flavors of language modelling (e.g. Pimentel et al., 2023). Interestingly, however, using more complex models for estimating word probabilities does not necessarily lead to better correlations with their lengths (cf. Meylan & Griffiths, 2021; Pimentel et al., 2023).

Table 2 Lengths per unit for the English word 'limitation', translated also into Mandarin Chinese. Phonemes are given in IPA. Strokes are given as numbers of strokes per character in writing.

Example	Unit	Length
English: limitation		
limit-ation	morpheme	2
li-mi-ta-tion	syllable	4
l-ɪ-m-ɪ-t-e-ɪ-ʃ-ə-n	phoneme	10
l-i-m-i-t-a-t-i-o-n	character	10
1-2-3-2-2-2-2-2-1-2	stroke	19
Chinese (simplified): 限度		
限-度	morpheme	2
限-度	syllable	2
ɕ-i-à-n-t-ù	phoneme	6
限-度	character	2
8-9	stroke	17
Chinese (Pinyin): xiàndù		
xiàn-dù	morpheme	2
xiàn-dù	syllable	2
ɕ-i-à-n-t-ù	phoneme	6
x-i-à-n-d-ù (x-i-a-`-n-d-u-`)	character	6 (8)
2-2-3-2-2-3 (2-2-2-1-2-2-2-1)	stroke	14 (14)

Source: The Chinese translation is provided by https://www.mdbg.net/chinese/dictionary.

3.2 Measuring Length

Given an estimation of probabilities per item in the vocabulary, a measure of magnitude (strokes, characters, phonemes, syllables, morphemes, duration) is needed. Typically, lengths of words are derived from textual material. These can differ considerably for a given word, depending on the writing system and the unit of measurement (see Table 2 for detailed examples). For example, the English word 'limitation' consists of two morphemes (limit-ation), but it takes 19 strokes to write it with a pen. In alphabetic writing systems, the length in characters will be a close approximation to the length in phonemes, but in Chinese writing it is less so (except for Pinyin). See Petrini et al. (2023) for further discussions of the methodological issues with measuring word lengths – in both written characters and durations in speech.

4 Two Conceptualizations of the Law of Abbreviation

In contrast to other laws, Zipf (1965) did not explicitly provide a formula for the law of abbreviation, i.e. how to capture the relationship between the "magnitude of words" and the "number of occurrences" mathematically. In this context, there is a common confusion between two strands of research:

- Research on *word length distributions*, that is, how often certain word lengths occur in a text (Figure 2 and Figure 3);
- research on the relationship between *frequencies* and respective *lengths* for *particular* word types (Table 1 and Figure 1).

Importantly, in the former case, the frequencies are calculated for *word lengths* – how often does a given length occur in a text? Here, the token counts of word types of a given length are cumulated. In the latter case, the frequencies are calculated for *particular* word types.

Note that in Figure 3 each dot corresponds to a given length and the probability of this length given a *cumulative word count*. For ex-

ample, the token count for all word types of length one. In Figure 1, on the other hand, each dot corresponds to the frequency and length of *exactly one word type*. *A priori* these are two different conceptualizations of the relationship between lengths and frequencies of word types.

Notably, Corral & Serra (2020) propose to link these two conceptualizations by assuming that there is an underlying bivariate joint probability mass function $g(l, f)$ representing the relationship between word lengths and word frequencies. The word frequency distribution $g(f)$ and the word length distribution $g(l)$ are then the univariate marginals of $g(l, f)$.

4.1 Word Length Distributions

There is a long list of publications on *word length distributions*. Since the mid 20th century different researchers have proposed and fitted a panoply of mathematical functions to the distributions of word lengths in various languages and texts. This includes the Poisson distribution, the lognormal distribution, the negative binomial distribution, the gamma distribution, etc. This constitutes a core area of research in the field of traditional *Quantitative Linguistics*. For an overview of this research see, for instance, Grzybek (2007) and Popescu et al. (2013). As a matter of fact, the study of word lengths is considerably older – reaching back to the mid 19th century (Grzybek, 2007, p. 15) – than the first formulations of the law of abbreviation by Zipf in 1935. In Table 4, a historical overview is provided. Studies on *word length distributions* are explicitly marked in gray.

4.2 The Frequency/Length Relationship for Particular Word Types

In the following, this chapter focuses on research into the relationship between frequencies and lengths of particular word types. This is what Zipf (1965, p. 38) refers to in his conclusion on *the law of abbreviation*: "In view of the evidence of the stream of speech we may say that the length of a word tends to bear an inverse relationship [footnote: not necessarily *proportionate*, possibly some nonlinear mathematical function] to its relative frequency; [...]". Note that this refers to *particular word types* and the relationship between their frequencies and lengths.

5 Statistical Models to Investigate the Frequency/Length Relationship

For assessing the relationship between relative frequencies of word types and their lengths there are, in general, two options: firstly, *hard* or *strong* versions of the law, namely, defining and fitting *mathematical equations* with a given number of parameters, and secondly, *soft* or *weak* versions of the law, that is, applying *correlation* metrics.

5.1 Model Fitting

Two major steps are necessary for model fitting in the context of Zipfian and other quantitative laws: a) define a mathematical equation which allegedly fits the respective relationship to be investigated; b) fit the model to the data – assessing goodness of fit. See Altmann & Gerlach (2016) for an in-depth discussion of model fitting in the context of linguistic laws in general.

Table 3 Estimations of λ_D values across different languages and measurement units of length.

Language	$\lambda_D^{\text{phonemes}}$	$\lambda_D^{\text{characters}}$	$\lambda_D^{\text{durations}}$
Catalan	0.49	0.53	23.8
English	0.5	0.6	20.6
Spanish	0.56	0.6	24.1

Source: Torre et al. (2019) and Hernández-Fernández et al. (2019).

5.1.1 Optimal Coding

In the context of model fitting, some research proposes to use mathematical formulations rooted in the wider framework of information theory (Hernández-Fernández et al., 2019; Torre et al., 2019). According to standard information theory (Cover & Thomas, 2006, p. 111) the optimal lengths l_i^* of information encoding items (i) with probabilities p_i are given by

$$l_i^* = -\log_D p_i,$$

where D is the size of an alphabet (or dictionary/vocabulary) of symbols. In the case of written English, this could be the set of word types in a given text – or other structural units (see also Section 6). Torre et al. (2019, p. 13) propose to model deviations from optimality by using a pre-factor which depends on the size of the alphabet, here termed λ_D, which gives

$$l_i \sim -\frac{1}{\lambda_D} \log_D p_i,$$

where $0 < \lambda_D \leq 1$. Relative frequencies can be plugged in as estimated probabilities $\hat{p}_i$. Rearranging the equation then gives

$$\hat{p}_i \sim D^{-\lambda_D l_i}.$$

Torre et al. (2019, p. 12) fit this equation to an English speech corpus, and show that measuring lengths of words in phonemes vs. characters in writing yield very similar results, while durations in time yield a different parameter-estimate (see Table 3). Applying the same methodology, Hernández-Fernández et al. (2019, p. 9) provide the estimates also for spoken Catalan and Spanish.

5.2 Correlation Metrics

As an alternative to – often cumbersome – model fitting, correlation metrics are commonly used to assess the law of abbreviation. This includes, in particular, the Pearson r, Spearman ρ, and Kendall τ coefficients. All of them fall in the range $[-1, 1]$, with *minus one* indicating a perfect negative correlation, *one* a perfect positive correlation, and *zero* a genuine lack of correlation. If the law of abbreviation applies for probabilities and lengths of items then negative correlation coefficients are expected.

In the following, simple definitions of these three metrics are given – while there are several variants found in the literature.[1]

5.2.1 Pearson correlation

The Pearson correlation coefficient for a given sample of items and their respective probabilities (p_i) and lengths (l_i) is defined as (cf. Ferrer-i-Cancho et al., 2022, p. 179):

[1] Most of the correlation metrics are implemented in common packages of statistical software such as R (R Core Team, 2024). The match between formula and implementation is normally given in the function's documentation.

$$r = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(l_i - \bar{l})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2}\sqrt{\sum_{i=1}^{n}(l_i - \bar{l})^2}},$$

where $\bar{p}$ and $\bar{l}$ are the mean probability and length across the sample. We have $r = -1$ if there is a perfectly *linear* and *negative* relation, and $r = 1$ if there is a perfectly *linear* and *positive* relation between probabilities and lengths. For Zipf's law of abbreviation to hold we expect $r < 0$.

5.2.2 Spearman Correlation

A non-parametric alternative is the Spearman rank correlation coefficient (Spearman, 2010, originally published in 1904). The p_i and l_i values are here *ranked* (from lowest to highest by convention). In other words, a function rank() maps values to ranks, i.e. $\text{rank}(p_i)$ and $\text{rank}(l_i)$. Importantly, the ranks need to be unique, i.e. ties in values have to be broken. The rank coefficient is then defined as (Hollander et al., 2014, p. 428)

$$\rho = 1 - \frac{6 \times \sum_{i=1}^{n} {d_i}^2}{n(n^2 - 1)}, \quad [1]$$

where d_i is the difference in the ranks, i.e. $d_i = \text{rank}(l_i) - \text{rank}(p_i)$, and n is the number of ranks (i.e. number of word types). Imagine that for each word w_i the rank in terms of length is exactly equivalent to the rank in terms of probability. In this case, all the d_is are zero, and we have $\rho = 1$. In other words, the probabilities and lengths entirely agree in terms of their rankings, such that there is a perfect positive relationship between them. On the other hand, if the ranks of probabilities are the exact inverse of the ranks of lengths, then the ratio in Equation 1 will equal *two*, and hence $\rho = -1$. Note that since this statistic is solely based on rankings, it is not necessarily reflecting a *linear* relationship between the values. Notice that Equation 1 is the definition of Spearman rank correlation when there are no ties in the data.

5.2.3 Kendall τ

Another non-parametric alternative is the Kendall rank correlation coefficient (Kendall, 1938). It is conceptually similar to the Spearman rank correlation, but works with so-called *concordant* and *discordant* pairs of values, rather than overall rankings. A simple definition can be given as (cf. Ferrer-i-Cancho et al., 2022, p. 176):

$$\tau = \frac{n_c - n_d}{\binom{n}{2}},$$

where n_c is the number of *concordant* pairs, and n_d is the number of *discordant* pairs. In a nutshell, we have a *concordant* pair if for $p_i > p_j$ we have $l_i > l_j$ and if for $p_i < p_j$ we have $l_i < l_j$, i.e. more likely elements of the vocabulary are longer. In contrast, we have a discordant pair if for $p_i > p_j$ we get $l_i < l_j$ and if for $p_i < p_j$ we get $l_i > l_j$, i.e. more likely elements are shorter. The difference between the concordant and discordant pairs is normalised by the overall number of pairs that can be chosen from the sample of n items, i.e. "n over 2". Intuitively, if the law of abbreviation holds, we expect to find discordant pairs. Note that there are more complex versions of Kendall τ which take into account ties in the data.

6 Beyond Written Words

This chapter has mostly considered orthographic words as the items for which probabilities and lengths are calculated. However, already in the original work by Zipf other structural features, e.g. syllables, were investigated as well. In studies since then, various structural levels (characters, inflectional markers, words etc.) have been checked for the presence of the law. Likewise, diverse units of "magnitude" have been considered (e.g. characters, strokes, phonemes, syllables, durations, etc.). Table 4 gives a condensed overview of selected studies on the law of abbreviation since Zipf's time.

7 Animal Communication

Current efforts to measure magnitudes beyond written words, e.g. durations in spoken language (Torre et al., 2017, 2019; Petrini et al., 2023), open up the possibility for inter-species comparisons. A growing body of research investigates linguistic laws in animal communication and other biological systems. For a general overview see Semple et al. (2022). The law of brevity in particular has been shown to emerge in the frequencies of call types and their durations for various species. This includes the duration-frequency relationship in Formosan macaques' vocal communication (Semple et al., 2010), male gibbon morning songs (Huang et al., 2020), and songs of lemurs (Valente et al., 2021). For vocal communication in different whale species, the results are mixed. Arnon et al. (2025) detect the law of brevity in humpback whale songs, while Youngblood (2025) can confirm it only in two out of five species investigated.

The law of brevity does not surface in calls of golden-backed uakaris and common marmosets (Bezerra et al., 2011). The failure to detect a significant duration-frequency relationship might here be related to small sample sizes (Ferrer-i Cancho & Hernández-Fernández, 2013). Interestingly, in the case of vocal communication in bats, the law surfaces for social calls, but not for distress vocalizations (Luo et al., 2013). In this case, the discrepancy does not seem related to data sparsity.

In summary, the law of brevity is not a universal of animal communication systems in general. On the other hand, it is also not unique to human speech and writing. It is a statistical property emerging in a subset of communication systems including human languages as well as some animal communication systems.

8 Possible Explanations

Along the same lines as for the *law of word frequencies*, Zipf (1949) interpreted the negative relationship between frequencies and lengths of words in the light of the *principle of least effort*. As speakers utter certain words more often, they reduce the overall effort by shortening those. Zipf's ideas on relating his statistical laws to coding efficiency have quickly come under criticism. Miller (1957, p. 314) illustrated that similar relationships hold for sequences of characters generated by random typing models, and that, as a consequence, their explanation does not have to invoke "least effort, least cost, maximal information [...]".

Note that standard information theory predicts the law of brevity as a result of the *principle of compression*. Namely, optimal coding requires minimization of the average code length – given a particular

Table 4 Selection of studies on the law of abbreviation since Zipf's original proposal. Studies are ordered by year of publication. Note: Studies on the distributions of word lengths (see also the discussion in Section 5) are here marked in gray.

Study	Structure	Magnitude Unit	Model/Method	Modality	Languages
Zipf (1965), first edition 1935	words	syllables	tabulation	writing	Chinese, Germ., Lat.
ibid.	words	phonemes	tabulation	writing	English
Čebanow (1947)†	words	syllables	Poisson distr.	writing	12
Elderton (1949)	words	syllables	geometric distr.	writing	English
Fucks (1955)	words	syllables	gen. Poisson distr.	writing	9
Herdan (1958)	words	phonemes	lognormal distr.	speech	English
ibid.	words	characters	lognormal distr.	speech	English
Zerzwadse et al. (1962)	words	syllables	gen. Poisson distr.	writing	Georgian
Bartkowiakowa & Gleichgewicht (1964)	words	syllables	gen. Poisson distr.	writing	Polish
Vranić & Matković (1965)◇	words	syllables	gen. Poisson distr.	writing	Croato-Serbian
Guiter (1974)	words	phonemes	logarithmic	writing	8
Grotjahn (1982)	words	syllables	neg. binomial distr.	writing	German
Hammerl (1990)	words	syllables	log./power law	writing	Polish
Meyer (1997)	words	syllables	hyper-Poisson distr.	writing	Inuktitut
Rottmann (1997)	words	syllables	hyper-Poisson distr.	writing	Old Church Slav.
Sigurd et al. (2004)	words	characters	gamma distr.	writing	English, Swedish
ibid.	words	phonemes	gamma distr.	writing	English, Swedish
ibid.	words	syllables	gamma distr.	writing	German
ibid.	sentences	words	gamma distr.	writing	English
Pustet & Altmann (2005)	morphemes	phonemes	Gegenbauer distr.	writing	Lakota
Strauss et al. (2007)	words	syllables	power law	writing	11
Piantadosi et al. (2011)	words	characters	Spearman ρ	writing	11
Popescu et al. (2013)	words	syllables	various distr.	writing	28
Popescu et al. (2014)*	syllables	phonemes	Zipf-Alekseev distr.	writing	German
ibid.	morphemes	phonemes	Zipf-Alekseev distr.	writing	German
ibid.	words	various	Zipf-Alekseev distr.	writing	43
ibid.	compounds	stems	Zipf-Alekseev distr.	writing	German
ibid.	rhythmic units	syllables	Zipf-Alekseev distr.	writing	German
ibid.	verse	words	Zipf-Alekseev distr.	writing	German
ibid.	sentences	clauses	Zipf-Alekseev distr.	writing	German
ibid.	speech acts	no. acts	Zipf-Alekseev distr.	speech	German
Narisong et al. (2014)	words	syllables	Poisson distr.	writing	Mongolian
ibid.	stems	syllables	Poisson distr.	writing	Mongolian
Bentz & Ferrer-i Cancho (2016)	words	characters	Kendall τ	writing	986
Chen & Liu (2016)	words	characters	various distr.	speech	Chinese
ibid.	words	phonemes	various distr.	speech	Chinese
ibid.	words	syllables	various distr.	speech	Chinese
ibid.	words	stroke	various distr.	writing	Chinese
ibid.	words	component	various distr.	writing	Chinese
ibid.	words	character	various distr.	writing	Chinese
Torre et al. (2017)	voice events	duration	power law	speech	16
Hernández-Fernández et al. (2019)	words	characters	exponential	speech	Catalan, Spanish
ibid.	words	phonemes	exponential	speech	Catalan, Spanish
ibid.	words	durations	exponential	speech	Catalan, Spanish
Torre et al. (2019)	words	characters	exponential	speech	English
ibid.	words	phonemes	exponential	speech	English
ibid.	words	durations	exponential	speech	English
Corral & Serra (2020)	words	characters	gamma distr.	written	English
Meylan & Griffiths (2021)	words	characters	Spearman ρ	writing	13
Guzmán Naranjo & Becker (2021)	infl. markers	phonemes	hurdle Poisson	writing	60
Wei et al. (2021)	words	characters	Zipf-Alekseev distr.	writing	Zhuang
ibid.	words	syllables	Zipf-Alekseev distr.	writing	Zhuang
Levshina (2022)	words	characters	Spearman ρ	writing	7
ibid.	words	phonemes	Spearman ρ	writing	4
Koshevoy et al. (2023)	characters	complexity	linear regr.	writing	27 writing systems
Petrini et al. (2023)	words	characters	r, τ	writing	22
ibid.	words	durations	r, τ	speech	46
ibid.	words	strokes	r, τ	writing	Chinese, Japanese
Pimentel et al. (2023)	words	characters	Spearman ρ	writing	13

† Indo-European languages according to Čebanov, whose work is here referred to via Grzybek (2007, pp. 26).

◇ This work is referred to via Grzybek (2007, pp. 58).

* This publication is a meta-analysis of data published in various studies for particular languages. See the references therein.

coding scheme. Recently, it has been shown that – counter-intuitively – random typing is also an optimal coding process (Ferrer-i-Cancho et al., 2022). Namely, given a probability distribution over characters and white spaces, random typing will create codes which are optimal in terms of their length/frequency relationship. Thus, *both* random typing *and* natural languages display properties of optimal coding.

However, it has been pointed out that random typing is not a psychologically plausible model of human learning and usage of words – regardless of its information-theoretic status (Howes, 1968; Manin, 2009; Ferrer-i-Cancho & Elvevåg, 2010; Piantadosi, 2014). Rather, experiments have shown that the law of abbreviation arises in form-meaning mappings of artificial languages when users are under pressure for both brevity and accuracy (Kanwal et al., 2017).

While the law of abbreviation surfaces across diverse languages and structural levels (see Table 4), it is not without exception in communicative systems more broadly. For instance, Miton & Morin (2019) do not find a significant relationship for motif frequency and motif complexity in European heraldry. They argue that in this particular symbol system – communicating identities via coats of arms – *iconicity* stands in the way of abbreviation/simplification.

Likewise, as discussed above, the law does not hold without exception in animal communication systems. It is an interesting avenue for future research to establish biological, ecological or cultural conditions under which the law surfaces, and under which it does not. This will ultimately shed light on the underlying forces explaining the law of brevity.

Acknowledgements

I would like to thank the editor Ramon Ferrer-i-Cancho as well as one anonymous reviewer 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, Zipf's law of 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.

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.

Bartkowiakowa, A., & Gleichgewicht, B. (1964). Zastosowanie dwuparametrowych rozkładów fucksa do opisu długości sylabicznej wyrazów w różnych utworach prozaicznych autorów polskich. *Zastosowania matematyki*, *7*(), 345–352.

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.

Bezerra, B. M., Souto, A. S., Radford, A. N., & Jones, G. (2011). Brevity is not always a virtue in primate communication. *Biology letters*, *7*(1), 23–25.

Chen, H., & Liu, H. (2016). How to measure word length in spoken and written Chinese. *Journal of Quantitative Linguistics*, *23*(1), 5–29.

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

Corral, Á., & Serra, I. (2020). The brevity law as a scaling law, and a possible origin of Zipf's law for word frequencies. *Entropy*, *22*(2), 224.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons.

Elderton, W. P. (1949). A few statistics on the length of English words. *Journal of the Royal Statistical Society. Series A (General)*, *112*(4), 436–445.

Eldridge, R. (1911). *Six thousand common English words: Their comparative frequency and what can be done with them*. Niagra Falls, N. Y.: Clement Press.

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. (2013). The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory*, *4*(1), 45–55.

Fucks, W. (1955). *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen*. Springer Fachmedien Wiesbaden.

Grotjahn, R. (1982). Ein statistisches Modell für die Verteilung der Wortlänge. *Zeitschrift für Sprachwissenschaft*, *1*(1), 44–75.

Grotjahn, R., & Altmann, G. (1993). Modelling the distribution of word length: Some methodological problems. In *Contributions to Quantitative Linguistics: Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier, 1991* (pp. 141–153).

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.

Guiter, H. (1974). Les relationes frequence - longueur - sens des mots (langues romanes et anglais). In A. Varvaro (Ed.), *XIV Congresso Internazionale di linguistica e filologia romanza* (pp. 373–381). John Benjamins.

Guzmán Naranjo, M., & Becker, L. (2021). Coding efficiency in nominal inflection: expectedness and type frequency effects. *Linguistics Vanguard*, *7*(s3), 20190075.

Hammerl, R. (1990). Länge–Frequenz, Länge–Rangnummer: Überprüfung von zwei lexikalischen Modellen. *Glottometrika*, *12*(), 1–24.

Herdan, G. (1958). The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. *Biometrika*, *45*(1-2), 222–228.

Hernández-Fernández, A., G. Torre, I., Garrido, J.-M., & Lacasa, L. (2019). Linguistic laws in speech: The case of Catalan and Spanish. *Entropy*, *21*(12), 1153.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric statistical methods*. Hoboken, New Jersey: John Wiley & Sons.

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

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*(), 145–155.

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

Kalimeri, M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakonos, F. K., & Papageorgiou, H. (2015). Word-length entropies and correlations of natural language written texts. *Journal of Quantitative Linguistics*, *22*(2), 101–118.

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*(), 45–52.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

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.

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*(), 325–333.
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.
Meyer, P. (1997). Word-length distribution in Inuktitut narratives: Empirical and theoretical findings. *Journal of Quantitative Linguistics*, *4*(1-3), 143–155.
Meylan, S. C., & Griffiths, T. L. (2021). The challenges of large-scale, web-based language datasets: Word length and predictability revisited. *Cognitive Science*, *45*(6), e12983.
Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, *70*(2), 311–314.
Miton, H., & Morin, O. (2019). When iconicity stands in the way of abbreviation: No Zipfian effect for figurative signals. *PLoS ONE*, *14*(8), e0220793.
Narisong, Jiang, J., & Liu, H. (2014). Word length distribution in Mongolian. *Journal of Quantitative Linguistics*, *21*(2), 123–152.
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*(), 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., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.
Pimentel, T., Meister, C., Wilcox, E., Mahowald, K., & Cotterell, R. (2023). Revisiting the optimality of word lengths. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 2240–2255). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.137 doi: 10.18653/v1/2023.emnlp-main.137.
Popescu, I.-I., Best, K.-H., & Altmann, G. (2014). Unified modeling of length in language. *Studies in quantitative linguistics*, *2*(), 124.
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.
Pustet, R., & Altmann, G. (2005). Morpheme length distribution in Lakota. *Journal of Quantitative Linguistics*, *12*(1), .
R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
Rottmann, O. A. (1997). Word-length counting in Old Church Slavonic. *Journal of Quantitative Linguistics*, *4*(1-3), 252–256.
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.
Sigurd, B., Eeg-Olofsson, M., & Van Weijer, J. (2004). Word length, sentence length and frequency – Zipf revisited. *Studia linguistica*, *58*(1), 37–52.
Spearman, C. (2010). The proof and measurement of association between two things. *International Journal of Epidemiology*, *39*(5), 1137–1150.
Strauss, U., Grzybek, P., & Altmann, G. (2007). Word length and word frequency. In P. Grzybek (Ed.), *Contributions to the science of text and language: Word length studies and related issues* (p. 277–294). Springer.
Torre, I. G., Luque, B., Lacasa, L., Kello, C. T., & Hernández-Fernández, A. (2019). On the physical origin of linguistic laws and lognormality in speech. *Royal Society open science*, *6*(8), 191023.
Torre, I. G., Luque, B., Lacasa, L., Luque, J., & Hernández-Fernández, A. (2017). Emergence of linguistic laws in human voice. *Scientific reports*, *7*(1), 43862.
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.
Čebanow, S. G. (1947). On conformity of language structures within the Indoeuropean family to Poisson's law. *Comptes rendus de l'Academie de science de l'URSS*, *55*(2), 99–102.
Vranić, V., & Matković, V. (1965). Mathematical theory of the syllabic structure of Croato-Serbian. *Rad JAZU (odjel za matematičke, fizičke i tehničke nauke; 10)*, *331*(), 181–199.
Wei, A., Lu, Q., & Liu, H. (2021). Word length distribution in Zhuang language. *Journal of Quantitative Linguistics*, *28*(3), 195–222.
Youngblood, M. (2025). Language-like efficiency in whale communication. *Science Advances*, *11*(6), eads6014.
Zerzwadse, G., Tschikoidse, G., & Gatschetschiladse, T. (1962). Die Anwendung der mathematischen Theorie der Wortbildung auf die georgische Sprache. *Grundlagenstudien aus Kybernetik und Geisteswissenschaft*, *3*(), 110-118.
Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, Massachusetts: Harvard 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.