Direct and indirect evidence of compression of word lengths.

Zipf's law of abbreviation revisited

Sonia Petrini[1] (0000-0002-0514-6223), Antoni Casas-i-Muñoz[2] (0000-0001-5690-316X), Jordi Cluet-i-Martinell[2] (0000-0003-4188-6728), Mengxue Wang[2] (0000-0002-8262-9333), Christian Bentz[3] (0000-0001-6570-9326), Ramon Ferrer-i-Cancho[1*] (0000-0002-7820-923X)

[1] Quantitative, Mathematical and Computational Linguistics Research Group. Departament de Ciències de la Computació, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain.
[2] Universitat Politècnica de Catalunya (UPC), Barcelona School of Informatics, Barcelona, Catalonia, Spain.
[3] Department of Linguistics, University of Tübingen, Tübingen, Germany.
[*] Corresponding author's email: rferrericancho@cs.upc.edu

DOI: https://doi.org/10.53482/2023_54_407

Abstract

Zipf's law of abbreviation, the tendency of more frequent words to be shorter, is one of the most solid candidates for a linguistic universal, in the sense that it has the potential for being exceptionless or with a number of exceptions that is vanishingly small compared to the number of languages on Earth. Since Zipf's pioneering research, this law has been viewed as a manifestation of a universal principle of communication, i.e. the minimization of word lengths, to reduce the effort of communication. Here we revisit the concordance of written language with the law of abbreviation. Crucially, we provide wider evidence that the law holds also in speech (when word length is measured in time), in particular in 46 languages from 14 linguistic families. Agreement with the law of abbreviation provides indirect evidence of compression of languages via the theoretical argument that the law of abbreviation is a prediction of optimal coding. Motivated by the need of direct evidence of compression, we derive a simple formula for a random baseline indicating that word lengths are systematically below chance, across linguistic families and writing systems, and independently of the unit of measurement (length in characters or duration in time). Our work paves the way to measure and compare the degree of optimality of word lengths in languages.

Keywords: word length, compression, law of abbreviation

1 Introduction

It has been argued that linguistic universals are a myth (Evans and Levinson, 2009), but this neglects the statistical regularities that the quantitative linguistic community has been investigating for many decades. A salient case is Zipf's law of abbreviation, the tendency of more frequent words to be shorter

(Zipf, 1949). It holds across language families (Bentz and Ferrer-i-Cancho, 2016; Koplenig et al., 2022; Levshina, 2022; Meylan and Griffiths, 2021; Piantadosi et al., 2011), writing systems (Sanada, 2008; Wang and Chen, 2015) and modalities (Börstell et al., 2016; Hernández-Fernández and Torre, 2022; Torre et al., 2019), and also when word length in characters is replaced by word duration in time (Hernández-Fernández et al., 2019). Furthermore, the number of species where a parallel of this law has been confirmed in animal communication is growing over time (Semple et al., 2022).[1] In language sciences, research on the law of abbreviation in languages measures word length in discrete units (e.g., characters) whereas, in biology, research on the law in other species typically uses duration in time. Here, we aim to reduce the gulf that separates these two traditions by promoting research on the law of abbreviation on word durations.

G. K. Zipf believed that the law of abbreviation constituted *indirect* evidence of the minimization of the cost of using words (Zipf, 1949). At present, Zipf's view is supported by standard information theory and its extensions: the main argument is that the minimization of L, the mean word length, that is indeed a simplification of Zipf's cost function,[2] leads to the law of abbreviation (Ferrer-i-Cancho et al., 2019; Ferrer-i-Cancho et al., 2013). Using the terminology of information theory, the minimization of mean word length is known as compression. Using the terminology of quantitative linguistics, L is the average length of tokens from a repertoire of n types, that is defined as

$$L = \sum_{i=1}^{n} p_i l_i, \tag{1}$$

where p_i and l_i are, respectively, the probability and the length of the i-th type. In practical applications, L is calculated replacing p_i by the relative frequency of a type, that is

$$p_i = f_i/T,$$

where f_i is the absolute frequency of a type and T is the total number of tokens, i.e.

$$T = \sum_{i=1}^{n} f_i.$$

This leads to a definition of L that is

$$L = \frac{1}{T}\sum_{i=1}^{n} f_i l_i.$$

At present, the mathematical link between the law of abbreviation and compression has been established under the assumption that words are coded optimaly so as to minimize L. If words are coded optimaly, the correlation between the frequency of a word and its duration cannot be positive (Ferrer-i-Cancho

[1]The interested reader can check the latest discoveries on this law in "Bibliography on laws of language outside human language" at https://cqllab.upc.edu/biblio/laws/.

[2]He referred to the cost function as "minimum equation" (Zipf, 1949).

et al., 2019). Thus, a lack of correlation between the frequency of a word and its duration does not imply absence of compression. Furthermore, it is not a warranted assumption that languages code words optimaly. Therefore, an approach to find *direct* evidence of compression getting rid of the assumption of optimal coding is required.

As a first approach, one could compare the value of L of a language against L_{max}, the maximum value that L could achieve in this language. The larger the gap between L and L_{max}, the higher the level of compression in the language. However, the problem is that L_{max} can be infinite *a priori*. To fix that problem, one could restrict L_{max} to be finite but then this raises the question of what should be the finite value of L_{max} and why. For these reasons, here we resort to the notion of random baseline, that here is defined assuming some random mapping of word types into strings. In previous research, the random baseline was defined by the average word length resulting from a shuffling of the current length/duration of types so as to check if L was smaller than expected by chance in that random mapping (Ferrer-i-Cancho et al., 2013; Heesen et al., 2019). Critically, an exact method to compute the random baseline, namely the expected word length in these shufflings, is missing.

The remainder of the article is organized as follows. In Section 2, we introduce the definition of L_r, the random baseline, that we will use to explore direct evidence of compression. In particular, we derive a simple formula for L_r that will simplify future research on compression in natural communication systems. In Section 3 and Section 4, we present, respectively, the materials and methods that will be used to provide further evidence of compression and the law of abbreviation in real languages with emphasis on word durations. In Section 4, we present a new unsupervised method to exclude words with foreign characters in line with good practices for research on linguistic laws and communicative efficiency (Meylan and Griffiths, 2021). In Section 5, we show that the law of abbreviation holds without exceptions in a wide sample of languages, independently of the unit of measurement of word length, namely characters or duration in time, providing further indirect evidence of compression in languages. In addition, the random baseline indicates that word lengths are systematically below chance, across linguistic families and writing systems, independently of the unit of measurement (length in characters or duration in time), providing direct evidence of compression. Finally, in Section 6, we discuss the findings in relation to the potential universality of the law of abbreviation and the universality of compression in languages. We also make proposals for future research.

2 A random baseline revisited

In our statistical setting, the null hypothesis states that compression (minimization of word lengths) has no effect on word lengths. The alternative hypothesis states that compression has an effect on word lengths as Zipf hypothesized. If the null hypothesis is rejected then word lengths are shorter than expected by chance.

Table 1: Matrix indicating the frequency and length of three types. The mean type length is $L = \frac{235}{125} = 1.88$.

i	f_i	l_i
1	100	2
2	20	1
3	5	3

Consider a matrix with two columns, f_i and l_i, that are used to compute the average word length L. The matrix in Table 1 gives $L = \frac{235}{125} = 1.88$. We consider the null hypothesis of a random mapping of probabilities into lengths, namely that the ordering of the f_i's or the l_i's in Table 1 is arbitrary and results from a random shuffling of one of these variables or both. We use f_i', l_i' and p_i' for the new values of f_i, l_i and p_i that result from one of these shufflings.

This null hypothesis was introduced in research on compression in human language and animal communication to test if L is significantly small using a permutation test (Ferrer-i-Cancho et al., 2013; Heesen et al., 2019). Later, it was used to estimate the degree of optimality of word lengths (Moreno Fernández, 2021; Pimentel et al., 2021). Our new contribution here is a precise mathematical characterization of the null hypothesis and the derivation of a simple formula the expected word length.

In the context of computing average word length, the matrix in Table 1 is equivalent to a matrix where the column f_i is replaced by a column with p_i thanks to

$$p_i' = \frac{f_i'}{T}.$$

Indeed, the null hypothesis has three variants

1. Single column shuffling. Only the column of f_i or p_i is shuffled.
2. Single column shuffling. Only the column of l_i is shuffled.
3. Dual column shuffling. The column of f_i or p_i and the column of l_i are both shuffled.

In each of the variants, all random shufflings of a specific column are equally likely. In case of dual shuffling, the shuffling of one column is independent of the shuffling of the other column. The outcome of a dual shuffling on Table 1 is shown in Table 2.

The random baseline, L_r, is the expected value of L under the null hypothesis.[3] L_r can be defined in more detail in two main equivalent ways:

[3]Notice that L is indeed the expected value of the length of a token but under a distinct setting (a distinct null hypothesis), where one picks a token uniformly at random over all tokens of a text and looks at its length.

Table 2: Matrix indicating the frequency and length of three types. The mean type length is $L = \frac{345}{125} = 2.76$.

i	f'_i	l'_i
1	20	2
2	100	3
3	5	1

1. The value of L that is expected if L is recomputed after pairing the f_i's and the l_i's at random and recomputing L. The new value of L depends on the variant of the null hypothesis. When shuffling the column for f_i in the matrix (Table 1), the new L is
$$L' = \frac{1}{T}\sum_{i=1}^{n} f'_i l_i.$$
When shuffling the column for l_i and recomputing L, the new L is
$$L' = \frac{1}{T}\sum_{i=1}^{n} f_i l'_i.$$
When shuffling both columns, the new L is
$$L' = \frac{1}{T}\sum_{i=1}^{n} f'_i l'_i.$$
2. The average value of L that is expected over all possible shufflings in one of the variants of the null hypothesis. In the example in Table 3, on shuffling only the l_i column,
$$L_r = \frac{\frac{155}{125} + \frac{170}{125} + \frac{235}{125} + \frac{265}{125} + \frac{330}{125} + \frac{345}{125}}{6} = \frac{155 + 170 + 235 + 265 + 330 + 345}{125 \cdot 6} = 2.$$

We use $\mathbb{E}[X]$ to refer to the expected value of a random variable X under some variant of the null hypothesis above. Then
$$L_r = \mathbb{E}[L'],$$
where L' is the value of L resulting from some shuffling.

In quantitative linguistics, the mean length of tokens (L) is also known as dynamic word length (Chen et al., 2015) and corresponds to the mean length of the words in a text. The mean length of types (M), defined as
$$M = \frac{1}{n}\sum_{i=1}^{n} l_i,$$
is also known as the static word length and corresponds to average length of the headwords in a dictionary (Chen et al., 2015). Interestingly, the following property states that L_r turns out to be M independently of the variant of the null hypothesis under consideration.

Property 2.1. *The expected value of L' under any variant of the null hypothesis is $L_r = M$.*

Proof. We analyze $\mathbb{E}[L']$ under each of the variants of the null hypothesis.

Dual shuffling. Applying the linearity of expectation and independence between the shuffling of the p_i column of the that of the l_i column, we obtain

$$\begin{aligned}\mathbb{E}[L'_1] &= \mathbb{E}\left[\sum_{i=1}^{n} p'_i l'_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[p'_i l'_i] \\ &= \sum_{i=1}^{n} \mathbb{E}[p'_i]\,\mathbb{E}[l'_i].\end{aligned}$$

Noting that

$$\begin{aligned}\mathbb{E}[p'_i] &= \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{n} \\ \mathbb{E}[l'_i] &= \frac{1}{n}\sum_{i=1}^{m} l_i = M,\end{aligned}$$

we finally obtain

$$\mathbb{E}[L'] = \sum_{i=1}^{n} \frac{M}{n} = M. \tag{2}$$

Single shuffling of the l_i column. Applying the linearity of expectation and the fact that the column of p_i remains constant, we obtain

$$\begin{aligned}\mathbb{E}[L'_1] &= \mathbb{E}\left[\sum_{i=1}^{n} p_i l'_i\right] \\ &= \sum_{i=1}^{n} p_i\,\mathbb{E}[l'_i].\end{aligned}$$

Recalling $\mathbb{E}[l'_i] = M$, we finally obtain

$$\mathbb{E}[L'] = M\sum_{i=1}^{n} p_i = M. \tag{3}$$

Single shuffling of the p_i column. Applying the linearity of expectation and the fact that the column of l_i remains constant, we obtain

$$\begin{aligned}\mathbb{E}[L'_1] &= \mathbb{E}\left[\sum_{i=1}^{n} p'_i l_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[p'_i] l_i.\end{aligned}$$

Recalling $\mathbb{E}[p'_i] = \frac{1}{n}$, we finally obtain

$$\mathbb{E}[L'] = \frac{1}{n}\sum_{i=1}^{n} l_i = M. \tag{4}$$

□

Table 3: All the $3! = 6$ permutations of the column l_i in Table 1 that can be produced. Each permutation is indicated with letters from A to F. L', the mean length of types in a shuffling, is shown at the bottom for each permutation.

		A	B	C	D	E	F
i	f_i	l'_i	l'_i	l'_i	l'_i	l'_i	l'_i
1	100	1	1	2	2	3	3
2	20	2	3	1	3	1	2
3	5	3	2	3	1	2	1
	L'	$\frac{155}{125} = 1.24$	$\frac{170}{125} = 1.36$	$\frac{235}{125} = 1.88$	$\frac{265}{125} = 2.12$	$\frac{330}{125} = 2.64$	$\frac{345}{125} = 2.76$

The previous finding indicates that the random baseline for L is equivalent to assuming that all word types are equally likely, namely, replacing each p_i by $1/n$.

3 Material

3.1 General information about corpora and languages

We investigate the relationship between the frequency of a word and its length in languages from two collections: Common Voice Forced Alignments (Section 3.2.1), hereafter CV, and Parallel Universal Dependencies (Section 3.2.2), hereafter PUD.

All the preprocessed files used to produce the results from the original collections are available in the repository of the article.[4]

PUD comprises 20 distinct languages from 7 linguistic families and 8 scripts (Table 4). CV comprises 46 languages from 14 linguistic families (we include 'Conlang', i.e. 'constructed languages', as a family for Esperanto and Interlingua) and 10 scripts (Table 5). Both PUD and CV are biased towards the Indo-European family and the Latin script. The typological information (language family) is obtained from Glottolog 4.6[5]. The writing systems are determined according to ISO-15924 codes[6]. In Table 4 and Table 5, we show the scripts using their standard English names. For example, most languages from the Indo-European family are written in Latin scripts. We also categorize Chinese Pinyin and Japanese Romaji as Latin scripts.

[4]In the *data* folder of https://github.com/IQL-course/IQL-Research-Project-21-22.
[5]https://glottolog.org/
[6]https://unicode.org/iso15924/iso15924-codes.html

Table 4: Summary of the main characteristics of the languages in the PUD collection. For each language, we show the linguistic family, the writing system (namely script name according to ISO-15924) and various numeric parameters: A, the observed alphabet size (number of distinct characters), n, the number of word types, and T, the number of word tokens.

Language	Family	Script	A	n	T
Arabic	Afro-Asiatic	Arabic	20	3309	11667
Indonesian	Austronesian	Latin	23	4501	16702
Russian	Indo-European	Cyrillic	23	4666	11749
Hindi	Indo-European	Devanagari	44	4343	20071
Czech	Indo-European	Latin	33	7073	15331
English	Indo-European	Latin	25	5001	18028
French	Indo-European	Latin	26	5214	20407
German	Indo-European	Latin	28	6116	18331
Icelandic	Indo-European	Latin	32	6035	16209
Italian	Indo-European	Latin	24	5606	21266
Polish	Indo-European	Latin	31	7188	15191
Portuguese	Indo-European	Latin	38	5661	21855
Spanish	Indo-European	Latin	32	5750	21067
Swedish	Indo-European	Latin	25	5624	16378
Japanese	Japonic	Japanese	1549	4852	24737
Japanese-strokes	Japonic	Japanese	1549	4852	24737
Japanese-romaji	Japonic	Latin	24	4849	24734
Korean	Koreanic	Hangul	379	6218	12307
Thai	Kra-Dai	Thai	50	3573	20860
Chinese	Sino-Tibetan	Han (Traditional variant)	2038	4970	17845
Chinese-strokes	Sino-Tibetan	Han (Traditional variant)	2038	4970	17845
Chinese-pinyin	Sino-Tibetan	Latin	50	4970	17845
Turkish	Turkic	Latin	28	6587	13799
Finnish	Uralic	Latin	24	6938	12701

Table 5: Summary of the main characteristics of the languages in the CV collection. For every language we show its linguistic family, the writing system (namely script name according to ISO-15924) and various numeric parameters: A, the observed alphabet size (number of distinct characters), n, the number of word types, and, T, the number of word tokens. 'Conlang' stands for 'constructed language', that is an artificially created language. This is not a family in the proper sense as Conlang languages are not related in the common linguistic family sense.

Language	Family	Script	A	n	T
Arabic	Afro-Asiatic	Arabic	31	6397	45825
Maltese	Afro-Asiatic	Latin	31	8058	44112
Vietnamese	Austroasiatic	Latin	41	370	938
Indonesian	Austronesian	Latin	22	3768	44210
Esperanto	Conlang	Latin	27	27759	406261
Interlingua	Conlang	Latin	20	5126	30504
Tamil	Dravidian	Tamil	29	1210	6439
Persian	Indo-European	Arabic	38	13115	1662508
Assamese	Indo-European	Assamese	43	971	1813
Russian	Indo-European	Cyrillic	32	31827	637686
Ukrainian	Indo-European	Cyrillic	34	14337	120760
Panjabi	Indo-European	Devanagari	37	84	98
Modern Greek	Indo-European	Greek	33	5813	37880
Breton	Indo-European	Latin	28	4228	38237
Catalan	Indo-European	Latin	39	79112	3294206
Czech	Indo-European	Latin	33	15518	147582
Dutch	Indo-European	Latin	23	10225	316498
English	Indo-European	Latin	28	173023	9828713
French	Indo-European	Latin	49	160243	3729370
German	Indo-European	Latin	30	148436	4230565
Irish	Indo-European	Latin	23	2251	22593
Italian	Indo-European	Latin	34	54996	811783
Latvian	Indo-European	Latin	27	7251	29456
Polish	Indo-European	Latin	32	25340	595411
Portuguese	Indo-European	Latin	27	11509	283048
Romanian	Indo-European	Latin	29	6423	33341
Romansh	Indo-European	Latin	26	9614	43792
Slovenian	Indo-European	Latin	24	5937	26304
Spanish	Indo-European	Latin	33	75010	1842474
Swedish	Indo-European	Latin	25	4371	62951
Welsh	Indo-European	Latin	22	11143	539621
Western Frisian	Indo-European	Latin	30	8383	63073
Oriya	Indo-European	Odia	41	764	1700
Dhivehi	Indo-European	Thaana	27	111	1284
Georgian	Kartvelian	Georgian	25	6505	12958
Basque	Language isolate	Latin	21	24748	458071
Mongolian	Mongolic	Mongolian	31	14608	70217
Kinyarwanda	Niger-Congo	Latin	26	133815	1939810
Abkhazian	Northwest Caucasian	Cyrillic	28	119	156
Hakha Chin	Sino-Tibetan	Latin	23	2499	17776
Chuvash	Turkic	Cyrillic	22	4311	13583
Kirghiz	Turkic	Cyrillic	30	10130	61844
Tatar	Turkic	Cyrillic	34	21823	144356
Yakut	Turkic	Cyrillic	28	7904	22577
Turkish	Turkic	Latin	31	8926	107686
Estonian	Uralic	Latin	23	28691	121549

3.2 The datasets

We measure word length in two main ways: *duration in time* and *length in characters*. Concerning Chinese and Japanese, we additionally consider the number of strokes and the number of characters of their romanization (i.e. Pinyin for Chinese and Romaji for Japanese).

Given these datasets, word durations are obtained only from CV. Word lengths in characters are obtained from both CV as well as from PUD. Word lengths in strokes, and word lengths in characters after romanization, are obtained only from PUD.

3.2.1 Common Voice Forced Alignments

The Common Voice Corpus[7] is an open source dataset of recorded voices uttering sentences in many different languages. The amount of data, as well as the source and topic of each sentence, depends considerably on the language and the corpus version. Specifically, the Common Voice Corpus 5.1 contains information on 54 languages and dialects.

Common Voice Forced Alignments (CVFA)[8] were created by Josh Meyer using the Montreal Forced Aligner[9] on top of the Common Voice Corpus 5.1. Kabyle, Upper Sorbian and Votic were left out of the alignments for an undocumented reason. Therefore, CVFA contains information on 51 languages.

In our analyses, Japanese and the three Chinese dialects were excluded as the forced aligner failed to correctly extract words from sentences. In addition, both Romansh dialects were fused into a single Romansh language. Indeed, given the nature of this corpus, all languages are likely to be represented by more than one dialect.

Notice that Abkhazian, Panjabi, and Vietnamese have a critically low number of tokens ($T < 1000$ in Table 5). However, we decided to include them in the analyses so as to understand their limitations related to corpus size.

3.2.2 Parallel Universal Dependencies

The Universal Dependencies (UD)[10] collection is an open source dataset of annotated sentences, in which the amount of data depends on each language. The Parallel Universal Dependencies (PUD) collection is a parallel subset of 20 languages from the UD collection, consisting of 1000 sentences. It allows for a cross-language comparison, controlling for content and annotation style.

In Table 4, we show the characteristics of the languages in PUD. For traditional Chinese and Japanese,

[7]https://commonvoice.mozilla.org/en/datasets
[8]https://github.com/JRMeyer/common-voice-forced-alignments
[9]https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner
[10]https://universaldependencies.org/

we also include word lengths in romanizations (Pinyin and Romaji respectively), as well as word lengths measured in strokes, resulting in a total of 24 language files. Notice that three Japanese words that are hapax legomena could not be romanized and thus the number of tokens and types varies slightly with respect to the original Japanese characters (Table 4).

4 Methodology

All the code used to produce the results is available in the repository of the article.[11]

4.1 The units of length

4.1.1 Duration

The duration of a word for a given language is estimated by computing the median duration in seconds across all its occurrences in utterances in the CV corpus. All words with equal orthographic form are assumed to be the same type. The median is preferred over the mean as it is less sensitive to outliers (that may be produced by forced alignment errors) and better suited to deal with heavy-tailed distributions (Hernández-Fernández et al., 2019). Given the oral nature of the data, we do expect to observe some variation in the duration of words, due to differences between individuals, and variation within a single individual. This is more generally in line with speakers acting as complex dynamical systems (Kello et al., 2010). For these reasons, median duration is preferred for research on the law of abbreviation in acoustic units (Torre et al., 2019; Watson et al., 2020).

4.1.2 Length in characters

Word length in characters is measured by counting every Unicode UTF-8 character present in a word. Special characters such as "=" were removed. Characters with stress accents are considered as different from their non-stressed counterpart (e.g. "a" and "à" are considered separate characters). Following best practices from (Meylan and Griffiths, 2021), characters were always kept in UTF-8.

4.1.3 Length in strokes

Japanese Kanji and Chinese Hanzi were turned into strokes using the *cihai* Python library.[12] In Japanese characters other than Kanji, namely Japanese Kana, the number of strokes in printed versus hand-written modality can differ (Chinese Hanzi and Japanese Kanji have the same number of strokes in printed version or hand-written version). Here we counted the number of strokes in printed form. Japanese Kana were converted into printed strokes by using a hand-crafted correspondence table, since Kana is not part

[11] In the *code* folder of https://github.com/IQL-course/IQL-Research-Project-21-22.
[12] https://github.com/cihai/cihai

of the CJK unified character system. This table was created by us and checked by a native linguist (S. Komori from Chubu University, Japan). It is available in the repository of the article.[13]

In case of discrepancies on the number of strokes for a given character, the most typical printed version was chosen.

4.1.4 Length in Pinyin and Romaji

Chinese Pinyin was obtained using the *cihai* package as above, while the Japanese Romaji was obtained with the *cutlet* Python library.[14] The latter uses Kunrei-shiki romanization (since it is the one used officially by the government of Japan) and the spelling of foreign words is obtained in its native reading (e.g. "カレー" is romanized as "karee" instead of "curry"). There are some particularities with the romanization of Kanji characters by *cutlet*. For example, in the case of the word "year" (年), it chose the reading of "Nen" instead of "Tosi", which would be the expected one.

A more systematic issue with Japanese romanization is that it does not provide means to indicate pitch accents, which are implicitly present in Kanji. For example, "日本" "Ni↑hon" ("Japan") is romanized as simply "Nihon". Therefore, the alphabet size of romanized Japanese is smaller than it should be, compared to other languages where, as stated before, stress accents are counted as distinctive features of characters.

4.2 Tokenization

Tokenization is already given in each dataset and we borrow it for our analyses. Thus tokenization methods are not uniform for CV and PUD and are not guaranteed to be uniform among languages even within each of these datasets.

4.3 Filtering of tokens

Examining our datasets, we noticed that in some text files there was a considerable number of unusual character strings, as well as foreign words (written in different scripts). These need to be filtered out in order to obtain a "clean" set of word types. To this end we filter out tokens following a two step procedure:

1. *Mandatory elementary filtering*. This filter consists of:
    - *Common filtering*. In essence, it consists of the original tokenization and the removal of tokens containing digits. In each collection, the original tokenizer yields tokens that may contain certain punctuation marks. Due to the nature of the CV dataset, the bulk of punctuation was

[13]In the *data/other* folder of https://github.com/IQL-course/IQL-Research-Project-21-22.
[14]https://github.com/polm/cutlet

   already removed via the Montreal Forced Aligner with some exceptions. For instance, single quotation (in particular “’”) is a punctuation sign that is kept within a word token in CV, as it is necessary for the formation of clitics in multiple languages, such as in English or French. In PUD, as a part of UD, contractions are split into two word types. “can’t” is split into “ca” “n’t” (in CV “can’t” would remain as just one token). In both collections, words containing ASCII digits are removed because they do not reflect phonemic length and can be seen as another writing system.

   - *Specific filtering*. In case of the PUD collection, we excluded all tokens with Part-of-Speech (POS) tag ‘PUNCT’. In case of the CV collection, we removed tokens tagged as <unk> or null tokens, namely tokens that either could not be read or that represent pauses.
   - *Lowercasing*. Every character is lowercased. In the case of CV, this is already given by the Montreal Forced Aligner, while in the case of PUD, tokens are lowercased by means of the *spaCy* Python package.[15]

2. *Optional filtering*. This is a new method that is applied after the previous filter and described in Section 4.4.

4.4 A new method to filter out unusual characters

It has been pointed out that “chunk” words and loanwords can distort the results of quantitative analyses of word lengths (Meylan and Griffiths, 2021). Indeed, especially the files of the Common Voice Corpus feature a considerable number of word tokens which do not consist of characters belonging to the primary alphabet of the respective writing system. Meylan and Griffiths (2021) proposed to use dictionaries to exclude such anomalous words. However, this is not feasible for our multilingual datasets, as loanword dictionaries are not available for this large number of diverse languages (Table 4 and Table 5). The Intercontinental Dictionary Series,[16] for example, contains only around half of the languages in our analysis, so it is not applicable to many of them. Hence, this approach would lead to a non-uniform treatment of different languages and texts. Selecting a matched set of semantic concepts across languages using a lexical database is also infeasible due to similar reasons.

Against this backdrop, we decided to develop an unsupervised method to filter out words which contain highly unusual characters. For a given language, the method starts by assuming that the strings (after the mandatory filtering illustrated above) contain characters of two types: characters of the working/primary alphabet as well as other characters. We hypothesize that the latter are much less frequent than the former.

[15] https://spacy.io/
[16] https://ids.clld.org/

Following this rationale, we apply the k-means algorithm of the *Ckmeans R* package[17] to split the set of characters into the two groups based on the logarithm of the frequency of the characters.[18] To maximize the power of the clustering method, we use the exact method with $k = 2$ for one dimension instead of the customary approximate method. We then keep the high frequency cluster as the real working alphabet and filter out the word tokens that contain characters not belonging to this high frequency cluster.

We illustrate the power of the method by showing working alphabets that are obtained on CV, that is the noisiest one of the collections.

In English, the working alphabet is defined by the 26 English letters and quotation marks (“’”, “’”). These quotation marks are used often in clitics, and as such are correctly identified as part of the encoding, since, for example, “can’t” and “cant” are different words in meaning, with “can’t” meaning “can not”, while “cant” is a statement on a religious or moral subject that is not believed by the person making the statement, with the differentiating feature being the “’”. Therefore, the working alphabet becomes 5 vowels (“a”, “e”, “i”, “o”, “u”), 21 consonants (“b”, “c”, “d”, “f”, “g”, “h”, “j”, “k”, “l”, “m”, “n”, “p”, “q”, “r”, “s”, “t”, “v”, “w”, “x”, “y”, “z”) and 2 kinds of quotation marks (“’”, “’”).

In Russian, the working alphabet comprises 9 vowels (“а”, “о”, “у”, “ы”, “э”, “я”, “ю”, “и”, “е”), a semivowel / consonant “й”, 20 consonants (“б”, “в”, “г”, “д”, “ж”, “з”, “к”, “л”, “м”, “н”, “п”, “р”, “с”, “т”, “ф”, “х”, “ц”, “ч”, “ш”, “щ”) and 2 modifier letters (“ъ”, “ь”).

In Italian, it comprises 5 vowels (“a”, “e”, “i”, “o”, “u”), 21 consonants (“b”, “c”, “d”, “f”, “g”, “h”, “j”, “k”, “l”, “m”, “n”, “p”, “q”, “r”, “s”, “t”, “v”, “w”, “x” , “y”, “z”) and 6 instances of the 5 vowels containing a diacritic mark (“à”, “è”, “é”, “ì”, “ò”, “ù”).

The unsupervised filter method filter is not applied to Chinese, Japanese and Korean as, given their nature, this would exclude letters that actually belong to the real alphabet. In Section B.1 we analyze the impact of the optional filter and provide arguments for not applying the unsupervised filter to these languages. As a compensation, strings that contain non-CJK characters are filtered out in Chinese and Japanese as a part of the optional filter. In Korean, only a few characters are not proper Hangul and thus such a complementary filtering is not necessary.

[17] https://cran.r-project.org/web/packages/Ckmeans.1d.dp/index.html

[18] The motivation for taking logarithms of frequencies is three-fold: First, this brings observations closer together. Note that the k-means algorithm prefers high-density areas. Second, this transforms the frequencies into a measure of surprisal, following standard information theory (Shannon, 1948). Third, manual inspection suggests that the logarithmic transformation is required to produce an accurate split.

4.5 Immediate constituents in writing systems

When measuring word length in written languages, we are using *immediate constituents* of written words. In Romance languages, the immediate constituents are letters of the alphabet, which are a proxy for phonemes. For syllabic writing systems (as Chinese in our dataset), these are characters that correspond to syllables. In addition, for Chinese and Japanese, we are considering two other possible units for word length, which are not immediate constituents, but alternative ways of measuring word lengths which could provide useful insights: strokes and letters in Latin script romanizations. That means that for each of these languages words are unfolded into three systems, one for each unit of encoding (original characters, strokes, romanized letters/characters). In the hierarchy from words to other units, only the original characters are immediate constituents.

4.6 Statistical testing

4.6.1 Correlation

When measuring the association between two variables, we use both Pearson correlation and Kendall correlation (Conover, 1999). Note that the traditional view of Pearson correlation as a measure of linear association and thus not suitable for non-linear association has been challenged (van den Heuvel and Zhan, 2022).

4.6.2 How to test for the law of abbreviation

We used a left-sided correlation test to verify the presence of the law of abbreviation. In a purely exploratory or atheoretic exploration, one should use a two-sided test. In an exploration guided by theory, namely regarding the law of abbreviation as a manifestation of compression, the test should be left-sided as theory predicts that $\tau(p, l)$ cannot be positive in case of optimal coding (Ferrer-i-Cancho et al., 2019).

4.6.3 How to test for compression

In the context of the null hypothesis of a random mapping of type probabilities into type lengths, testing that compression (minimization of L) has some effect on actual word lengths is easy because L is a linear function of r, the Pearson correlation between word length and word probability (Appendix A). In particular,

$$L = ar + L_r,$$

where $a = (n-1)s_p s_l$, being n the number of types and s_p and s_l, respectively,the standard deviation of type probabilities and type lengths. In such random mappings, L_r, s_p and s_l remain constant and then testing if r is significantly small is equivalent to testing if L is significantly small (notice $a \geq 0$).

4.6.4 Controlling for multiple testing

When performing multiple correlation tests at the same time, it becomes easier to reject the null hypothesis simply by chance. To address this problem we used a Holm-Bonferroni correction to p-values.[19] We applied the correction when checking the law of abbreviation in the languages of a collection, so as to exclude the possibility that the law of abbreviation is found many times simply because we are testing it in many languages.

5 Results

In Section 1, we highlighted the importance of distinguishing between direct and indirect evidence of compression. Against this theoretical backdrop, here we first investigate the presence of Zipf's law of abbreviation in languages. Then we investigate direct evidence of compression with the help of the new random baseline.

5.1 The law of abbreviation revisited

We investigate the presence of the law of abbreviation by means of left-sided correlation tests for the association between frequency and length. We use both Kendall correlation, as suggested by theory on the origins of the law (Ferrer-i-Cancho et al., 2019), and Pearson's. For each language, we show the significance level of the relationship, color-coded by the value of the correlation coefficient. Figure 1 (a,b) indicates that the law holds in all languages – regardless of the definition of word length – when Kendall τ correlation is used. In both collections, we find Kendall τ correlation coefficients significant at the 99% confidence level, except for Dhivehi in the CV collection when length is measured in characters, and Abkhazian, Dhivehi, Panjabi and Vietnamese when length is measured in duration. However, note that these are all still significant at the 95% confidence level. When Pearson correlation is used instead, Figure 1 (c) shows that the picture remains the same in PUD. The main findings are the same also in CV (Figure 1 (d)), but when length is measured in duration Panjabi ceases to be significant at the 95% confidence level. Overall, we only fail to find the law of abbreviation in Panjabi given word durations, and using Pearson correlation. This is most probably related to undersampling, as this particular language only features 98 tokens (Table 5).

[19] https://stat.ethz.ch/R-manual/R-devel/library/stats/html/p.adjust.html

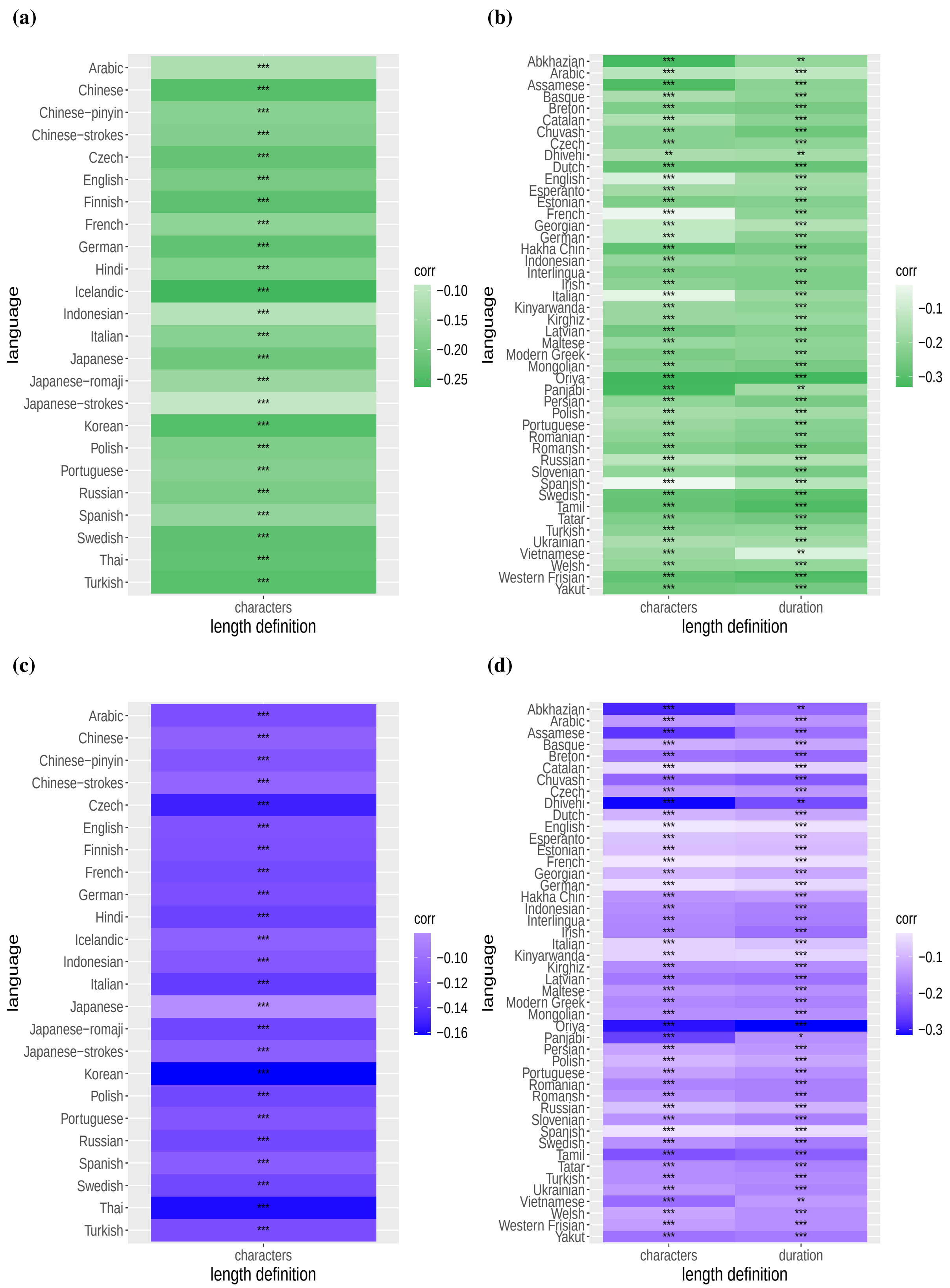

Figure 1: The correlation between frequency and length across languages. '***' indicates a Holm-Bonferroni corrected p-value lower than or equal to 0.01, '**' indicates lower than or equal to 0.05 but smaller than 0.1 and '*' indicates lower than or equal to 0.1. Here '*' symbols are not used to indicate significance but p-value ranges. (a) Kendall τ correlation in PUD (word length in characters). (b) Kendall τ correlation in CV (left: word length in characters; right: word length in duration). (c) Same as (a) with Pearson r correlation. (d) Same as (b) with Pearson r correlation.

5.2 Real word lengths versus the random baseline

We investigate the relationship between the actual mean word length (L) and the random baseline (L_r). We find that $L < L_r$ for all languages in every collection (Figure 2 and Tables B3, B4, B5). Interestingly, there is a large gap between L and L_r in the majority of languages, which is more compelling in CV with word durations (Figure 2). Exceptions to the large gap – as in the case of Panjabi and Abkhazian when length is measured in duration – mainly concern languages with reduced sample sizes. The result holds even when alternative units of measurement are considered for Chinese and Japanese.

Figure 2 is reminiscent of Figure 4 of Pimentel et al. (2021) but our setting is much simpler (it only involves L and L_r).

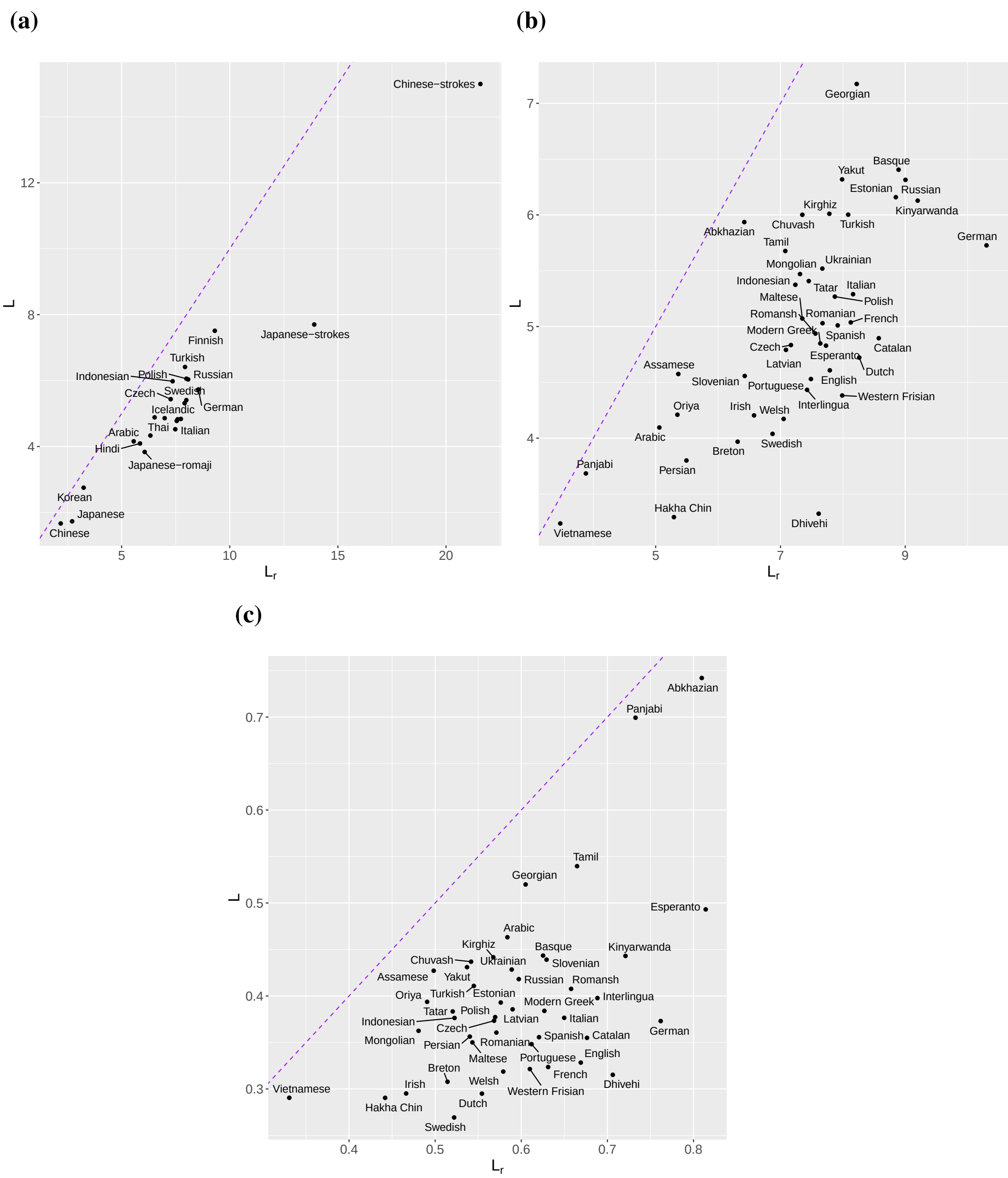

Figure 2: Mean word length (L) as a function of the random baseline (L_r) in languages. Every point stands for a language. The diagonal (long dashed line) indicates the line $L = L_r$. Languages with $L < L_r$ are located below the diagonal. (a) Languages in PUD with word length measured in characters (or strokes for Chinese and Japanese). (b) Languages in CV with word length measured in characters. (c) Languages in CV with word length measured in duration (seconds).

5.3 Impact of disabling the filter of words that contain "foreign" characters

All results presented in this section have been obtained after applying the new method to filter out highly unusual characters and words described in Section 4.4. If the filter is disabled, we obtain some slight changes in the values, but the qualitative results summarized above remain the same.

6 Discussion

6.1 The universality of Zipf's law

The first step of our analysis consisted in checking the universality of the law of abbreviation in the languages of our samples through a Kendall τ correlation test. Here, we introduced two methodological improvements with respect to previous research: using the Bonferroni-Holm correction for p-values, as well as word length in time given spoken utterances, rather than just characters in written form (Bentz and Ferrer-i-Cancho, 2016). We also computed Pearson correlations for two reasons: (a) to verify the robustness of the conclusions and (b) to check the significance of the gap between L and L_r (the case of (b) is addressed in the next subsection). We find that the law of abbreviation holds in nearly all languages in our sample at a 95% confidence level, independently from how word length is measured, and even after controlling for multiple testing. The only exception is Panjabi in CV, but only when length is measured in duration and Pearson r correlation is used. Panjabi is also the language suffering most from under-sampling (only 98 tokens). Therefore, Panjabi cannot be considered a true exception to the law of abbreviation.

Given the rather scarce evidence of the law of abbreviation in word durations in human language (Torre et al., 2019), we have taken step forward by providing evidence of it in 46 languages from 14 linguistic families. The massive agreement of the law of abbreviation even when orthographic word lengths are replaced by word durations in human languages provides stronger support for the law of abbreviation as a potentially universal pattern of human languages with respect to previous research relying on word length in characters (Bentz and Ferrer-i-Cancho, 2016) and often on a small number of linguistic families (Koplenig et al., 2022; Levshina, 2022; Meylan and Griffiths, 2021; Piantadosi et al., 2011).

6.2 Direct evidence of compression

We have found that word lengths are shorter than expected by chance ($L < L_r$) in all languages in every collection (Figure 2). Such a systematic finding is unlikely to be accidental and strongly indicates that compression is acting in all languages in our sample. Crucially, the finding holds independently of how word length is measured. The ample evidence of compression even when orthographic word lengths are replaced by word durations in human languages provides stronger support for compression as a universal principle of the organization of languages with respect to previous research relying on word length in characters (Ferrer-i-Cancho et al., 2013).

It could be argued that these findings constitute evidence of compression in ensembles of language but not in individual languages. The reason is that $L < L_r$ does not imply that the difference between the actual word length and the random baseline is statistically significant for a single language. However, we have shown that the Pearson correlation is indeed a linear function of L and L_r (Appendix A) and thus L is significantly small in every language where the law of abbreviation has been confirmed using a Pearson correlation test.

Finally, the direct correspondence we have established between the average length of types (M) and the random baseline sheds new light on previous research. For instance, it has been shown that $M < L$ in Chinese characters in six time periods spanning two millennia (Chen et al., 2015, Fig. 4), which now can be reinterpreted as a sign of compression of word lengths in Chinese in light of our theoretical findings.

Future research

In this article, we have introduced a new random baseline and unveiled a systematic gap between that random baseline and real mean word lengths that we have interpreted as direct evidence of compression. Figure 2 suggests that the gap is wider when word lengths are measured in duration rather than in characters. However, we have not quantified the magnitude of that gap and we have neither taken into consideration the gap between actual mean words lengths and the minimum baseline, that would be defined as the minimum word length that could be achieved under certain constraints (Cover and Thomas, 2006; Ferrer-i-Cancho et al., 2019; Pimentel et al., 2021). Future research should quantify the first gap in relation to the minimum baseline. As the random baseline is crucial to asses the degree of optimality of word lengths, we have paved the way for exploring the degree of optimality of word lengths in characters or duration in languages.

Authors' contributions

SP: Conceptualization, Formal Analysis, Investigation, Software, Supervision, Validation, Visualization, Writing-original draft, Writing-review & editing; ACM: Data curation, Resources, Software; JCM: Writing-review & editing, Resources; MW: Writing-review & editing; CB: Conceptualization, Writing-review & editing; RFC: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project Administration, Supervision, Writing-original draft, Writing-review & editing.

Acknowledgments

This article is one of the products of the research project of the 1st edition of the master degree course "Introduction to Quantitative Linguistics" at Universitat Politècnica de Catalunya. We are specially grateful to two students of that course: L. Alemany-Puig for helpful discussions and computational support, and M. Michaux for comments on early versions of the manuscript. We thank M. Farrús and A. Hernández-Fernández for advice on voice datasets. We also thank S. Komori for advice on Japanese, Y. M. Oh for advice on Korean and S. Semple for helping us to improve English. Last but not least, we are very grateful to an anonymous reviewer for very helpful comments.

SP is funded by the grant 'Thesis abroad 2021/2022' from the University of Milan. CB was partly funded by the *Deutsche Forschungsgemeinschaft* (FOR 2237: Words, Bones, Genes, Tools - Tracking Linguistic, Cultural and Biological Trajectories of the Human Past), and the *Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung* (Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora, 176305). RFC is supported by a recognition 2021SGR-Cat (01266 LQMC) from AGAUR (Generalitat de Catalunya).

References

Balasubrahmanyan, V. K., Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, *3*(3), 177–228. https://doi.org/10.1080/09296179608599629

Bentz, C., Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In C. Bentz, G. Jäger, I. Yanovich (Eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen. http://hdl.handle.net/10900/68639

Börstell, C., Hörberg, T., Östling, R. (2016). Distribution and duration of signs and parts of speech in Swedish Sign Language. *Sign Language & Linguistics*, *19*, 143–196. https://doi.org/10.1075/sll.19.2.01bor

Chen, H., Liang, J., Liu, H. (2015). How does word length evolve in written Chinese? *PLOS ONE*, *10*(9), 1–12. https://doi.org/10.1371/journal.pone.0138567

Conover, W. J. (1999). *Practical nonparametric statistics* [3rd edition]. Wiley.

Cover, T. M., Thomas, J. A. (2006). *Elements of information theory* [2nd edition]. Wiley.

Deng, W., Allahverdyan, A. E., Li, B., Wang, Q. A. (2014). Rank–frequency relation for Chinese characters. *The European Physical Journal B*, *87*(2), 47. https://doi.org/10.1140/epjb/e2014-40805-2

Evans, N., Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*, 429–492. https://doi.org/10.1017/s0140525x0999094x

Ferrer-i-Cancho, R., Bentz, C., Seguin, C. (2019). Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, *29*(2), 165–194. https://doi.org/10.1080/09296174.2020.1778387

Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, *37*(8), 1565–1578. https://doi.org/10.1111/cogs.12061

Heesen, R., Hobaiter, C., Ferrer-i-Cancho, R., Semple, S. (2019). Linguistic laws in chimpanzee gestural communication. *Proceedings of the Royal Society B: Biological Sciences*, *286*, 20182900. https://doi.org/10.12775/3991-1.039

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). https://doi.org/10.3390/e21121153

Hernández-Fernández, A., Torre, I. G. (2022). Compression principle and Zipf's Law of brevity in infochemical communication. *Biology Letters*, *18*, 20220162. https://doi.org/10.1098/rsbl.2022.0162

Kello, C. T., Brown, G. D. A., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., Orden, G. C. V. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*(5), 223–232. https://doi.org/10.1016/j.tics.2010.02.005

Koplenig, A., Kupietz, M., Wolfer, S. (2022). Testing the relationship between word length, frequency, and predictability based on the German reference corpus. *Cognitive Science*, *46*(6), e13090. https://doi.org/10.1111/cogs.13090

Levshina, N. (2022). Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy*, *24*(2). https://doi.org/10.3390/e24020280

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. https://doi.org/10.1111/cogs.12983

Moreno Fernández, J. (2021). The optimality of word lengths (Master's thesis). Barcelona School of Informatics. Barcelona. http://hdl.handle.net/2117/361054

Naranan, S., Balasubrahmanyan, V. K. (1993). Information theoretic model for frequency distribution of words and speech sounds (phonemes) in language. *Journal of Scientific and Industrial Research*, *52*, 728–738.

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. https://doi.org/10.1073/pnas.1012551108

Pimentel, T., Nikkarinen, I., Mahowald, K., Cotterell, R., Blasi, D. (2021). How (non-)optimal is the lexicon? *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2021.naacl-main.350

Sanada, H. (2008). *Investigations in japanese historical lexicology*. Peust & Gutschmidt Verlag.

Semple, S., Ferrer-i-Cancho, R., Gustison, M. (2022). Linguistic laws in biology. *Trends in Ecology and Evolution*, *37*(1), 53–66. https://doi.org/10.1016/j.tree.2021.08.012

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*, 379–423, 623–656.

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. https://doi.org/10.1098/rsos.191023

van den Heuvel, E., Zhan, Z. (2022). Myths about linear and monotonic associations: Pearson's r, Spearman's ρ, and Kendall's τ. *The American Statistician*, *76*(1), 44–52. https://doi.org/10.1080/00031305.2021.2004922

Wang, Y., Chen, X. (2015). Structural complexity of simplified Chinese characters. In A. Tuzzi J. M. M. Benesová (Eds.), *Recent contributions to quantitative linguistics* (pp. 229–239). De Gruyter. https://doi.org/10.1515/9783110420296-019

Watson, S. K., Heesen, R., Hedwig, D., Robbins, M. M., Townsend, S. W. (2020). An exploration of Menzerath's law in wild mountain gorilla vocal sequences. *Biology Letters*, *16*, 20200380. https://doi.org/10.1098/rsbl.2020.0380

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Addison-Wesley.

Appendices

Appendix A Theory

Here we review the relationship between L, L_r and Pearson correlation

Given two random variables x and y and a sample of n points, $\{(x_1, y_1), ..., (x_i, y_i), ..., (x_n, y_n)\}$, the sample covariance is defined as

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right),$$

where $\bar{x}$ is the sample mean of x and $\bar{y}$ is the sample mean for y, i.e.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Now consider than the random variables are p (the probability of a type) and l (the length/duration of a type) instead of x and y. Then our sample of n points is $\{(p_1, l_1), ..., (p_i, l_i), ..., (p_n, l_n)\}$, one point per type. Accordingly, the covariance between p and l in a sample of points is

$$s_{pl} = \frac{1}{n-1}\left(\sum_{i=1}^{n} p_i l_i - n\bar{p}\bar{l}\right).$$

Recalling the definition of L (Equation 1) and noting that $\bar{p} = \frac{1}{n}$ and $\bar{l} = M = L_r$ (recall Property 2.1), we finally obtain

$$s_{pl} = \frac{1}{n-1}(L - L_r).$$

The sample Pearson correlation is

$$r = \frac{s_{xy}}{s_x s_y},$$

where s_x and s_y are the sample standard deviation of x and y, i.e.

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Proceeding as we did for the covariance, we find that the Pearson correlation between p and l is

$$r = \frac{L - L_r}{(n-1)s_p s_l}.$$

Then it is easy to see that L is a linear function of the Pearson correlation r or s_{pl}. For instance,

$$L = ar + b,$$

where

$$a = (n-1)s_p s_l$$
$$b = L_r.$$

Other linear relationships can be shown similarly.

Appendix B Analysis

We here present complementary analyses, tables and plots.

B.1 The impact of the unsupervised filter

Table B1 and Table B2 show the impact of the unsupervised filter in the optional filter. PUD is a controlled setting for the impact of the filter because it is a collection where tokens are of high quality compared to CV. Thus we expect that the impact of the optional filter is low in PUD. Unexpectedly, the number of tokens reduces substantially (a reduction of the order of thousands) in Chinese, Japanese and Korean. An additional drastic reduction in the observed alphabet size in these languages strongly suggests that the optional filter is not adequate for them. For these reasons, we believe we should not apply the unsupervised filter to these languages because their writing system is essentially a syllabary. We suspect that the actual need for the exclusion could be a combination of sampling problems relating to a large alphabet size (compared to the Latin script) and a heavy- tailed rank distribution that breaks the optional filter. It is well-known that the rank distribution of Chinese characters is long-tailed, spanning two orders of magnitude (Deng et al., 2014), while that of phonemes (the counterpart of letters in many languages using the Latin script) is exponential-like (Balasubrahmanyan and Naranan, 1996; Naranan and Balasubrahmanyan, 1993). However, that issue should be the subject of future research.

In CV, we find that the optional filter has a similar impact in languages concerning the reduction in the number of tokens but higher impacts concerning the reduction of the alphabet sizes, suggesting that presence of strings with strange characters. The three languages with the most marked reduction in alphabet size are French, Spanish, German and Italian, with an alphabet size greater then 100.

Table B1: The impact of the unsupervised filter in the PUD collection. For every language, we show its linguistic family, the writing system (namely script name according to ISO-15924) and various numeric parameters after applying the mandatory filter but before applying the unsupervised filter, that are A, the observed alphabet size (number of distinct characters), n, the number of types, and, T, the number of tokens. A', n' and T' are the respective values of A, n and T after applying the unsupervised filter.

Language	Script	Family	A	A'	n	n'	T	T'
Arabic	Arabic	Afro-Asiatic	47	39	6600	6596	18214	18201
Indonesian	Latin	Austronesian	39	23	4596	4501	16819	16702
Russian	Cyrillic	Indo-European	61	31	7358	7113	15870	15588
Hindi	Devanagari	Indo-European	84	50	4920	4716	21184	20796
Czech	Latin	Indo-European	49	33	7360	7073	15700	15331
English	Latin	Indo-European	39	25	5082	5001	18135	18028
French	Latin	Indo-European	48	26	5593	5214	21084	20407
German	Latin	Indo-European	39	28	6215	6116	18446	18331
Icelandic	Latin	Indo-European	43	32	6175	6035	16385	16209
Italian	Latin	Indo-European	42	24	5944	5606	21815	21266
Polish	Latin	Indo-European	47	31	7329	7188	15386	15191
Portuguese	Latin	Indo-European	47	38	5678	5661	21873	21855
Spanish	Latin	Indo-European	39	32	5765	5750	21083	21067
Swedish	Latin	Indo-European	39	25	5842	5624	16653	16378
Japanese	Japanese	Japonic	1549	609	4990	3345	24899	22538
Japanese-strokes	Japanese	Japonic	1549	609	4852	3345	24737	22538
Japanese-romaji	Latin	Japonic	23	19	4984	4860	24892	24743
Korean	Hangul	Koreanic	1002	401	8031	6424	14475	12540
Thai	Thai	Kra-Dai	89	52	3818	3599	21642	21121
Chinese	Han (Traditional variant)	Sino-Tibetan	2038	814	5224	3154	18129	15436
Chinese-strokes	Han (Traditional variant)	Sino-Tibetan	2038	814	4970	3154	17845	15436
Chinese-pinyin	Latin	Sino-Tibetan	49	44	5224	5038	18129	17885
Turkish	Latin	Turkic	42	28	6793	6587	14092	13799
Finnish	Latin	Uralic	39	24	7076	6938	12853	12701

Table B2: The impact of the unsupervised filter in the CV collection. The content is the same as in Table B1. 'Conlang' stands for 'constructed language', that is an artificially created language. This is not a family in the proper sense as Conlang languages are not related in the common linguistic family sense.

Language	Script	Family	A	A'	n	n'	T	T'
Arabic	Arabic	Afro-Asiatic	44	31	7497	6397	49448	45825
Maltese	Latin	Afro-Asiatic	40	31	8148	8058	44272	44112
Vietnamese	Latin	Austroasiatic	86	41	574	370	1300	938
Indonesian	Latin	Austronesian	28	22	3817	3768	44336	44210
Esperanto	Latin	Conlang	38	27	27932	27759	406725	406261
Interlingua	Latin	Conlang	27	20	5552	5126	31428	30504
Tamil	Tamil	Dravidian	44	29	1525	1210	7580	6439
Persian	Arabic	Indo-European	105	38	13240	13115	1665428	1662508
Assamese	Assamese	Indo-European	60	43	1115	971	2000	1813
Russian	Cyrillic	Indo-European	54	32	31921	31827	638782	637686
Ukrainian	Cyrillic	Indo-European	44	34	14399	14337	120984	120760
Panjabi	Devanagari	Indo-European	48	37	95	84	110	98
Modern Greek	Greek	Indo-European	46	33	5834	5813	37926	37880
Breton	Latin	Indo-European	41	28	4322	4228	38493	38237
Catalan	Latin	Indo-European	67	39	79213	79112	3294506	3294206
Czech	Latin	Indo-European	44	33	16032	15518	150312	147582
Dutch	Latin	Indo-European	41	23	10666	10225	320992	316498
English	Latin	Indo-European	97	28	173522	173023	9829660	9828713
French	Latin	Indo-European	244	49	162740	160243	3732822	3729370
German	Latin	Indo-European	152	30	150362	148436	4235094	4230565
Irish	Latin	Indo-European	31	23	2311	2251	22751	22593
Italian	Latin	Indo-European	110	34	55480	54996	812604	811783
Latvian	Latin	Indo-European	35	27	7792	7251	30358	29456
Polish	Latin	Indo-European	38	32	25365	25340	595613	595411
Portuguese	Latin	Indo-European	41	27	13049	11509	295042	283048
Romanian	Latin	Indo-European	36	29	6449	6423	33370	33341
Romansh	Latin	Indo-European	40	26	9801	9614	44192	43792
Slovenian	Latin	Indo-European	28	24	5994	5937	26402	26304
Spanish	Latin	Indo-European	186	33	75617	75010	1843646	1842474
Swedish	Latin	Indo-European	30	25	4454	4371	63282	62951
Welsh	Latin	Indo-European	43	22	11488	11143	547345	539621
Western Frisian	Latin	Indo-European	42	30	8419	8383	63127	63073
Oriya	Odia	Indo-European	59	41	921	764	1929	1700
Dhivehi	Thaana	Indo-European	40	27	155	111	1388	1284
Georgian	Georgian	Kartvelian	34	25	7945	6505	15481	12958
Basque	Latin	Language isolate	28	21	24998	24748	460188	458071
Mongolian	Mongolian	Mongolic	36	31	14844	14608	70638	70217
Kinyarwanda	Latin	Niger-Congo	96	26	135328	133815	1945038	1939810
Abkhazian	Cyrillic	Northwest Caucasian	37	28	150	119	189	156
Hakha Chin	Latin	Sino-Tibetan	28	23	2515	2499	17806	17776
Chuvash	Cyrillic	Turkic	36	22	5565	4311	16270	13583
Kirghiz	Cyrillic	Turkic	38	30	10497	10130	62687	61844
Tatar	Cyrillic	Turkic	47	34	22313	21823	145458	144356
Yakut	Cyrillic	Turkic	42	28	8041	7904	22795	22577
Turkish	Latin	Turkic	37	31	8957	8926	107910	107686
Estonian	Latin	Uralic	34	23	30135	28691	123895	121549

B.2 Mean word length and the law of abbreviation

In Table B3, Table B4 and Table B5, we show the mean word length (L) and the random baseline (L_r) as well as the outcome of the correlation test between length and frequency for PUD and for CV when length is measured in characters and also in duration, respectively.

Table B3: Mean word length and the correlation between frequency and length in PUD. Word length is measured in number of characters. Mean word length (L) is followed by the random baseline (L_r). Each correlation statistic (Kendall τ or Pearson r) is followed by p-values after applying Holm-Bonferroni correction (rather than being the direct output of the correlation test).

language	family	script	L	L_r	τ	τ_{pvalue}	r	r_{pvalue}
Arabic	Afro-Asiatic	Arabic	4.03	5.54	-0.13	8.32×10^{-32}	-0.13	1.12×10^{-20}
Czech	Indo-European	Latin	5.44	7.27	-0.22	1.20×10^{-113}	-0.15	2.47×10^{-36}
English	Indo-European	Latin	4.87	7.00	-0.20	2.52×10^{-66}	-0.12	6.98×10^{-17}
French	Indo-European	Latin	4.81	7.47	-0.16	2.44×10^{-49}	-0.12	4.24×10^{-19}
German	Indo-European	Latin	5.74	8.56	-0.23	1.25×10^{-108}	-0.12	3.85×10^{-21}
Indonesian	Austronesian	Latin	5.96	7.35	-0.11	6.37×10^{-21}	-0.12	6.53×10^{-15}
Italian	Indo-European	Latin	4.85	7.64	-0.16	4.09×10^{-54}	-0.13	8.45×10^{-23}
Polish	Indo-European	Latin	6.07	8.00	-0.19	1.12×10^{-80}	-0.13	2.78×10^{-26}
Portuguese	Indo-European	Latin	4.35	7.47	-0.20	9.96×10^{-67}	-0.12	1.12×10^{-17}
Russian	Indo-European	Cyrillic	6.04	8.08	-0.19	4.58×10^{-88}	-0.13	4.85×10^{-26}
Spanish	Indo-European	Latin	4.83	7.59	-0.16	4.10×10^{-51}	-0.11	1.89×10^{-17}
Swedish	Indo-European	Latin	5.41	7.99	-0.23	3.99×10^{-101}	-0.13	6.28×10^{-21}
Turkish	Turkic	Latin	6.43	7.94	-0.24	4.26×10^{-124}	-0.12	4.20×10^{-23}

Table B4: Mean word length and the correlation between frequency and length in CV. Word length is measured in number of characters. Content is the same as in B3. 'Conlang' stands for 'constructed language', that is an artificially created language. This is not a family in the proper sense, and Conlang languages are not related in the common family sense.

language	family	script	L	L_r	τ	τ_{pvalue}	r	r_{pvalue}
Abkhazian	Northwest Caucasian	Cyrillic	5.94	6.42	-0.32	4.48×10^{-5}	-0.29	1.43×10^{-3}
Arabic	Afro-Asiatic	Arabic	4.10	5.06	-0.14	5.32×10^{-43}	-0.14	2.04×10^{-28}
Assamese	Indo-European	Assamese	4.57	5.36	-0.31	4.73×10^{-31}	-0.27	3.09×10^{-17}
Basque	Language isolate	Latin	6.41	8.89	-0.16	2.68×10^{-262}	-0.11	6.95×10^{-69}
Breton	Indo-European	Latin	3.97	6.31	-0.24	4.93×10^{-86}	-0.19	4.09×10^{-35}
Catalan	Indo-European	Latin	4.90	8.58	-0.15	0.00	-0.05	9.53×10^{-51}
Chuvash	Turkic	Cyrillic	6.00	7.35	-0.22	5.49×10^{-74}	-0.21	3.80×10^{-43}
Czech	Indo-European	Latin	4.83	7.17	-0.22	1.75×10^{-295}	-0.13	1.69×10^{-58}
Dhivehi	Indo-European	Thaana	3.32	7.61	-0.16	1.65×10^{-2}	-0.31	1.24×10^{-3}
Dutch	Indo-European	Latin	4.72	8.26	-0.28	0.00	-0.10	1.35×10^{-24}
English	Indo-European	Latin	4.61	7.79	-0.07	0.00	-0.03	3.45×10^{-45}
Esperanto	Conlang	Latin	4.83	7.73	-0.18	0.00	-0.08	1.12×10^{-41}
Estonian	Uralic	Latin	6.16	8.85	-0.24	0.00	-0.09	2.55×10^{-48}
French	Indo-European	Latin	5.04	8.13	-0.04	3.57×10^{-85}	-0.04	8.56×10^{-46}
Georgian	Kartvelian	Georgian	7.17	8.22	-0.12	3.67×10^{-31}	-0.10	3.47×10^{-15}
German	Indo-European	Latin	5.73	10.30	-0.12	0.00	-0.04	4.21×10^{-59}
Hakha Chin	Sino-Tibetan	Latin	3.29	5.29	-0.29	4.31×10^{-72}	-0.15	3.88×10^{-13}
Indonesian	Austronesian	Latin	5.37	7.24	-0.20	1.13×10^{-59}	-0.16	2.73×10^{-21}
Interlingua	Conlang	Latin	4.43	7.43	-0.24	8.95×10^{-101}	-0.16	7.39×10^{-31}
Irish	Indo-European	Latin	4.20	6.58	-0.21	2.38×10^{-41}	-0.17	5.18×10^{-15}
Italian	Indo-European	Latin	5.29	8.16	-0.06	2.24×10^{-67}	-0.06	4.19×10^{-49}
Kinyarwanda	Niger-Congo	Latin	6.13	9.20	-0.19	0.00	-0.06	3.32×10^{-117}
Kirghiz	Turkic	Cyrillic	6.01	7.78	-0.19	1.45×10^{-141}	-0.16	6.13×10^{-57}
Latvian	Indo-European	Latin	4.79	7.09	-0.26	5.81×10^{-160}	-0.18	1.36×10^{-53}
Maltese	Afro-Asiatic	Latin	5.07	7.35	-0.20	2.32×10^{-107}	-0.14	1.58×10^{-36}
Modern Greek	Indo-European	Greek	4.85	7.64	-0.24	3.73×10^{-124}	-0.16	1.77×10^{-34}
Mongolian	Mongolic	Mongolian	5.47	7.31	-0.23	1.73×10^{-263}	-0.15	2.31×10^{-76}
Oriya	Indo-European	Odia	4.21	5.35	-0.33	2.00×10^{-28}	-0.31	2.94×10^{-17}
Panjabi	Indo-European	Devanagari	3.68	3.88	-0.32	8.69×10^{-4}	-0.26	8.60×10^{-3}
Persian	Indo-European	Arabic	3.80	5.49	-0.21	2.38×10^{-229}	-0.12	2.06×10^{-45}
Polish	Indo-European	Latin	5.27	7.87	-0.17	8.03×10^{-292}	-0.10	2.47×10^{-58}
Portuguese	Indo-European	Latin	4.53	7.49	-0.19	1.09×10^{-168}	-0.13	1.21×10^{-41}
Romanian	Indo-European	Latin	5.03	7.67	-0.21	3.27×10^{-97}	-0.17	6.46×10^{-41}
Romansh	Indo-European	Latin	4.94	7.56	-0.24	5.91×10^{-184}	-0.15	5.42×10^{-48}
Russian	Indo-European	Cyrillic	6.31	9.00	-0.13	7.75×10^{-225}	-0.09	3.03×10^{-52}
Slovenian	Indo-European	Latin	4.56	6.43	-0.21	1.47×10^{-88}	-0.15	4.71×10^{-29}
Spanish	Indo-European	Latin	5.01	7.92	-0.03	5.95×10^{-32}	-0.04	3.48×10^{-29}
Swedish	Indo-European	Latin	4.04	6.87	-0.28	6.91×10^{-129}	-0.15	1.62×10^{-22}
Tamil	Dravidian	Tamil	5.68	7.08	-0.28	1.01×10^{-35}	-0.23	5.21×10^{-16}
Tatar	Turkic	Cyrillic	5.41	7.45	-0.24	0.00	-0.16	3.15×10^{-118}
Turkish	Turkic	Latin	6.00	8.09	-0.22	1.32×10^{-158}	-0.16	2.51×10^{-48}
Ukrainian	Indo-European	Cyrillic	5.52	7.67	-0.16	3.01×10^{-136}	-0.14	1.74×10^{-61}
Vietnamese	Austroasiatic	Latin	3.24	3.47	-0.19	2.98×10^{-5}	-0.20	1.96×10^{-4}
Welsh	Indo-European	Latin	4.17	7.05	-0.21	2.40×10^{-185}	-0.12	4.39×10^{-38}
Western Frisian	Indo-European	Latin	4.38	7.99	-0.29	1.19×10^{-244}	-0.13	2.62×10^{-33}
Yakut	Turkic	Cyrillic	6.32	7.99	-0.26	5.48×10^{-185}	-0.19	2.12×10^{-65}

Table B5: Mean word length and the correlation between frequency and length in CV. Word length is measured in duration. Content is the same as in B4.

language	family	script	L	L_r	τ	τ_{pvalue}	r	r_{pvalue}
Abkhazian	Northwest Caucasian	Cyrillic	0.74	0.81	-0.20	1.23×10^{-2}	-0.21	2.52×10^{-2}
Arabic	Afro-Asiatic	Arabic	0.46	0.58	-0.12	1.75×10^{-40}	-0.15	2.00×10^{-31}
Assamese	Indo-European	Assamese	0.43	0.50	-0.22	1.25×10^{-17}	-0.19	3.14×10^{-9}
Basque	Language isolate	Latin	0.44	0.63	-0.21	0.00	-0.12	1.29×10^{-78}
Breton	Indo-European	Latin	0.31	0.51	-0.25	1.92×10^{-107}	-0.20	4.94×10^{-39}
Catalan	Indo-European	Latin	0.35	0.68	-0.21	0.00	-0.06	8.70×10^{-69}
Chuvash	Turkic	Cyrillic	0.44	0.54	-0.26	1.18×10^{-116}	-0.22	6.89×10^{-49}
Czech	Indo-European	Latin	0.37	0.57	-0.21	6.40×10^{-295}	-0.14	5.07×10^{-70}
Dhivehi	Indo-European	Thaana	0.32	0.71	-0.17	2.40×10^{-2}	-0.24	1.51×10^{-2}
Dutch	Indo-European	Latin	0.29	0.55	-0.28	0.00	-0.12	1.47×10^{-33}
English	Indo-European	Latin	0.33	0.67	-0.17	0.00	-0.04	4.83×10^{-62}
Esperanto	Conlang	Latin	0.49	0.81	-0.18	0.00	-0.09	1.25×10^{-47}
Estonian	Uralic	Latin	0.39	0.58	-0.23	0.00	-0.09	4.65×10^{-55}
French	Indo-European	Latin	0.32	0.63	-0.21	0.00	-0.04	7.25×10^{-71}
Georgian	Kartvelian	Georgian	0.52	0.61	-0.15	6.51×10^{-51}	-0.12	9.17×10^{-21}
German	Indo-European	Latin	0.37	0.76	-0.22	0.00	-0.05	1.57×10^{-96}
Hakha Chin	Sino-Tibetan	Latin	0.29	0.44	-0.25	9.56×10^{-64}	-0.14	1.10×10^{-11}
Indonesian	Austronesian	Latin	0.38	0.52	-0.22	1.29×10^{-76}	-0.17	8.41×10^{-26}
Interlingua	Conlang	Latin	0.40	0.69	-0.24	9.77×10^{-114}	-0.18	3.80×10^{-36}
Irish	Indo-European	Latin	0.30	0.47	-0.24	1.42×10^{-55}	-0.19	1.02×10^{-19}
Italian	Indo-European	Latin	0.38	0.65	-0.19	0.00	-0.08	4.19×10^{-87}
Kinyarwanda	Niger-Congo	Latin	0.44	0.72	-0.21	0.00	-0.06	7.54×10^{-101}
Kirghiz	Turkic	Cyrillic	0.44	0.57	-0.20	1.38×10^{-159}	-0.16	1.63×10^{-55}
Latvian	Indo-European	Latin	0.39	0.59	-0.23	1.70×10^{-141}	-0.19	4.58×10^{-60}
Maltese	Afro-Asiatic	Latin	0.35	0.54	-0.21	9.96×10^{-140}	-0.15	3.89×10^{-42}
Modern Greek	Indo-European	Greek	0.38	0.63	-0.21	3.20×10^{-105}	-0.17	1.46×10^{-37}
Mongolian	Mongolic	Mongolian	0.36	0.48	-0.25	0.00	-0.15	3.12×10^{-73}
Oriya	Indo-European	Odia	0.39	0.49	-0.33	2.21×10^{-31}	-0.32	1.59×10^{-18}
Panjabi	Indo-European	Devanagari	0.70	0.73	-0.18	4.63×10^{-2}	-0.15	8.44×10^{-2}
Persian	Indo-European	Arabic	0.36	0.54	-0.25	0.00	-0.14	4.66×10^{-58}
Polish	Indo-European	Latin	0.38	0.57	-0.17	0.00	-0.12	2.88×10^{-82}
Portuguese	Indo-European	Latin	0.35	0.61	-0.22	1.15×10^{-243}	-0.15	4.38×10^{-59}
Romanian	Indo-European	Latin	0.36	0.57	-0.23	2.49×10^{-127}	-0.17	2.82×10^{-43}
Romansh	Indo-European	Latin	0.41	0.66	-0.26	7.70×10^{-248}	-0.17	2.06×10^{-64}
Russian	Indo-European	Cyrillic	0.42	0.60	-0.15	2.13×10^{-299}	-0.10	5.30×10^{-75}
Slovenian	Indo-European	Latin	0.44	0.63	-0.25	3.37×10^{-146}	-0.17	3.04×10^{-40}
Spanish	Indo-European	Latin	0.36	0.62	-0.14	0.00	-0.05	2.05×10^{-41}
Swedish	Indo-European	Latin	0.27	0.52	-0.29	1.03×10^{-156}	-0.18	4.76×10^{-32}
Tamil	Dravidian	Tamil	0.54	0.66	-0.31	2.06×10^{-48}	-0.22	5.35×10^{-14}
Tatar	Turkic	Cyrillic	0.38	0.52	-0.26	0.00	-0.17	8.68×10^{-141}
Turkish	Turkic	Latin	0.41	0.54	-0.21	7.11×10^{-158}	-0.16	9.43×10^{-50}
Ukrainian	Indo-European	Cyrillic	0.43	0.59	-0.18	3.01×10^{-176}	-0.16	3.53×10^{-86}
Vietnamese	Austroasiatic	Latin	0.29	0.33	-0.07	4.63×10^{-2}	-0.14	1.40×10^{-2}
Welsh	Indo-European	Latin	0.32	0.58	-0.20	6.25×10^{-197}	-0.15	3.21×10^{-58}
Western Frisian	Indo-European	Latin	0.32	0.61	-0.31	0.00	-0.15	8.59×10^{-43}
Yakut	Turkic	Cyrillic	0.43	0.54	-0.25	2.41×10^{-186}	-0.18	9.50×10^{-58}