The optimality of word lengths. Theoretical foundations and an empirical study

Sonia Petrini[1], Antoni Casas-i-Muñoz[2], Jordi Cluet-i-Martinell[2], Mengxue Wang[2], Christian Bentz[3,4], Ramon Ferrer-i-Cancho[1*]

[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 Language Science and Technology, Saarland University, Saarbrücken, Saarland, Germany.
[4] Chair of Multilingual Computational Linguistics, University of Passau, Passau, Bavaria, Germany.
[*] Corresponding author's email: rferrericancho@cs.upc.edu

DOI: https://doi.org/10.53482/2026_60_434

ABSTRACT

Zipf's law of abbreviation, namely the tendency of more frequent words to be shorter, has been viewed as a manifestation of compression, i.e. the minimization of the length of forms – a universal principle of natural communication. Although the claim that languages are optimized has become trendy, attempts to measure the degree of optimization of languages have been rather scarce. Here we present two optimality scores that are dualy normalized, namely, they are normalized with respect to both the minimum and the random baseline. We analyze the theoretical and statistical advantages and disadvantages of these and other scores. Harnessing the best score, we quantify the degree of optimality of word lengths per language. This includes parallel texts in 20 languages of 9 families, written in 8 scripts, as well as spoken data for 46 languages of 12 families, two constructed languages, and one isolate. Our analyses indicate that languages are optimized to 62 or 67 percent on average (depending on the source) when word lengths are measured in characters, and to 65 percent on average when word lengths are measured in time. In general, spoken word durations are more optimized than written word lengths in characters. Our work paves the way to measure the degree of optimality of the vocalizations or gestures of other species, and to compare them against written, spoken, or signed human languages.

Keywords: word length, compression, optimality score, law of abbreviation

1 Introduction

Language universals – properties that apply to all languages on Earth – are "vanishingly few", as exceptions can often be found (Evans and Levinson, 2009). Some of the promising candidates are linguistic laws (Semple et al., 2022; Zipf, 1949), though these have taken a rather secondary role in mainstream linguistics since Zipf's pioneering research. One of the most robust patterns is Zipf's law of abbreviation: more frequent words tend to be shorter (Bentz and Ferrer-i-Cancho, 2016; Zipf, 1949). It has been

claimed that there is an even stronger law which relates the length of a word with its co-text of occurrence (Piantadosi et al., 2011), but further research has demonstrated that this new law is weaker and not supported across all languages tested (Koplenig et al., 2022; Levshina, 2022b; Meylan and Griffiths, 2021). Therefore, Zipf's law of abbreviation is, at present, one of the strongest laws of communication in terms of theoretical understanding (Ferrer-i-Cancho, Bentz, and Seguin, 2022; Ferrer-i-Cancho, Debowski, and Moscoso del Prado Martín, 2013; Kanwal et al., 2017), and in terms of empirical support across languages (Bentz and Ferrer-i-Cancho, 2016), linguistic levels (Hernández-Fernández et al., 2019; Koshevoy et al., 2023; Torre et al., 2019) and across species (Semple et al., 2022).

However, while Zipf's law of abbreviation – and other linguistic laws – have been investigated under the umbrella term of "efficient communication", there is generally a lack of precise quantification of the *degree of efficiency*. How optimized are natural languages in terms of word lengths across the board? How optimized is one language compared to another? The first question relates to the universal communicative pressures shaping natural languages. The second question relates to the diversity of encoding strategies within the bounds of these universal pressures. Research in this direction is currently hampered by a lack of clear formalizations of baselines. In other words, we have to ask: optimal compared to *what*?

This question is particularly urgent given that natural languages occur in different modalities: spoken, signed, and written. Since languages are often investigated based on written documents, this raises the question to what extent different writing systems influence the measured optimality of word lengths. To illustrate this point, take the following Mandarin Chinese example and the respective parallel sentence in English from the *Parallel Universal Dependencies* (PUD):

Mandarin Chinese (cmn)[1]

(1) 學生通過實際運用來學習科學課程的内容。

學生	通過	實際	運用	來	學習	科學	課程	的	内容	。
xuéshēng	tōngguò	shíjì	yùnyòng	lái	xuéxí	kēxué	kèchéng	de	nèiróng	.
student	through	practice	use.INF	PRT	learn.INF	science	course	POSS	content	.

'Students learn science content by applying it.'

Since this is a parallel text, the overall content is kept approximately constant between the Mandarin Chinese and English sentences. However, the word lengths in UTF-8 characters are vastly different. On average, we have $\frac{18}{10} = 1.8$ Han characters per word, $\frac{56}{10} = 5.6$ Latin characters per word in the Pinyin

[1] This sentence is taken from the file zh_pud-ud-test.conllu (sent_id = n01004009) of the PUD (https://universaldependencies.org/). The tokenization and transliteration into Pinyin is here taken directly from the PUD. The glossing is provided with reference to the online dictionary at https://www.mdbg.net/chinese/dictionary. INF: infinitive verbal form; PRT: particle, here roughly translating as 'in order to'; POSS: possessive particle.

transliteration, and $\frac{39}{7} \sim 5.6$ characters per word in the English parallel sentence. To further complicate the picture, we can also count the distinct strokes the Han characters are composed of. For instance, the first word written in Han characters 學生 *xuéshēng* 'student' consists of overall 21 strokes.

A similar problem arises when we compare word lengths in written characters with spoken durations. Table 1 gives examples deriving from the Common Voice (CV)[2] corpus for French and Spanish. While the French words are slightly longer on average in terms of UTF-8 characters (5.4 versus 5 characters/word), they are in fact shorter on average in spoken durations averaged across speakers (0.29 versus 0.34 seconds/word). Given this state of affairs, we propose to measure the optimality of word lengths with scores which are normalized to be independent of the alphabet size (in UTF-8 characters), the number of word tokens, as well as the number of word types in a text. These can then be used for a less biased comparison of word length optimization across different languages, writing systems, and modalities.

The remainder of the article is organized as follows. In Section 2, we discuss the information-theoretic background of the proposed scores (Section 2.1), and introduce three optimality scores (Section 2.2), as well as the specific baselines, namely, the *minimum baseline*, and the *random baseline* (Section 2.3). These are the building blocks of our optimality scores. In Section 3 and Section 4, we present, respectively, the materials and methods to apply these scores to real languages. In Section 5, we show that one of the scores (Ψ) is the best optimality score according to a wide range of criteria. Ψ indicates that languages are optimized to 62% or 67% on average (depending on the source) when word length is measured in characters, and to 65% when word length is measured in durations. We also find that word lengths in durations are more optimized than word lengths in characters – when language samples are sufficiently large. Moreover, Chinese and Japanese word lengths turn out to be more optimized when measured in characters rather than in strokes (or in Latin characters after romanization). Finally, in Section 6, we discuss the repercussions of our findings on a range of research questions related to the research program on the optimality of languages more generally (Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022). As an outlook, we derive further proposals for future research.

2 Measuring the optimality of word lengths

2.1 Information-theoretic background

Here we aim to shift the main focus from the surface of linguistic behavior (e.g. linguistic laws), and the all-or-nothing logics of absolute universals, to the principles underlying a general theory of natural communication (Ferrer-i-Cancho, 2018; Ferrer-i-Cancho and Gómez-Rodríguez, 2021b; Semple et al., 2022). In this setting, the principles that govern communication across species are the true universals which manifest themselves in production (potentially with exceptions), and may require specific

[2] https://commonvoice.mozilla.org/en/datasets

Table 1: Cognates in French and Spanish with number of written characters, and mean duration in seconds as measured across speakers in the Common Voice (CV) corpus.

Language	Word	No. char.	IPA	Duration
French	est	3	ɛ	0.14
	une	3	ynə	0.17
	sont	4	sõ	0.23
	comme	5	kɔmə	0.22
	informations	12	ɛ̃fɔʁmasjõ	0.67
	Average	5.4		0.29
Spanish	es	2	es	0.19
	una	3	una	0.21
	son	3	son	0.25
	como	4	komo	0.27
	informaciones	13	informaθjones	0.78
	Average	5		0.34

experimental conditions to appear (Ferrer-i-Cancho and Gómez-Rodríguez, 2021a; Ferrer-i-Cancho and Hernández-Fernández, 2013). This shift of focus requires a split between principles on one hand, and their manifestations on the other, as well as a theoretical understanding of when and why the manifestations of a principle reach the surface of the observable world. Since Zipf's pioneering research, the law of abbreviation has been argued to be a manifestation of word length minimization (Ferrer-i-Cancho, Bentz, and Seguin, 2022; Ferrer-i-Cancho, Hernández-Fernández, et al., 2013; Zipf, 1949), currently known as the principle of compression (Ferrer-i-Cancho, Hernández-Fernández, et al., 2013; Semple et al., 2022). In its simplest form, the principle of compression can be formulated as the minimization of 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

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

Table 2: Five most frequent words for English and Mandarin Chinese in the PUD corpus.

English	f_i	l_i	Chinese	f_i	l_i^{hanzi}	l_i^{pinyin}
the	1441	3	的 *de* (possessive particle)	1362	1	2
of	620	2	在 *zài* (preposition 'in')	415	1	3
in	510	2	了 *le* (aspective particle)	380	1	2
to	481	2	一 *yī* ('one')	249	1	2
and	456	3	是 *shì* (copular 'be')	215	1	3

As a concrete example, take the frequency and length distributions for English and Mandarin Chinese words given in Table 2. To calculate the average length L according to Equation 1, we have to multiply each length l_i of a given word type with its probability p_i. The probability, in turn, is derived as the relative frequency of a given type, i.e. a given f_i over the total number of tokens (sum of f_i's). In fact, L can be seen as the *average length of word tokens*, since we take token frequencies into account via the p_i's. Given these definitions, it is clear that L is influenced by the type of writing system (compare Hanzi and Pinyin lengths), the number of tokens T (i.e. text size), the number of types (i.e. vocabulary size), and also the alphabet size A – there will be many more different Han characters than English and Pinyin characters. Writing in a more diverse character set allows for coding of frequent words with single characters, rather than combinations of characters.

The claim that languages are optimized for efficient communication has become trendy (Gibson et al., 2019; Levshina, 2022a; Liu et al., 2017). However, attempts to actually measure the degree of optimization in a given dimension have been rather scarce (Coupé et al., 2019; Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022; Ferrer-i-Cancho and Bentz, 2018; Koplenig, 2021). Here we aim to quantify the degree of optimality of average word lengths L as a window to the strength of compression in languages.

2.2 Optimality scores

The degree of optimality of syntactic dependency distances – reflecting the manifestation of the dependency distance minimization principle – has been investigated for two decades (Ferrer-i-Cancho, 2004; Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022; Gulordava and Merlo, 2015, 2016; Hawkins, 1998; Tily, 2010). In contrast, the degree of optimality of word lengths has been investigated only recently (Ferrer-i-Cancho and Bentz, 2018; Moreno Fernández, 2021; Pimentel et al., 2021). This degree of optimization has been measured with the η-score (Borda, 2011; Ferrer-i-Cancho and Bentz, 2018; Pimentel et al., 2021)[3]:

$$\eta = \frac{L_{min}}{L}, \tag{2}$$

[3]For Pimentel et al. (2021), see Fig. 5.

where L is the average length of tokens, as defined in Equation 1 above, and L_{min} is the *minimum baseline*, namely, the minimum value that L can achieve – given certain assumptions. We say that a language is $x\%$ optimal with respect to η if $\eta = x/100$ (we apply the same convention to other scores). The analyses using η indicate that languages are 30% optimal on average under *non-singular* coding and 40% optimal on average under *unique decodability* (Ferrer-i-Cancho and Bentz, 2018).

Here we will investigate the degree of optimality of languages with two new scores that we develop from recent research on the optimality of dependency distances (Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022). One is

$$\Psi = \frac{L_r - L}{L_r - L_{min}}, \tag{3}$$

where L_r is the random baseline for L. The other is

$$\Omega = \frac{\tau}{\tau_{min}}, \tag{4}$$

where τ is the Kendall correlation coefficient between p_i and l_i, and τ_{min} is the minimum baseline for τ. For further mathematical details and properties of these scores we refer the reader to Section A.

2.3 The baselines

When it comes to defining baselines, researchers often turn to randomly generating data. For instance, Miller (1957) famously proposed to use "monkey typing" as a generative process, and then compare its output to natural language in terms of Zipfian laws. However, there are two main arguments against using random typing as a baseline:

Firstly, in stark contrast to common expectations, Ferrer-i-Cancho, Bentz, and Seguin (2022) prove that random typing is actually an *optimal* coding process. In fact, this is a logical consequence of how random typing is defined: when randomly drawing characters of a given alphabet with predefined probabilities, the overall probability of a string p_i is a function of its length (l_i). If q is the probability of producing a character (not a word delimiter) and the alphabet contains N characters that are equally likely, then the probability of a type of length 1 is

$$p_i(1) = \frac{1-q}{N}.$$

For longer types, types of length l ($l > 1$), we have that

$$p_i(l) = \frac{q}{N} p_i(l-1).$$

Hence, longer types are less likely because $\frac{q}{N}$ is a number between 0 and 1. In other words, there is a strong link between string length and string probability built into the generative process of random typing *a priori*. In particular, more frequent types are shorter as expected from the compression principle.

Secondly, random typing is psychologically implausible. Humans (or animals) do not randomly churn out phonemes or graphemes when communicating. Rather, there is a host of factors – phonotactic, morphosyntactic, semantic, pragmatic, cognitive, sociolinguistic – which govern *what* is said and *how*. While it is impossible to take all of these into account, in the following we aim to define baselines which are more realistic than random typing.

2.3.1 The random baseline

Many different pressures will act on word lengths and probabilities in the course of language change. We do not attempt to model these here. Rather, we simply assume that both the lengths of words and their probabilities are a given – at the point in time when we measure them in corpora. Our null hypothesis is then a random one-to-one mapping of word probabilities onto word lengths, which leads to the random baseline being defined as (Petrini et al., 2023)

$$L_r = \frac{1}{n} \sum_{i=1}^{n} l_i. \tag{5}$$

A random one-to-one mapping can be obtained in three equivalent ways (1) by shuffling the p_i's, (2) by shuffling the l_i's and (3) by shuffling each of them (Petrini et al., 2023). This baseline has been referred to as "shuffle coding" in related work (Pimentel et al., 2021). Note that L_r actually corresponds to the mean length of *word types*, whereas the mean word length L corresponds to the mean length of *word tokens*. If word lengths are randomly mapped onto word probabilities as explained above, the expected value of word lengths is equal to the value when all word types are equally likely, that is, the relative frequency of a word is irrelevant to its length. See Petrini et al. (2023) for further details on this random baseline, as well as a comparison showing that $L < L_r$ on the same languages used for the present article.

2.3.2 The minimum baseline

L_{min} is the minimum value that L can achieve making certain assumptions. For example, if the only assumptions are $l_i \geq 0$ and

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

then $L_{min} = 0$ (Ferrer-i-Cancho, Bentz, and Seguin, 2022). In plain words, not communicating at all would reduce L_{min} to zero, and hence lead to maximum compression. Of course, this is not a realistic scenario if we assume that there are pressures to communicate. We might introduce another constraint,

namely, that strings of length zero are not considered valid types, but rather that $l_i \geq 1$. In this case $L_{min} = 1$. That is, the minimum average length of words should be one. As a matter of fact, however, not even scripts with a rich set of characters will allow for just words of length one (remember the 1.8 Han characters per word from above).

To derive a more realistic minimum baseline, we follow standard information theory and its extensions. Here, the p_i's are assumed to be given, and optimization is assumed to operate only on the l_i's (Ferrer-i-Cancho, Bentz, and Seguin, 2022; Ferrer-i-Cancho and Bentz, 2018). Arguably, this also makes sense from a linguistic point of view. How often we use a word is guided by many factors, its length will play a minor role – if any. For example, if we have a conversation about kitchen appliances, we might have to refer to the concept of a *refrigerator* a certain number of times. We cannot simply replace this word by other words – *oven*, *freezer*, *microwave*, etc. – just because these are shorter. However, if we have to refer to the concept often, we might shorten its length to *fridge*. So in this case, compression acts on the length of the word *given* its probability.

Previous research on the optimality of word lengths has taken into account different kinds of coding schemes (e.g. non-singular coding and uniquely decodable coding) from standard information theory (Ferrer-i-Cancho and Bentz, 2018; Moreno Fernández, 2021). A limitation of these approaches is that, in order to calculate L_{min}, one has to assume that word length is discrete, and that all characters have the same weight. Also, one has to choose an alphabet size (for instance, one has to decide if the alphabet size will be fixed or will depend on the language). Here we focus on computing L_{min} according to the so-called *rank ordering* (RO) method (Moreno Fernández, 2021), which does not suffer from the limitations above. This method has been referred to as "Zipfian coding" in related work (Pimentel et al., 2021). In particular, L_{min}^{RO} is obtained when the current values of l_i are reassigned to frequencies (or equivalently probabilities) so as to minimize L. Namely, L is minimized when probabilities are sorted decreasingly and lengths are sorted increasingly (Ferrer-i-Cancho, Bentz, and Seguin, 2022). In this case, the i-th most frequent type gets the i-th shortest length, hence the name *rank ordering*. Throughout this paper and appendices, L_{min} normally refers to L_{min}^{RO} – unless indicated otherwise. In parallel to L_{min}, τ_{min} is also defined by the rank ordering principle. See Section B and Section C.1 for further mathematical details and discussions of this and other minimum baselines.

In a nutshell, using the rank ordering minimum baseline, we hold that the effort of communication would be minimized if indeed there was a perfect inverse match between the probabilities of words and their length, namely, if there was a perfect agreement with Zipf's law of abbreviation. The question then is how far away from this optimum a given language is.

2.4 The properties of the baselines and optimality scores

In summary, we investigate optimality scores, i.e. η, Ψ and Ω, which are defined using the baselines L_r and L_{min}^{RO}. By choosing the previous minimum baselines, the scores measure the *closeness to a perfect law of abbreviation*. A perfect fit to the law of abbreviation means that the lengths are arranged optimally given the probabilities of words. Importantly, note that the minimum and the random baseline share various statistical properties with the original source:

1. the distribution of word frequencies and the distribution of word probabilities,
2. the number of tokens and the number of types,
3. the alphabet size in the case of written language (and the repertoire of phonemes and larger constructs in the case of spoken language).

In extension, our optimality scores assume that all these statistical characteristics are fixed; only the one-to-one mapping of word probabilities into word lengths can be changed. A fundamental question is whether a given comparison between two languages is supported by the mathematical properties of the optimality scores used. In the following, we discuss different scenarios of how frequency/length distributions used in communication can relate to one another, and what this implies for the usage of our optimality scores.

2.5 General problems when comparing communication systems

Borrowing the general setting of previous research into Zipf's law of abbreviation (Ferrer-i-Cancho, Bentz, and Seguin, 2022), we might reduce a communication system to a mapping from types to codes, which may not be discrete. The key of the mapping is the code lengths, represented as natural numbers when we measure the length of words in characters, or real numbers when we measure the duration of a vocalization or a gesture (Semple et al., 2022). In this general setting, suppose that we have two communication systems, A and B.

2.5.1 The weak recoding problem

In this recoding problem, B is just a transformation of A assigning a new code, e.g. a new string, to each type of A (hence preserving the frequency of every type from A in B while changing their length). Then suppose that $s(X)$ is the optimality score of a communication system X. The question is, under what conditions $s(A) = s(B)$? This question cannot be answered unless we clarify what we consider to be relevant or not for $s(A) = s(B)$. Invariance under *increasing linear transformation*, for instance, is a possible requirement: we will have $s(A) = s(B)$ when the score satisfies this type of invariance, i.e. the new lengths in B are an increasing linear transformation of those in A. Similarly, another condition

for $s(A) = s(B)$ can be invariance under *strictly increasing transformation*. A set of mathematical properties including the aforementioned ones is discussed in Section A with regards to the proposed optimality scores. In this article in particular, we will deal with three instances of the weak recoding problem:

1. **Length in characters versus duration in the same language**. In this instance, the lengths in characters in one of the communication systems, say A, have been replaced by the corresponding durations in time to produce the other, say B. This corresponds to the mapping of word lengths in characters to durations for Spanish and French examples in Table 1.

2. **Length in Chinese/Japanese characters versus other discrete lengths (romanizations and strokes)**. Here, lengths in Chinese/Japanese characters are replaced by lengths in strokes or in romanizations, namely Latin script conversions (via Pinyin for Chinese and Romaji for Japanese). Such replacements certainly lead to an increase in L (see Table 2 as well as Table D1). However, a key question is whether the *degree of optimality* will also change as a consequence of this increase in L.

3. **Removal of vowels in Latin scripts or romanizations**. The motivation for this recoding is to check if the scores are robust to varying decisions with regard to the design of a writing system. For instance, writing systems differ in how they code for vowels. Catalan – as all Romance languages – uses a Latin script, where both consonants and vowels are represented. In contrast, the primary writing system of Arabic is an abjad, where only consonants are represented (unless *fatha* diacritics are used to indicate vowels). Removing vowels in Romance languages certainly leads to a decrease in L, but, again, the key question is whether their degree of optimality will also change.

Suppose now that coding in A is non-singular, namely no two distinct types are assigned the same code. After replacing the strings in A by new strings, it may turn out that B ceases to be non-singular. That could happen because two types in A end up having the same code or one type is assigned the empty string. This motivates the need to define a strong recoding problem where non-singularity is preserved.

2.5.2 Strong recoding problem

In this recoding problem, B is initially obtained from A as in the weak recoding problem, but additionally, all types in B that are assigned the same string are merged into the same type, and types that are assigned the empty string are dropped. Notice that strong recoding may change the distribution of word frequencies, a feature that the weak recoding problem preserves. Notice that if A is non-singular then B will also be non-singular. All the instances of the weak recoding problem presented above can be investigated under the framework of the strong recoding problem as well. However, in this article, we

investigate directly only the weak recoding problem for the sake of simplicity and due to pressure for space.

2.5.3 The free comparison problem

Above, we have introduced problems where B stems from A. The free comparison problem appears when A and B are *a priori* two different communication systems, hence the number of types and their distributions differ. This setting raises the question of whether the optimality scores of two languages with distinct writing systems or distinct evolutionary histories are comparable. This is a big research challenge for analyses of the degree of optimality in written languages, given the disparity of writing systems of the world (Daniels, 1990), which is also reflected in the dataset in the present study (Table 3 and Table 4). Some more specific questions in this context are:

1. Can we compare Romance languages against Standard Arabic, although the former code for vowels and the latter essentially do not? We will show that removing vowels in Romance languages does not have a big impact on the values of the optimality scores, suggesting that the original scores for Romance languages are comparable to those of Standard Arabic.

2. Can we compare Romance languages – which use a script that essentially codes for phonemes – against syllabic/logographic writing systems, namely systems that use a character for every syllable? Examples of syllabic/logographic writing systems are Chinese Hanzi, where every character stands for a syllable, or abugida writing systems.

We here aim to design scores that abstract away from cultural differences, hence providing an insight into the actual degree of optimality of word lengths. Ultimately, an entirely "fair" comparison may be impossible to accomplish. Our take is that we define scores that can at least give reliable results under ideal conditions.

2.6 Desirable properties of scores

The score fulfilling most mathematical properties (i.e. most checks in Table A1 of Section A) is not necessarily the best score. Further properties are also desirable.

Firstly, our baselines, and therefore the scores that are defined on top of them, assume that the number of tokens (T), the number of types (n), and the alphabet size (A) are fixed, and that only the mapping of probabilities into words can be changed (Section 2.4). In this setting, a desirable property of a score is that it is not influenced by these characteristics. If a score is heavily correlated with any of these parameters across languages, it can be argued that, rather than observing differences in the degree of word length optimality of languages, we are observing differences in text length (in tokens), or in the size of the alphabet, or the vocabulary. A most critical dependence would be on the number of tokens. This

would be particularly worrying when using non-parallel corpora, where higher variation in text length is expected in comparison to parallel texts.

Secondly, an important question is whether a score converges to a stable value, given a sufficiently large language sample, and how fast this convergence is. This approaches the problem of dependency on the number of tokens within a given language. The equivalent problem across languages is tackled by the first point above.

Thirdly, another important and related question is whether we can replace more complex scores by simpler ones, regardless of all the theoretical arguments summarized in Table A1. In the extreme case, could one of the best scores, say Ψ, be replaced simply by L because the former is strongly correlated with the latter?

All these questions are difficult to investigate theoretically and will be tackled empirically at the beginning of Section 5 so as to help choose a primary score for reporting results.

3 Material

We borrow a dataset with information about word length and word frequency from recent research (Petrini et al., 2023), available in the repository of this article (Petrini et al., 2026). The dataset has been extracted from two collections of texts: *Common Voice Forced Alignments*,[4] hereafter CV, and *Parallel Universal Dependencies*,[5] hereafter PUD. PUD comprises 20 distinct languages from 9 linguistic families and 8 scripts (Table 3). CV comprises 46 languages from 12 linguistic families, two constructed languages, one isolate (Basque), and overall 10 scripts (Table 4). The typological information (language family) is obtained from Glottolog 4.6.[6] The writing systems are determined according to ISO-15924 codes. See Table 3 and Table 4 for basic statistical properties of the datasets. These are reproduced for convenience from Petrini et al. (2023).[7]

[4] https://commonvoice.mozilla.org/en/datasets, for the forced alignments see https://github.com/JRMeyer/common-voice-forced-alignments.

[5] https://universaldependencies.org/

[6] https://glottolog.org/

[7] Notice that the counterpart of this table in Petrini et al. (2023), Table 3, was distorted by bugs that were fixed to generate the version in this article.

Table 3: Summary of the main characteristics of the languages in the PUD collection. For each language, we show the linguistic family, the writing system (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	39	6596	18201
Indonesian	Austronesian	Latin	23	4221	16305
Russian	Indo-European	Cyrillic	31	7113	15588
Hindi	Indo-European	Devanagari	50	4716	20796
Czech	Indo-European	Latin	33	7069	15286
English	Indo-European	Latin	25	5001	18021
French	Indo-European	Latin	26	5211	19812
German	Indo-European	Latin	28	6108	18003
Icelandic	Indo-European	Latin	32	6035	16207
Italian	Indo-European	Latin	24	5528	19935
Polish	Indo-European	Latin	32	7204	15216
Portuguese	Indo-European	Latin	37	5621	20367
Spanish	Indo-European	Latin	32	5689	20602
Swedish	Indo-European	Latin	25	5624	16369
Japanese	Japonic	Japanese	1577	4852	24737
Japanese-strokes	Japonic	Japanese	1549	4852	24737
Japanese-romaji	Japonic	Latin	28	4849	24734
Korean	Koreanic	Hangul	1002	8031	14475
Thai	Kra-Dai	Thai	52	3599	21121
Chinese	Sino-Tibetan	Han (Traditional variant)	2071	4970	17845
Chinese-strokes	Sino-Tibetan	Han (Traditional variant)	2038	4970	17845
Chinese-pinyin	Sino-Tibetan	Latin	54	4970	17845
Turkish	Turkic	Latin	28	6373	13512
Finnish	Uralic	Latin	24	6933	12691

Table 4: Summary of the main characteristics of the languages in the CV collection. For every language we show its linguistic family, the writing system (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

3.1 The dataset

The dataset provides the length of a word in characters (in PUD and CV) and its duration (in CV only) in two variants, median duration and mean duration (median is used by default, but mean is also used in certain cases as a control). It also provides the length in strokes and in romanizations for Japanese and Chinese (Romaji and Pinyin, respectively) in PUD. The traditional writing systems of Japanese and Chinese yield very short word lengths in characters, while the number of distinct characters is very large, especially compared to Western languages with mostly alphabetic writing systems (Chen et al., 2015; Joyce et al., 2012).[8]

3.1.1 Common Voice Forced Alignments

Notice that Abkhazian, Panjabi, and Vietnamese have a critically low number of tokens (less than 1000 tokens according to Table 4). However, we decided to include them in the analyses so as to better understand problems with sample size.

3.1.2 Parallel Universal Dependencies

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 3). Thus, Japanese in strokes follows the setting of the weak recoding problem only approximately.

4 Methodology

All the code used to produce the results is available in the repository of this article (Petrini et al., 2026). The bulk of the methods are borrowed from a preceding article, including the unsupervised method to filter words with unusual or "foreign" characters (Petrini et al., 2023). Next we highlight methods or variants thereof which are specific to this article.

4.1 Tokenization

Tokenization into word forms is provided by the respective corpora. Both the PUD and CV (forced alignments) provide their own splits of sentences in written and spoken form into tokens. There are some issues with tokenization which are relevant for our results. To illustrate this, take the same example sentence as above in Japanese.

[8] See Table 3 for alphabet size and Table D1 for average word length in Chinese and Japanese.

Japanese (jpn)[9]

(2) それ適用することで、生徒は科学の内容を学習する。

それ	適用	する	こと	で	、	生徒	は	科学	の	内容	を	学習
sore	tekiyou	suru	koto	de	,	seito	wa	kagaku	no	naiyou	o	gakusyuu
it	applying	to.carry.out	matter	PRT	,	student	TOP	science	POSS	content	OBJ	learning
する	。											
suru	.											
to.carry.out												

'Students learn science content by applying it.'

In Japanese, particles marking topic (は *wa*), possession (の *no*), and direct objects (を *o*) are here treated as separate words, while another grammatical analysis might treat them as affixes to the respective nouns (e.g. 生徒は *seito-wa*), hence creating longer word types.

In general, the UD corpus follows a syntactic definition of "word". According to its designers, "the basic units of annotation are syntactic words (not phonological or orthographic words), which means that we systematically want to split off clitics, as in Spanish *dámelo* = *da me lo*, and undo contractions, as in French *au* = *à le*".[10] Although it is possible to recover the unsplit versions from the PUD corpora, we here use only the split versions. After all, this is the default design choice of UD. Also, this will enable further research on the interplay between word length and syntactic dependency structures (Ferrer-i-Cancho and Gómez-Rodríguez, 2021b).[11]

4.2 Weak recoding problem

When measuring word length in Latin script compared to strokes in Japanese or Chinese, we simply preserve the original character types but recompute their length, in agreement with the definition of the weak recoding problem we introduced in Section 2.5. We do not recompute types according to distinct strings in Latin script (that would be the strong recoding problem, which is not the target of this article). It is possible and expected that distinct character types are assigned the same string in Latin script. For instance, 書く, 格, 角, 核, 欠く, 画, 各, 佳句, 確, 斯く, 昇く, 殻, 隔, 膈 are all written as 'kaku' in Romaji. Hence, Japanese romanizations do not comply with the setting of the strong recoding problem.

[9] This (part of a) sentence is taken from the file ja_pud-ud-test.conllu (sent_id = n01004009). The tokenization is here taken directly from the PUD. The transliteration into Romaji is taken from the repository of Petrini et al. (2023). The glossing is provided with reference to the online dictionary at https://jisho.org/ as well as the Japanese grammar by Akiyama and Akiyama (2002). PRT: case particle, here to be translated as 'by'; TOP: topic marker (similar to nominative case); POSS: possessive marker (similar to genitive case); OBJ: direct object marker (similar to acusative case).

[10] https://universaldependencies.org/u/overview/tokenization.html.

[11] Due to a difference in pre-processing, both the split and the unsplit versions contributed word tokens in the analyses of (Petrini et al., 2023). Here, only the split version contributes the word tokens for analyses.

4.3 Immediate constituents in writing systems

When measuring word length in written languages, we are mainly using *immediate constituents* of written words. In Romance languages, we are measuring word length in letters of the alphabet, the immediate constituents in the written form, which are a proxy for phonemes. For syllabic writing systems (as Chinese in our dataset), the immediate constituents of words are characters that correspond to syllables. In addition, for Chinese and Japanese, we are considering two other possible word length units: strokes and letters in Latin script romanizations. In the hierarchy from words to other units, only the original characters are immediate constituents. This 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). For simplicity – and to avoid that these two languages are hence over-represented in statistical analyses on the correlation between scores and certain parameters at the level of languages (Section 2.6) – we will use only their immediate constituents (namely original characters). Thus, the results based on original characters are compared with those for the other languages.

5 Results

In Section 1, we highlighted the importance of distinguishing between principles and their manifestations. In Section 2, we have addressed the problem of how to measure the degree of compression of word lengths through the relationship between the frequency of a type and its length, examining distinct scores for the optimality of word lengths. In Section C, we show that η, Ψ, and Ω are indices of the intensity of compression because they always reach a value of *one* when the system is fully optimized according to the minimum baseline (rank ordering) in Section 2.3. In contrast, Pearson r and Kendall τ correlations fail to give a constant value when the system is fully optimized. Therefore, r and τ are suitable to investigate the manifestation of compression, i.e. the law of abbreviation (Petrini et al., 2023), but are a poor reflection of the actual degree of optimality with respect to the minimum baseline.

In this results section, we first provide some intuitive understanding of the optimality scores. Secondly, we establish that Ψ is the best score in terms of mathematical properties (Table A1), and other desirable properties (Section 2.6). Thirdly, we choose Ψ accordingly to measure the degree of optimality of languages. Lastly, we compare Ψ against the other optimality scores (excluding the correlation scores, r and τ, for the reasons mentioned above).

5.1 Understanding the optimality scores

Here we focus on a qualitative understanding of the optimality scores η, Ψ and Ω as a preparation for the results that are presented below. All these scores are ratios of the form y/x (Equation 2, Equation 3, and Equation 4). That is, they give a *percentage* of word length optimization with respect to some baseline(s).

Figure 1, Figure 2, and Figure 3 show y as a function of x for each of them. A deeper understanding of the expected behavior of these scores requires taking into account the mathematical properties of the scores (Section A) that are indicated in parenthesis for the mathematically oriented reader. By definition, all the data points have to fall below the identity line ($y = x$) because the scores are designed to satisfy $y \leq x$ (the maximum value of these scores is 1). In case of optimal coding (minimum word length), languages would fall on the identity line (because these scores exhibit constancy under optimal coding). Figure 1, Figure 2, and Figure 3 indicate that languages are not coded optimally.

The expected behavior of the scores in other conditions depends on the score. For example, in the absence of any word length effect (for or against compression), languages are expected to be on the abscissa axis (the line $y = 0$) in case of Ψ and Ω (because these scores exhibit stability under the null hypothesis). In case of η, languages are going to be distributed according to $x = L = L_r$ (because η is not stable under the null hypothesis).

In case of some compression of word lengths, languages must be over the line $y = 0$ in case of Ψ and Ω – which is what these figures show. In the case of η, a comparison of its actual value against its expected value under the null hypothesis – that is L_{min}/L_r – is required (because η lacks stability under the null hypothesis). Figure 2 and Figure 3 reveal the presence of some degree of compression. In fact, points are in general closer to the diagonal than to the abscissa axis.

A central score in the analyses to come is Ψ. The percentage Ψ yields is with respect to $L_r - L_{min}$, that is the gap between the *minimum possible* mean word length (obtained by assigning the i-th most frequent word the i-th shortest length in a language and recomputing mean word length) and the *random* mean word length (obtained by shuffling at random word lengths or word probabilities in a language). In more detail, Ψ is the percentage of this gap ($L_r - L_{min}$) in relation to the gap $L_r - L$, i.e. the gap between the *actual* word lengths and the *random* mean word length. As an example, in Mandarin Chinese (when the units are Han characters), $\Psi = 72.3\%$ is the percentage of the gap $x = L_r - L_{min} = 2.18 - 1.47 = 0.71$ in relation to the gap $y = L_r - L = 2.18 - 1.67 = 0.41$) (Figure 1 and Table D1).

Finally, notice that similar scores, or their components, have already appeared implicitly in previous research on the optimality of word lengths (Pimentel et al., 2021). In Figure 4 of Pimentel et al. (2021), one finds the gap $L_r - L$, the numerator of Ψ (Equation 3), in the x-axis of the left panel as well as $1/\eta$ in the y-axis of the right panel (but we used *rank ordering* to define L_{min} for η as in Equation 2).

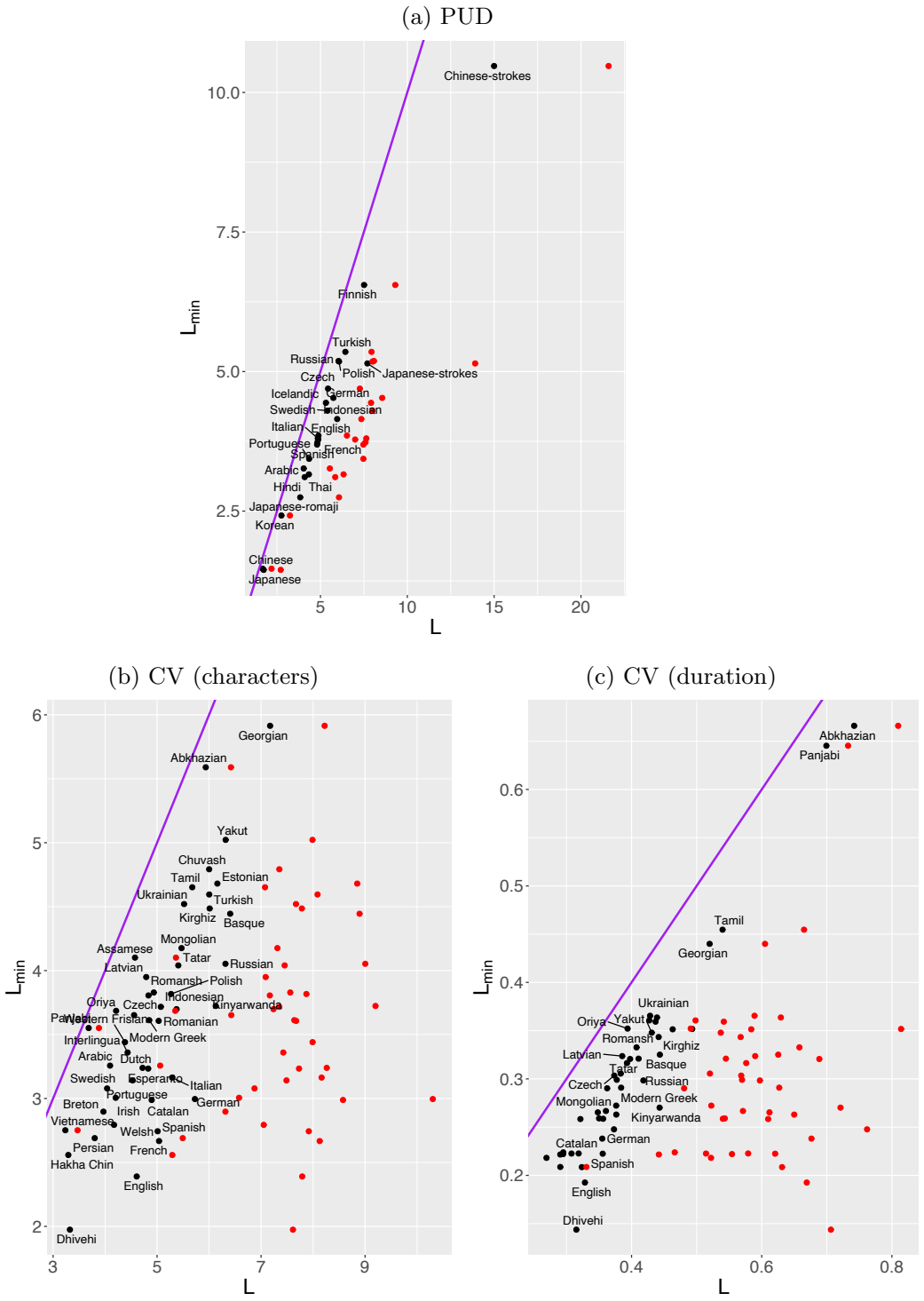


Figure 1: The ingredients of the ratio η: L_{min} (the minimum baseline) versus L (the actual word length). The purple line is the identity function (slope of 1 and zero intercept). Red points indicate the points that are expected under the null hypothesis (where $L = L_r$ is expected). (a) PUD collection. (b) CV collection with length measured in characters. (c) CV collection with length measured in duration.

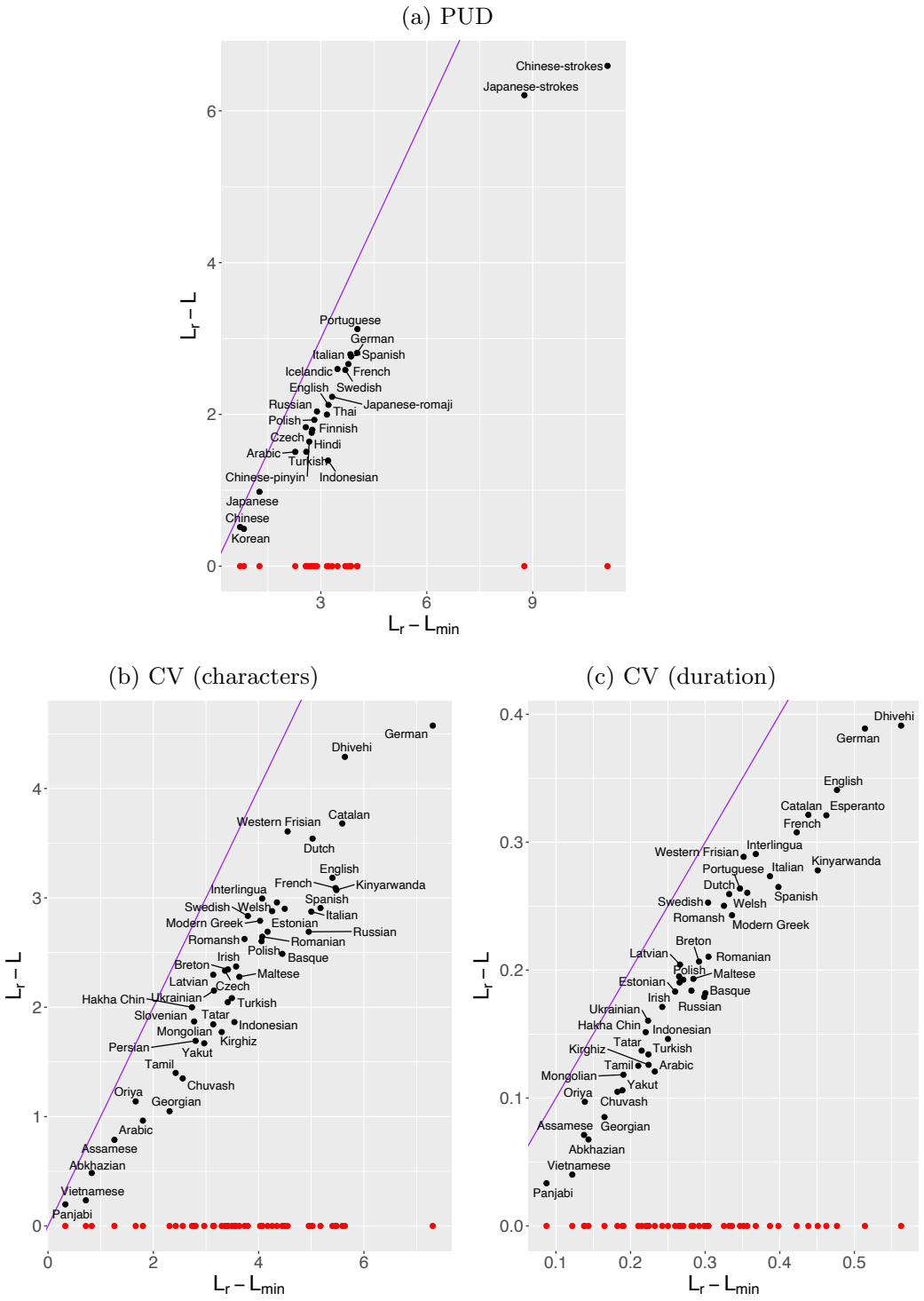


Figure 2: The ingredients of the ratio Ψ: the gap $L_r - L$ versus the gap $L_r - L_{min}$. The purple line is the identity function (slope of 1 and zero intercept). Red points indicate the points that are expected under the null hypothesis (where $L = L_r$ is expected). (a) PUD collection. (b) CV collection with length measured in characters. (c) CV collection with length measured in duration.

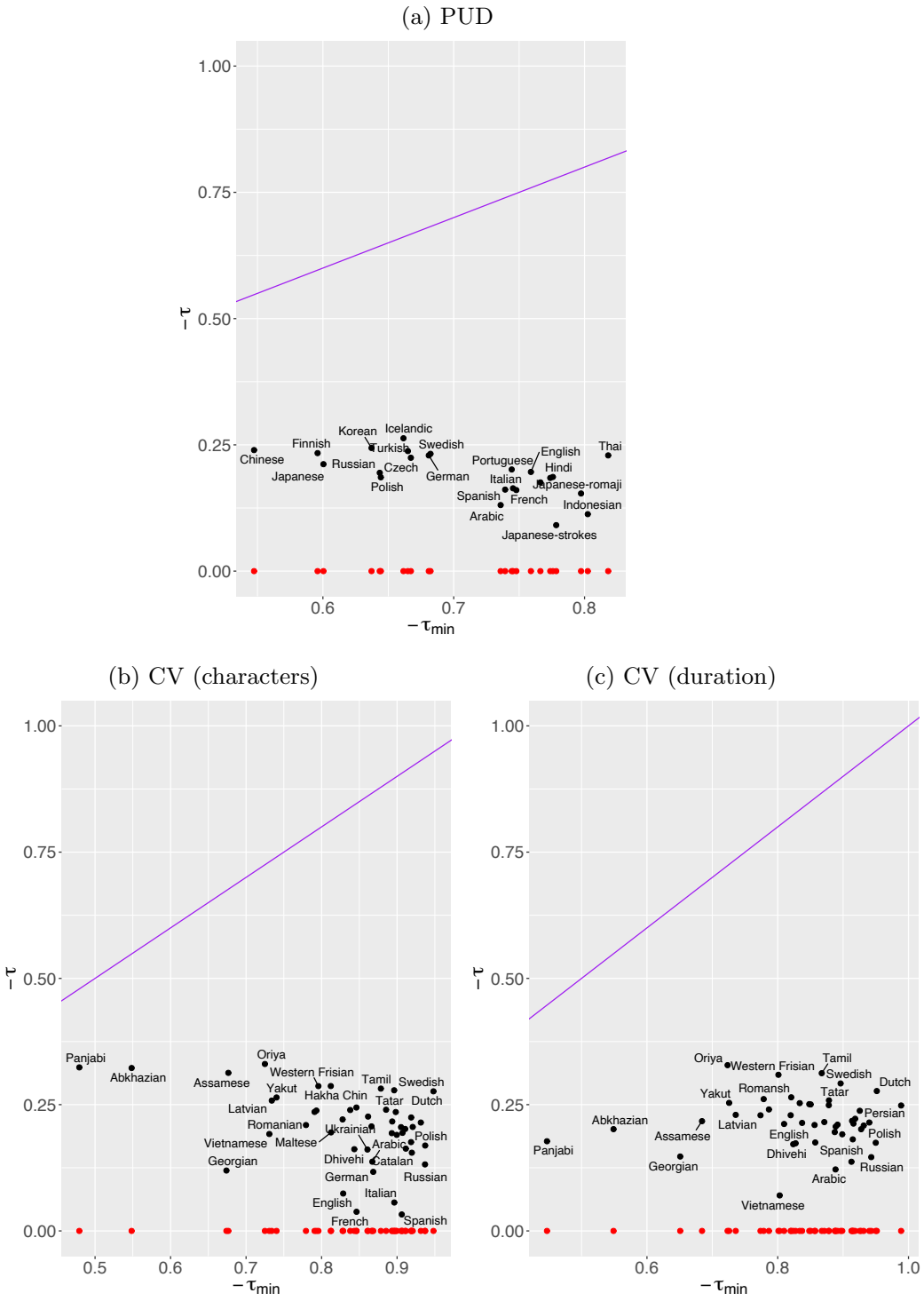


Figure 3: The ingredients of the ratio Ω: negative Kendall τ correlation between frequency and length, versus the negative τ_{min}, the minimum baseline for τ. Here we use the definition of Ω Equation 6 with $\tau_r = 0$, hence the minus sign in the correlations displayed by each axis. Red points indicate the points that are expected under the null hypothesis (where $\tau = \tau_r = 0$ is expected). The purple line is the identity function (slope of 1 and zero intercept). (a) PUD collection. (b) CV collection with length measured in characters. (c) CV collection with length measured in duration.

5.2 Desirable properties of the scores

In order to identify the best score for cross-linguistic research on the degree of optimality, we take into account what we consider to be desirable properties for an optimality score, in addition to those that can be established by a purely mathematical analysis (Table A1).

First, a score should not be related to the basic language parameters – namely observed alphabet size (number of distinct characters, A), observed vocabulary size (number of types, n), and sample size (number of tokens, T) – which are assumed to be constant in the definition of the baselines. The only score consistently satisfying these properties across all collections and length definitions is Ψ, while both η and Ω show some degree of significant association with A and T when data is more heterogeneous, i.e. in the CV collection (Figure E1).

Second, the optimality score of a given language should not depend too strongly on the size of its available sample, meaning that it should converge, and ideally it should converge fast. Again, the score which gets closest to this ideal behaviour is Ψ. It varies within a rather small range, and appears to approach convergence in different languages. In contrast, both η and Ω keep their decreasing trend even for large sample sizes, yet with different speeds (Figure E2, Figure E3 and Figure E4).

Third, a complex score should not be replaceable by a simpler score, namely it should not be significantly and strongly associated to it. When using parallel data, none of our scores is replaceable by L, the simplest score. We only find that Ω could be replaced by η (Figure E5). Thus Ψ is the best complex score for parallel data. See Section E for further details on the analysis of the desirable properties.

5.3 The distribution of optimality values across languages

Here we review the distributions of optimality values for each of the scores. For the scores on word length in characters, we excluded strokes and romanizations for Chinese and Japanese for the sake of homogeneity (i.e. we use only immediate written word constituents). In Figure 4, we show the density plots for both collections, color-coded by definition of length. All scores show a rather narrow distribution, both in PUD and in CV, despite its heterogeneity and large size in terms of languages. The bulk of the values is far from zero. However, this is only informative for scores whose expected value under the null hypothesis is zero, that is Ψ and Ω. The large observed values of η, on the other hand, are not informative in this sense. Interestingly, the values of Ψ are farther from zero than those of Ω.

In Table 5, we report the main summary statistics. In all cases, Ψ and Ω take positive values suggesting some degree of optimization in every language. Indeed, the correlation tests on the same data support that Ψ or Ω are significantly large in each individual language (Petrini et al., 2023). The magnitude strongly depends on the score used. Concerning Ψ, our preferred score for cross-linguistic analysis, the

mean values computed in the three considered scenarios (in a range from 0 to 100%) are close to each other: 67% in PUD with length measured in characters, 62% in CV when considering characters, and 66% when considering duration.

In Table D1, Table D2 and Table D3, we detail the values of all scores and those of their ingredients for each language and collection.

Table 5: Summary statistics of optimality scores η, Ψ and Ω for every collection and definition of length. For length in characters, we only use immediate word constituents for the sake of homogeneity. Accordingly, scores in strokes or in romanizations for Chinese and Japanese in PUD are excluded. Refer to Table D1,Table D2 and Table D3 for values on individual languages.

Score	Collection	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sd
η	PUD-characters	0.69	0.78	0.80	0.81	0.85	0.88	0.05
	CV-characters	0.52	0.69	0.75	0.73	0.80	0.96	0.10
	CV-medianDuration	0.46	0.72	0.76	0.76	0.81	0.92	0.09
Ψ	PUD-characters	0.41	0.65	0.69	0.67	0.71	0.78	0.08
	CV-characters	0.33	0.57	0.63	0.62	0.68	0.79	0.09
	CV-medianDuration	0.33	0.60	0.69	0.66	0.73	0.83	0.11
Ω	PUD-characters	0.11	0.23	0.29	0.29	0.35	0.44	0.08
	CV-characters	0.04	0.19	0.24	0.26	0.30	0.68	0.12
	CV-medianDuration	0.09	0.22	0.25	0.26	0.31	0.45	0.07

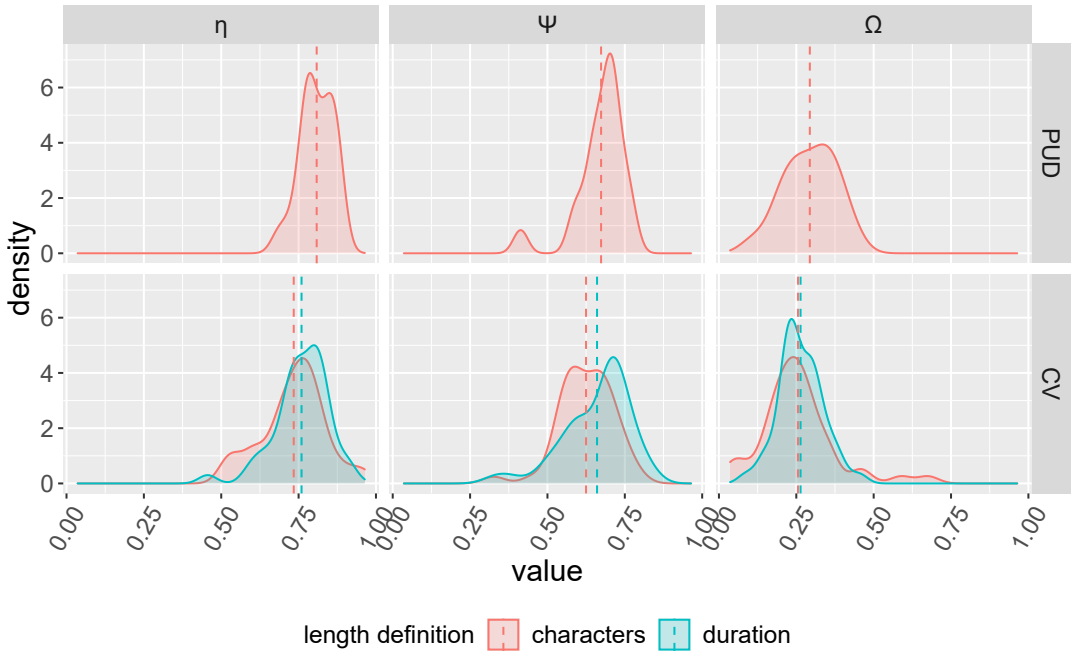


Figure 4: Density distribution of the scores η, Ψ and Ω for length in characters or duration over languages in the PUD and CV collections. For length in characters, we only use immediate word constituents for the sake of homogeneity. Accordingly, scores in strokes or in romanizations for Chinese and Japanese in PUD are excluded. Refer to Table D1, Table D2 and Table D3 for values on individual languages. The vertical dashed lines show mean values (the values of the means are shown in Table 5).

5.4 Sorting languages by their degree of optimality

Here we focus on PUD, as it is the only parallel corpus, and on Ψ, for the sake of simplicity given its mathematical and statistical properties. Sorting languages in CV might lead to misinterpretations, as its intrinsic heterogeneity hampers valid cross-linguistic comparisons (see Section E for problems and risks of drawing conclusions from CV).

In Figure 5(a), we show the languages in the PUD collection sorted decreasingly by Ψ. In Figure 5(b), we display a breakdown of the composition of Ψ intended to help understand its definition as a ratio of two differences.

Tables showing the concrete values of the scores and further details on both collections can be found in Section D.1.

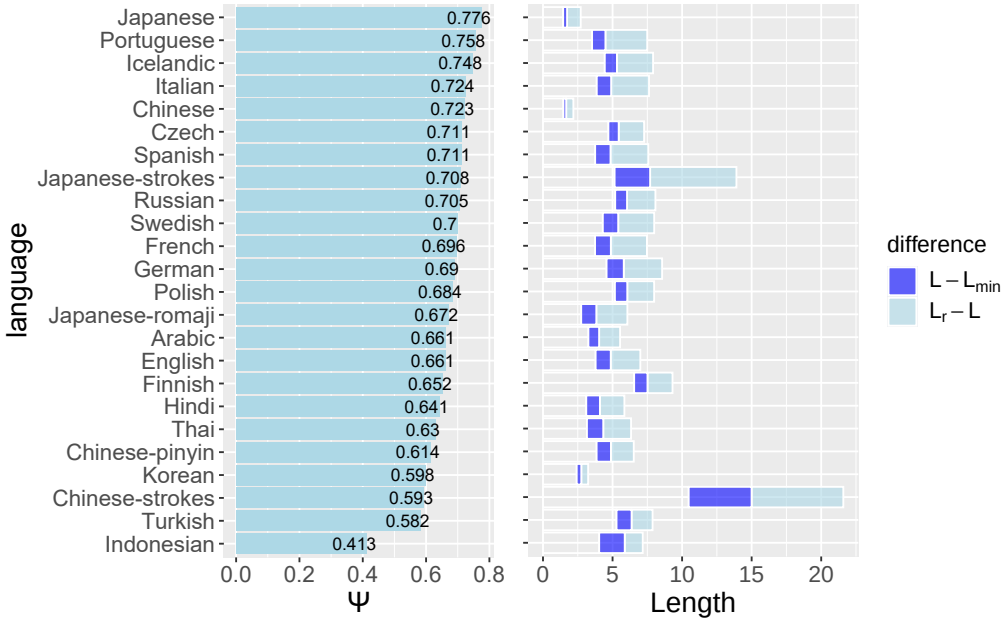


Figure 5: Optimality of word lengths in PUD according to Ψ with length measured in characters. (a) Absolute values of Ψ. (b) Visual depiction of the composition of $\Psi = (L_r - L)/(L_r - L_{min})$ using bars that consist of two segments, a dark blue one followed by a light blue one. The bar starts at L_{min}, changes color at L, and ends at L_r. The dark blue segment of the bar is the difference $L - L_{min}$. This quantity is positive because $L_{min} \leq L$ by definition. The light blue segment is the difference $L_r - L$. This quantity is positive because $L < L_r$ due to compression (Petrini et al., 2023). Hence the length of the whole bar is the difference $L_r - L_{min}$, and the length of the light blue segment over the whole represents Ψ.

5.5 The weak recoding problem

5.5.1 Length in characters versus duration in the same language

To understand the relation between the optimality of a language in written and in oral form, we compare the values of Ψ when length is measured in duration against its values when length is measured in characters (Figure 6). This comparison is possible only in the CV collection. In Figure 6, we also show the outcome of fitting a linear model. Both when considering length in median and in mean duration, we find that the scores are generally preserved as supported by a significant and positive Pearson correlation over 0.74, and a linear regression that yields a slope near 1, as well as a small positive intercept. Interestingly, the great majority of languages show more optimization in oral than in written form (dots above the purple identity line). The intensity of this phenomenon seems to be related with sample size, in the sense that the languages with the smallest samples tend to be below the identity line. Recall, however, that we found no significant correlation between Ψ and T (Section 2.6).

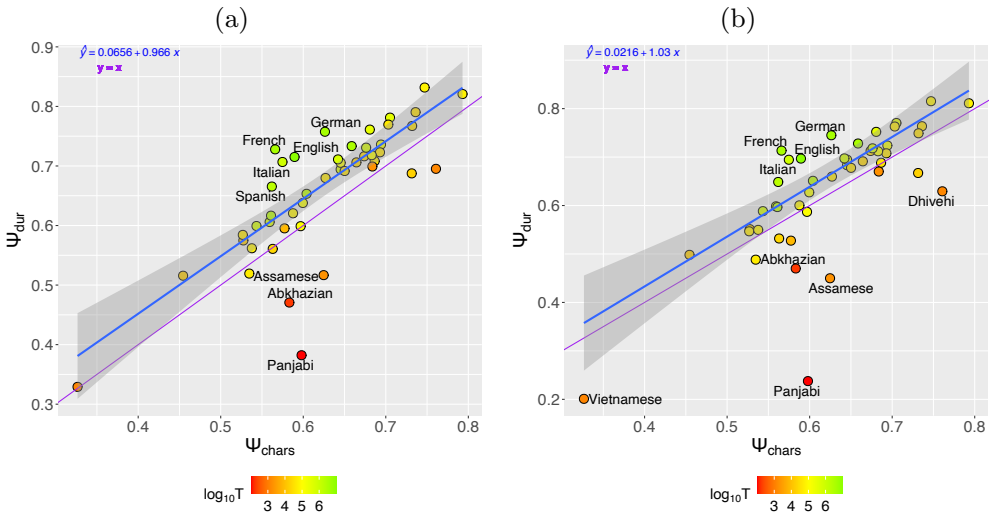


Figure 6: Ψ measured in duration (Ψ_{dur}) versus Ψ measured in characters (Ψ_{chars}) in the CV collection. Languages are color-coded by number of tokens in logarithmic scale to indicate orders of magnitude in sample size. We fit a robust linear model by Theil-Sen regression (blue line), which is then compared to the identity function, $\Psi_{dur} = \Psi_{chars}$ (purple line). 95% confidence intervals for the regression line are shown as a gray band. The choice of robust linear regression is motivated by the presence of undersampled languages whose scores deviate from the remainder of languages in this collection. For written language, only immediate word constituents are used (Japanese and Chinese are not included in CV). We report the parameters of the linear model (slope and intercept), Pearson correlation (r) and the standard error of the regression (S). (a) Word length defined as median duration. $y = 0.066 + 0.966x$, $r = 0.794$ with p-value 4.5×10^{-11} and $S = 0.227$. (b) Word length defined as mean duration. $y = 0.022 + 1.03x$, $r = 0.745$ with p-value 3.0×10^{-9} and $S = 0.239$.

5.5.2 Length in Chinese/Japanese characters versus other discrete lengths (Latin script characters and strokes)

Here we analyze the impact of replacing word length in Chinese and Japanese characters, the immediate word constituents in written form, by word length in strokes or Latin script characters after romanization. We find that the value of Ψ reduces in all cases (Figure 5). However, the value of the optimality score is still significantly large according to the corresponding correlation test, even after the replacement (Petrini et al., 2023, Figure 1 (a, c)). In particular, romanization affects both languages in a comparable way (scores decrease by 15% in Chinese, and 13% in Japanese), while measuring length in strokes has a stronger effect in the case of Chinese (18% decrease) compared to Japanese (9% decrease).

5.5.3 Removal of vowels in Latin script languages or romanizations

We also investigate the impact of removing vowels on the 15 languages using Latin scripts (including Chinese Pinyin and Japanese Romaji) in the PUD collection, comparing the original value of the score against its new value after this particular transformation. Figure 7 shows the new scores as a function of the original ones, alongside the results of a linear regression. For both Ψ and Ω, the best fit of the linear model gives a slope very close to 1, and a small offset close to 0. In contrast, η yields a slope of 1.4 and an intercept of –0.4. Moreover, the standard error S of its linear regression is more than twice the values found for Ψ and Ω. Undoubtedly, η is not robust to the removal of vowels. Among all the scores, Ω has

the highest Pearson correlation and the lowest standard error, suggesting it is the score that is the least sensitive to vowel removal.

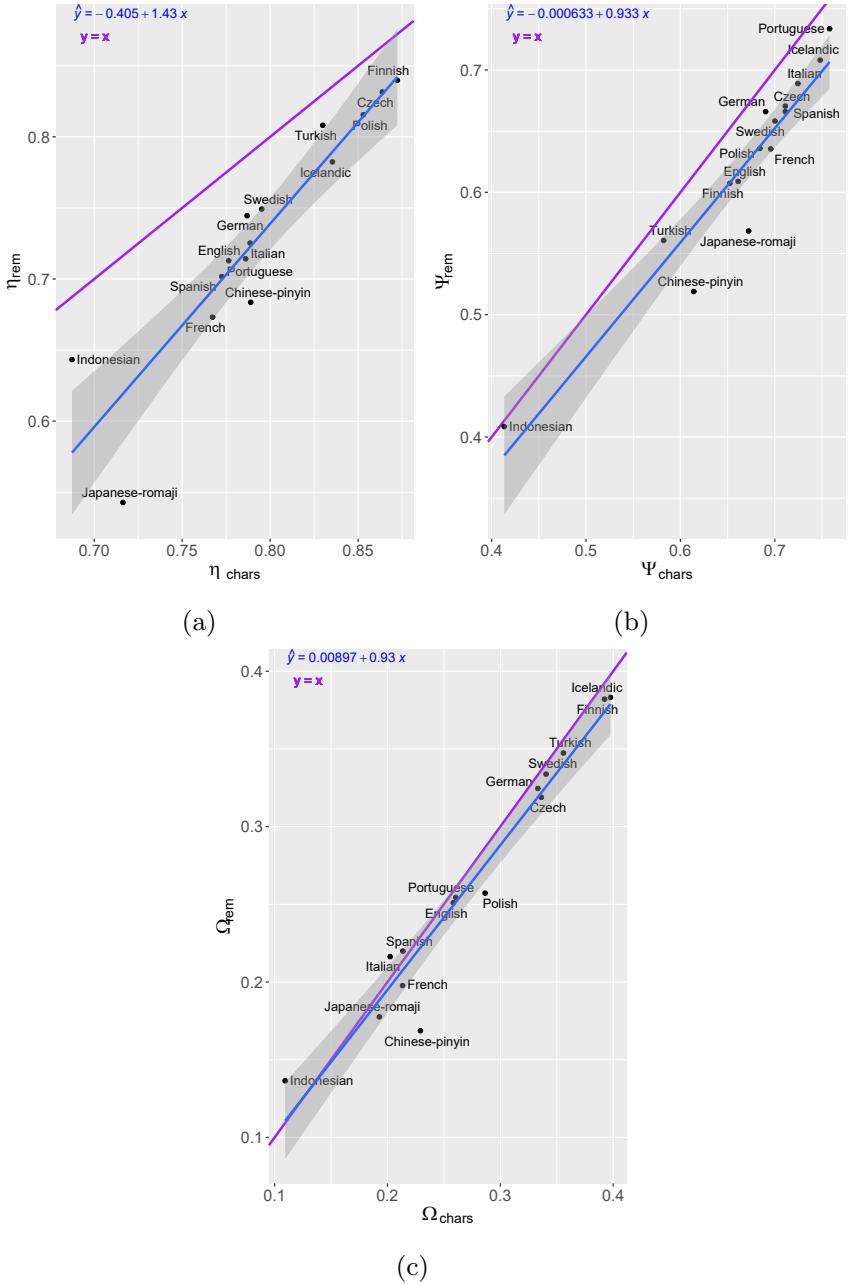


Figure 7: The value of a score after removing vowels, indicated with the subindex *rem* (e.g. Ψ_{rem}), as a function of the original value with all characters (e.g. Ψ_{chars}) in languages using the Latin script in the PUD collection. For a given language, the new value of the score is obtained after removing vowels that may include different accents or tones. The purple line is the identity function (slope of 1 and zero intercept), while the blue line is the least squares linear regression. 95% confidence intervals for the regression line are shown as a gray band. For each score, we indicate the parameters of the best fit of a linear model (slope and intercept), as well as the Pearson correlation coefficient (r) and the standard error of the regression (S). (a) η score. $y = -0.405 + 1.43x$, $r = 0.919$ with p-value 1.3×10^{-6} and $S = 0.170$. (b) Ψ score. $y = -0.001 + 0.93x$, $r = 0.972$ with p-value 4.9×10^{-8} and $S = 0.083$. (c) Ω score. $y = 0.009 + 0.93x$, $r = 0.952$ with p-value 1.4×10^{-9} and $S = 0.062$.

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

All results presented in this section have been obtained after applying a specifically developed method to filter out highly unusual characters and words – as described by Petrini et al. (2023). If the filter is disabled, we obtain some slight changes in the values, but the qualitative results remain the same.

6 Discussion

The degree to which languages are optimized is receiving increasing attention (Coupé et al., 2019; Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022; Ferrer-i-Cancho and Bentz, 2018; Koplenig, 2021). Quite generally, linguistic research is subject to a tension between acknowledging the astonishing uniformity of languages at higher levels of abstraction, and the empirical diversity which languages exhibit in a myriad of structural features from phonology to syntax. In the domain of coding efficiency, it has been argued that distinct languages exhibit similar degrees of efficiency (Coupé et al., 2019). In this article, we contribute to each of these complementary views. Questions 1–2 are related to observations of homogeneity, whereas Question 3 and Question 4 shed some light on diversity.

Question 1. What are the best optimality scores for cross-linguistic research?

We measured optimality with three scores: η, Ψ, and Ω. As shown in Table A1, the two latter scores are endowed with better statistical and mathematical properties, and are thus more reliable than η. Concerning the comparison between the remaining two, we argue that Ψ is endowed with some additional desirable properties which make it better suited for the scope of cross-linguist research.

Desirable properties

First, an ideal score should not be sensitive to the basic language parameters that are assumed constant in the definition of the baselines, such as alphabet size, vocabulary size, and sample size. To check whether this requirement holds for the considered scores, we computed the Kendall τ (and Pearson r) correlation between them and the mentioned basic parameters (Figure E1). We want to stress that PUD is the only parallel corpus in the analysis, and the correlations computed in CV might suffer from confounding effects.

On the one hand, no significant linear or non-linear association could be detected for Ψ – a result consistent across collections and length definitions. On the other hand, both η and Ω show some significant negative correlations with observed alphabet size and sample size in the CV collection. These two properties are, however, also strongly associated with one another. Indeed, the observed alphabet size depends on the amount of seen data (Figure E1 (b,c,e,f)), and the correlation of η and Ω with A could potentially just be a side-effect of the relationships above, by transitivity. Nevertheless, the most critical correlation is the one observed with sample size T, as it implies that these two scores tend to give larger values in

under-sampled languages, which we do not consider a desirable property for a score. This can also be observed when looking at the convergence speed of the scores. As shown in Figure E2 for PUD, Figure E3 for CV when length is measured in characters, and Figure E4 for CV when length is measured in duration, both η and Ω keep their decreasing trend even for large sample sizes. In contrast, Ψ generally varies within a smaller range, and importantly, it seems to reach a quite stable value in many languages, especially in CV where larger sample sizes are available. While the convergence analysis allows one to capture the evolution of the scores for different sample sizes within a language, it also gives an understanding of the behaviour of scores across languages with different sample sizes. In fact, since η and Ω tend to decrease, they will be likely to have higher values in languages with a smaller sample, thus leading to potentially misleading conclusions in non-parallel data, or data with excessive heterogeneity in terms of sample size. Moreover, Ψ and Ω have a very similar shape up to around 10^2 tokens, after which the latter generally drops down. This highlights the importance of using parallel corpora, as well as large enough samples.

Of course, the Ψ scores obtained on non-parallel data are still the reflection of optimization of a language represented by a certain doculect, rather than of the language in its entirety. However, this score is less likely to be influenced by the heterogeneity in terms of text sizes than Ω and η. For this reason, despite its relation with the Pearson correlation – which poses a potential challenge on the ability of Ψ to grasp a non linear relation between length and frequency – we argue that this score is more suited for non-parallel data. Besides, Ω could theoretically have more power in detecting a non-linear association, but its observed relationship with the number of tokens in CV (characters) raises serious concerns about its reliability in non-parallel data. Moreover, the steep decreasing trend observed in the convergence analysis suggests that, even when dealing with texts of a comparable size, Ω will be more strictly related to text size.

Complementary arguments about the best score in the context of the recoding problems are given in Question 5.

Possible alternatives

In our derivation of the new optimality scores for word lengths, namely Ψ and Ω, we followed the template to design a new score for dependency distances (Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022). Equivalent scores may be obtained by other means. We cannot exclude the possibility that there are simpler scores or existing scores that have the same statistical properties as our Ψ score for word lengths.

A promising candidate for future research is the γ index, a variant of Kendall τ correlation (Goodman and Kruskal, 1963), whose estimator is

$$\gamma = \frac{n_c - n_d}{n_c + n_d}.$$

The inverted sign γ, i.e. $-\gamma$, seems to have the same mathematical properties as Ω but may satisfy the desirable properties of Ψ. γ is able to achieve -1 when the number of concordant pairs is $n_c = 0$ whereas τ only achieves that when ties are missing (Conover, 1999).

Question 2. Is compression a universal principle? – Or, are there languages showing no optimization at all?

Despite the different definitions and properties, each of the three considered optimality scores reaches 1 when a language is fully optimized, while a value of 0 when there is no optimization at all is only expected for Ψ and Ω. The distribution of the values of the scores (Figure 4) and, with greater detail, the summary statistics (Table 5), indicate that Ψ and Ω always take values that are larger than 0, with magnitudes depending on how optimization is measured (the score and the units of length). This finding indicates that, globally, languages are shaped by compression (it is unlikely that all languages yield a positive value just by chance). However, to ensure that compression has shaped each of the languages, a test of the significance of the score is required. We will focus on Ψ given our discussion above on the best score for cross-linguistic research. We have shown that testing if Ψ is significantly high is equivalent to testing the law of abbreviation using a one-sided Pearson correlation test, whose outcomes have already been shown (Petrini et al., 2023, Figure 1 (c, d)).

All languages in the sample show some significant degree of optimization in written form, independently of the units chosen to measure the latter. The test of Ψ yields significance at the 99% confidence level for all the languages in PUD. Concerning oral form (duration), the only exception is Panjabi in the CV collection, for which we could not find a significantly small correlation. Thus its Ψ score is not significantly large despite being greater than 0. This is likely related to severe under-sampling, as this particular sample contains only 98 tokens. Indeed, the only languages which are significant at the 95% (rather than 99%) confidence level are Abkhazian, Dhivehi, and Vietnamese, all having very small samples compared to the other languages of the CV collection. Therefore, we conclude that optimization is a universal principle, that manifests in our ensemble of languages when a large enough sample of a given language is available.

Question 3. What is the degree of optimization of languages?

It has been argued that language production is not optimal but "good enough" (Goldberg and Ferreira, 2022). A necessary condition for the frequency-length association to be "good enough" is that the degree of optimality is statistically significant, which is the case (Petrini et al., 2023, Figure 1 (c, d)). A further step is quantifying how "good" the mapping is, an issue addressed by inspecting the absolute value of Ψ in a scale from 0 to 100% as its values are significantly large.

By looking at the summary statistics in Table 5 and at the probability density plots in Figure 4, we can observe how the distribution of Ψ is similar across the three scenarios (PUD, CV considering characters, CV considering duration). Indeed, their median, mean, and standard deviation values are comparable across collections and definitions of length. In particular, the mean value of Ψ in parallel data is 67%, while for CV it ranges between 62% (with length measured in characters) and 66% (with length measured in duration). Moreover, half of the languages in PUD show an optimization score larger than 70%, while this value changes slightly to 69% in CV when length is measured in duration, and drops to 63% when length is measured in characters.

This robustness suggests that the values of Ψ that we observe are a reasonable approximation of the real degree of optimization of natural languages, at least within the scope of the families and scripts considered in this particular study.

In sum, the degree of optimization of word lengths in languages ranges between 60% and 70% for the majority of the languages, both in parallel and non-parallel data. These findings are in line with the claim that languages exhibit a similar degree of coding efficiency in spite of their diversity (Coupé et al., 2019). Interestingly, our measurements of word length are on a scale from 0 to 100% as a result of dual normalization, showing a high concentration of values within a narrow range.

Finally, although we have excluded Ω to inspect the degree of optimization of languages for the sake of simplicity, we would like to make a point about the possible origin of the low values (compared to Ψ) that Ω exhibits. We are certain that they are not due to the choice of τ as opposed to Spearman ρ correlation in the definition of Ω (Equation 4). See Section D.2 for further details and a further discussion.

Question 4. What are the most and the least optimized languages in our sample?

By means of Ψ, we are able to quantify the level of optimization of languages taking advantage of the fact that it is not correlated with the basic parameters (A, n, T) of each language. However, a truly "fair" comparison of the values of Ψ in languages is hard to establish even given parallel texts. For one thing, text sizes in PUD are not large enough for Ψ to reach convergence in all cases, meaning that the scores computed on our sample are likely not fully representative of the real ones (Figure E2). This still has to be kept in mind when comparing languages by optimization degree even knowing that Ψ decays slowly once the text sample is sufficiently large. Besides, large samples are available for CV but parallelism is lacking and sample sizes are very heterogeneous, making cross-linguistic comparisons problematic. For these reasons, here we focus on PUD and make emphasis on the relative order of the languages when sorted by Ψ rather than on the concrete values, assuming that the ranks of languages by Ψ are more stable than their absolute value of Ψ when sample size is increased.

Figure 5 shows the ranking of languages in PUD obtained by sorting their Ψ values in decreasing order. Note that not every difference in the scores of two languages is necessarily significant. Future research should address the problem of the statistical significance of the differences between the optimality scores of languages (Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022).

Having said this, the languages vary in optimization between c. 41% in Indonesian and c. 78% in Japanese according to immediate word constituents (Figure 5)[12]. Japanese is the language showing the highest degree of optimization but only when word lengths are measured in characters (the immediate constituents of its written word forms), with $\Psi = 77.6\%$. The writing systems used in Japanese, Chinese, and Korean clearly lead to lower values of L_{min}, L, and L_r (when length is measured in characters) compared to other scripts, and are contained in a rather small interval (Figure 5 (b)). However, these three languages have a very different position in the ranking, meaning that the reduced variability in word length is not *a priori* an advantage or a disadvantage. Together with Japanese, at the top of the ranking we find Chinese and a cluster of Romance languages: Portuguese, Italian, Spanish, and French. Romance languages show a degree of optimality that ranges between $71 - 75\%$ (above English, that has 66%) and also show a very similar score composition (Figure 5 (b)).

To get a better impression of the linguistic factors governing higher or lower optimality scores, we give examples of the most frequent and least frequent words of Japanese (high optimality), English (middle optimality), and Indonesian (low optimality) in Table 6. Given relatively constant content (parallel texts), Japanese has many high frequency particles of minimum length (length one in UTF-8 characters), and relatively few long words of length up to ten. In fact, all the longest words in this example are actually loanwords from German or English transliterated into Japanese characters. Note, though, that these loanwords do not bias the results, since they are likewise represented in all the other texts – due to their parallel nature. In comparison, Indonesian uses particles of lower frequency[13] and generally higher lengths (both compared to original and romanized Japanese), and it has infrequent words of higher lengths reaching up to 20 characters. The latter is certainly related to the productive usage of prefixes and suffixes which are counted towards word length. It is possible that Indonesian and Japanese would fall closer together in the optimality range if Indonesian was written with a syllabary, and if Indonesian pre- and suffixes were interpreted as separate particles.

Also, note that Japanese Kanji are borrowed from traditional Chinese Hanzi, but in Japanese, Hiragana and Katakana characters are added to the repertoire. The finding that written Japanese is more optimized

[12]The range is based on immediate constituents of characters, this is why we are neglecting the minimum Ψ achieved by Chinese characters in strokes, that is an even lower value than that of Indonesian.

[13]We here give raw frequency counts. The overall number of tokens in Japanese is c. 25 000, and in Indonesian c. 17 000. So the relative frequency of the most frequent particle in Japanese would be $\frac{1753}{25000} = 0.07$, and for Indonesian $\frac{541}{17000} = 0.03$. Hence, the observation that Indonesian has lower frequency particles also holds for *relative frequencies*.

Table 6: Most frequent words and longest hapax legomena in Japanase, English and Indonesian texts of the PUD corpus. The Japanese hapax legomena can be interpreted as follows: パプアニュ一ギニア ('Papua New Guinea'); フォルクスワ一ゲン ('Volkswagen'); ディジボ一デンベルク ('Disibodenberg'); オ一バ一マルスベルク ('Obermarsberg'); エンタ一テインメント ('entertainment'). These interpretations are based on the Japanese-English dictionary at https://jisho.org/. The Indonesian hapax legomena can be interpreted as follows: men-dokument-asi-kan-nya: *men-* is a verbal prefix for active voice, *dokument* is a loanword, i.e. 'document', *-asi* verbal suffix especially for loanwords, *-kan* verb suffix (active/passive voice), *-nya* suffix with various functions, here likely pronominal, i.e. 'document this/it'. These interpretations are based on the Indonesian-English dictionary at https://dictionary.cambridge.org/dictionary/indonesian-english/ as well as the Indonesian reference grammar by Sneddon (2010).

Japanese	f_i	l_i	**English**	f_i	l_i	**Indonesian**	f_i	l_i
の *no* (poss. particle 'of')	1753	1	the	1441	3	yang (rel. pronoun 'that')	541	4
に *ni* (loc. particle 'at')	1162	1	of	620	2	dan (conjunction 'and')	447	3
は *wa* (focus particle)	1157	1	in	510	2	di (adposition 'at')	422	2
た *ta* ('did/do')	918	1	to	481	2	nya (pronom. suffix)	381	3
を *o* (direct object particle)	851	1	and	456	3	untuk (adposition 'for')	221	5
...	...	...	...	...	...	...	...	...
パプアニュ一ギニア	3	9	conservationists	2	16	mendokumentasikannya	1	20
フォルクスワ一ゲン	2	9	catastrophically	1	16	mendokumentasikan	1	17
ディジボ一デンベルク	1	10	hydroelectricity	1	16	mendemonstrasikan	1	17
オ一バ一マルスベルク	1	10	technologically	1	15	mempertemukannya	1	16
エンタ一テインメント	1	10	indistinguishable	1	17	menggulingkannya	1	16

than written Chinese when using the same unit (characters, romanizations and strokes) suggests that syllabaries along with Japanese Kanji result in a higher degree of optimization of written language than the single use of Hanzi in Chinese, according to our metrics. However, this issue deserves further investigation as there is a hidden parameter that we are not controlling for: the level of granularity in the definition of a word in Chinese and Japanese (by level of granularity we mean the criteria to segment a sequence of characters into words). The current differences in optimality may reduce if the same level of granularity was used.

Question 5. What is the optimality of languages after recoding?

We investigated the effect of recoding on optimization by assessing the degree to which ranks of scores are preserved when the respective recoding transformation is applied. In particular, we addressed the issue from three different viewpoints.

Length in characters versus duration in the same language

By means of the CV corpus, we were able to explore the effect on optimality of measuring a word length in duration rather than in characters. Most languages are above the identity line, indicating that they show more optimization in word durations (Figure 6). Undersampled languages deviate from that trend. Especially when median duration is used (Figure 6 (a)) the slope is very close to 1 (0.966), and the intercept is small (0.066), meaning that the scores are simply shifted up by a small amount.

However, there is a positive relation between the number of tokens and the extent to which a language is more optimized in oral rather than written form, as measured by the ratio Ψ_{dur}/Ψ_{chars}. Notice that not

only the undersampled languages are below the diagonal, but also the larger the sample the higher the ratio Ψ_{dur}/Ψ_{chars} (the Pearson correlation between T and the ratio Ψ_{dur}/Ψ_{chars} is $r = 0.389$; p-value 0.008). Therefore, the larger the sample, the stronger the evidence for word durations being more optimized than word lengths. The opposite finding in certain languages (lower optimality for word durations) is likely due to undersampling.

These considerations notwithstanding, it has not escaped our attention that French, a language whose alphabetic writing system is commonly observed to deviate from actual phonetic form, exhibits one of the highest contrasts between optimality in written form and optimality in oral form, with the latter being considerably higher. However, we cannot exclude that this finding is partly driven by the larger sample size of French with respect to other Romance languages (Table 4).

Length in Chinese/Japanese characters versus other discrete lengths (Latin script characters and strokes)

We also recoded Chinese and Japanese using strokes and romanizations – instead of the original characters of the main analyses. This way we investigate the effect of measuring lengths in different units. It is tempting to jump to the conclusion that word lengths have to be more optimized given Chinese/Japanese characters than given romanizations or strokes. Namely, there is an obvious increase in plain word lengths for the latter. However, this line of thinking sweeps under the carpet that, after recoding, the random and minimum baselines have also changed. For this reason, the results given optimality scores built upon these baselines are hard to intuit *a priori.*

Having said this, we indeed find a drop in the optimality scores for both alternative discrete lengths in both languages (Figure 5), yet the degree of optimization is still significant (Petrini et al., 2023, Figure 1 (c)). While romanization leads to a comparable effect in decreasing the scores in the two languages (the optimality of Japanese drops from 78% to 67% and that of Chinese drops from 72% to 61%), the impact of using strokes as a unit leads to a larger effect in Chinese (Japanese drops from 78% to 71% while Chinese drops from 72% to 60%), as shown in Figure 5. This is likely related to Japanese Hiragana and Katakana characters being made up of fewer strokes than Chinese Hanzi characters.

Overall, these results suggest that compression operates on characters rather than on strokes. Characters (or components within characters) seem to form units within which the number of strokes is less relevant for efficiency considerations. Finally, notice that the Latin-like scripts that are commonly used to type characters on digital devices (cell phones or computers) in Chinese or Japanese, do not reach the same level of optimality as the original characters. A direct translation of Chinese/Japanese characters into a Latin alphabet plus some additional diacritics results in a reduction of the optimality of these languages in written form.

Of course, we should not forget that our optimality scores ultimately reflect the link between frequency and length of words. In some writing systems this link is stronger, in others it is weaker. How exactly this relates to learning how to read and write in these systems is a separate matter. For example, in the case of writing characters, bear in mind that Chinese and Japanese speakers cannot, in general, type a single Hanzi or Kanji character by touching just one key. For that reason, in the context of writing, the comparison would be fair if they were always able to type a single character by touching just one key. All these considerations together, suggest that the pathways Western and Eastern writing systems took are not arbitrary, but (to some extent) guided by coding efficiency at the level of immediate constituents of words.

Removal of vowels in Latin script languages or romanizations

We finally consider a scenario in which recoding implies applying a decreasing transformation to word lengths, namely removing vowels. Given the invariance properties of Ψ and Ω (that η lacks; Table A1), we would expect them to yield approximately the same values for the original and the new scores computed by excluding vowels (while η be less robust to that transformation). Consistently, the linear regression slopes for both Ψ and Ω are close to 1, while that of η is about 1.4. Furthermore, η has both the largest standard regression error and the largest intercept (in absolute values).

All the scores experience a decrease after the removal of vowels. Note that this is counterintuitive. Removal of vowels generally shortens words, i.e. L. We might jump to the conclusion that a writing system without vowels must be more optimal, but this is not the case.

Importantly, Ψ and Ω show greater stability, but the latter seems to be the most capable of preserving the scores under weak recoding, having the lowest standard error, the highest Pearson correlation, and the smallest intercept. This is reasonable, given that Ω is endowed with invariance also under non-linear transformations, and there is no reason to assume that vowels would reduce word lengths in a linear way. Indeed, while on average a syllable contains one vowel, syllables can have different lengths, and thus be affected in different ways by the vowels' removal. In conclusion, Ψ and Ω are both robust in the face of vowel removal, with Ω being most robust in this respect.

An important conclusion is that we have demonstrated the existence of scores which abstract away from certain fundamental characteristics in the design of a writing system, e.g. how vowels are coded (abjads as in Standard Arabic, where only consonants are coded, or the Latin script in Romance languages, which codes for both consonants and vowels). Crucially, the choice of the optimality score can be guided by the writing characteristics which we want to abstract from. In the larger picture, this helps with resolving the customary tension between seeing languages as uniform, or seeing each language as unique, a tension which characterizes language research.

Question 6. Why do languages deviate from optimality?

We found that word lengths in languages are not 100% optimal. In previous research on the optimality of word lengths, we have argued that perceptibility and distinguishability factors as well as phonotactic constraints are likely to prevent languages from reaching theoretical optimality (Ferrer-i-Cancho, Bentz, and Seguin, 2022). Along these lines, it has been demonstrated that the gap between actual average word length and the optimal one vanishes after controlling for morphological and graphotactical constraints (Pimentel et al., 2021). However, while these constraints might, for words in isolation, to a certain extent explain the actual level of compression, it is yet to be understood why languages with shorter syntactic dependencies between words also tend to have shorter words (Ferrer-i-Cancho and Gómez-Rodríguez, 2021b). This suggests that the online memory pressures leading to syntactic dependency distance minimization are also responsible for compression of word lengths. Thus, compression seems to stem from cognitive constraints beyond word units.

In addition to these factors acting on individuals (a speaker or a listener), the nature of the cultural process by which languages are optimized may also divert them from full optimality (Kanwal et al., 2017). Languages are distributed and conventional communication systems. Reaching full optimality would imply coordinated changes in a system that may not be possible in a real community of speakers.

In this article, we have considered a minimum baseline that preserves the strings that are already being used by a language and simply attempts to reorder their corresponding lengths so as to minimize average word length. Once a communication system has been established, such reordering is a serious challenge in a conventional system like language, as this implies changing the meaning of the strings. Languages can only increase their optimality gradually. Two speakers can change locally their conventions to increase their optimality but the challenge is to propagate those changes over the entire population. Among the gradual changes, some look easier than others for a cognitive system. Optimizing a system implies changing the lengths of the strings that are being assigned to each type. The lower the resemblance between the new string and the old string, the higher the cognitive effort. Thus language users will have difficulties to interpret a string with its new meaning or to change their mental mapping of meanings to strings. We actually see every-day examples of word length reduction of the cognitively easy kind: acronyms and abbreviations for frequently used expressions are used to communicate more efficiently (e.g. 'org.' to say 'organization', 'asap' to say 'as soon as possible'). That is the case of certain areas of specialization (e.g. 'CPU' to say 'central processing unit' in Computer Science, 'ACL' to say 'anterior cruciate ligament' in Anatomy), communication on social media (e.g. 'idk' to say 'I don't know', 'brb' to say 'be right back'), or when taking personal notes.

However, there is also preliminary evidence for the opposite trend: word lengths seem to have been increasing over historical time in Chinese and Arabic (Chen et al., 2015; Milička, 2018). This is relevant to our study for three reasons. First, it indicates a potential diachronic pattern of word lengths at odds with the law of abbreviation, which is a synchronic observation. Second, it suggests that the contribution to the cost of communication stemming from word lengths has been increasing over time in these languages. Third, it highlights the necessity for scores which are comparable over time. In fact, we have defined above word length scores which are less affected by the scarcity of texts samples. These are hence amenable to diachronic studies where textual material from earlier time periods is often sparse.

In sum, all the arguments above raise the question if full optimality could be reached just by gradual, local, distributed and cognitively easy changes. We thus have to ponder if reaching full optimality is actually necessary for a language. The actual degree of optimality of word lengths in languages (more than 50% on scale from 0 to 100% according to Ψ) may just be "good enough" in language production (Goldberg and Ferreira, 2022).

Future research

In this article, we have introduced two new scores that require further theoretical and empirical research. Although we have focused on Ψ for the sake of simplicity, further attention to Ω and its variants is necessary. We have made a contribution to the problem of the best score for cross-linguistic research, but the problem should be the subject of further research.

For simplicity, we have investigated only the weak recoding problem. The next step is initiating research on the strong recoding problem, which may provide further arguments for the choice of the best score.

We have approached the degree of optimality of word lengths only synchronically. However, we have established some foundations for research on the evolution of communication. In particular, we hope that Ψ contributes to revise the finding that word lengths have been increasing over centuries in Chinese and Arabic based on L, namely simple average word lengths (Chen et al., 2015; Milička, 2018), and extending the analysis to more languages. At present, these findings suggest that languages may have been evolving against the principle of compression.

Finally, our investigation of the degree of optimality of word durations has paved the way for exploring the degree of optimality of the durations of vocalizations or gestures of other species (Semple et al., 2022).

Data accessibility

Data and code for this research work are stored in GitHub:
https://github.com/IQL-course/IQL-Research-Project-21-22.

For a permanent version of the software used in the paper see:
https://doi.org/10.5281/zenodo.18890298.

Authors' contributions

CRediT (Contributor Roles Taxonomy) contributions for each author are as follows.

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

Acknowledgements

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 also thank J. Moreno-Fernández for contributing to early developments of this research program (Moreno Fernández, 2021). 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. Finally, we thank an anonymous reviewer, A. Goldberg and L. Debowski for valuable feedback. RFC was supported by a recognition 2021SGR-Cat (01266 LQMC) from AGAUR (Generalitat de Catalunya).

RFC and CB were funded by the grant PID2024-155946NB-I00 funded by Ministerio de Ciencia, Innovación y Universidades (MICIU), Agencia Estatal de Investigación (AEI/10.13039/501100011033) and the European Social Fund Plus (ESF+).

CB was funded by the *Deutsche Forschungsgemeinschaft* (FOR 2237: Words, Bones, Genes, Tools - Tracking Linguistic, Cultural and Biological Trajectories of the Human Past), the *Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung* (Non-randomness in Morphological Diversity: A Computational Approach Based on Multilingual Corpora, 176305), and 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. The sponsors had no role in study design; collection, analysis, and interpretation of data; writing of the paper; and decision to submit it for publication.

Abbreviations

CV Common Voice Forced Alignements

NS Non-singular coding

PUD Parallel Universal Dependencies

RO Rank Ordering

UD Uniquely decodability

References

Akiyama, N., Akiyama, C. (2002). *Japanese grammar*. Barron's.

Bentz, C., Alikaniotis, D., Cysouw, M., Ferrer-i-Cancho, R. (2017). The entropy of words – learnability and expressivity across more than 1000 languages. *Entropy*, 19(6). https://doi.org/10.3390/e19060275

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

Borda, M. (2011). *Fundamentals in information theory and coding*. Springer.

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

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

Coupé, C., Oh, Y. M., Dediu, D., Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9), eaaw2594. https://doi.org/10.1126/sciadv.aaw2594

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

Daniels, P. T. (1990). Fundamentals of grammatology. *Journal of the American Oriental Society*, 110(4), pp. 727–731. https://doi.org/10.2307/602899

DeGroot, M. H., Schervish, M. J. (2002). *Probability and statistics* (3rd). Wiley.

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

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, 70, p. 056135. https://doi.org/10.1103/physreve.70.056135

Ferrer-i-Cancho, R. (2018). Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25(3), pp. 207–237. https://doi.org/10.1080/09296174.2017.1366095

Ferrer-i-Cancho, R., Bentz, C., Seguin, C. (2022). Optimal coding and the origins of Zipfian laws. *Journal of Quantitative Linguistics*, 29(2), pp. 165–194. https://doi.org/10.1080/09296174.2020.1778387

Ferrer-i-Cancho, R., Debowski, L., Moscoso del Prado Martín, F. (2013). Constant conditional entropy and related hypotheses. *Journal of Statistical Mechanics*, p. L07001. https://doi.org/10.1088/1742-5468/2013/07/l07001

Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2021a). Anti dependency distance minimization in short sequences. a graph theoretic approach. *Journal of Quantitative Linguistics*, 28(1), pp. 50–76. https://doi.org/10.1080/09296174.2019.1645547

Ferrer-i-Cancho, R., Gómez-Rodríguez, C., Esteban, J. L., Alemany-Puig, L. (2022). Optimality of syntactic dependency distances. *Physical Review E*, 105(1), p. 014308. https://doi.org/10.1103/physreve.105.014308

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). https://doi.org/10.1524/glot.2013.0004

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), pp. 1565–1578. https://doi.org/10.1111/cogs.12061

Ferrer-i-Cancho, R., Lusseau, D. (2009). Efficient coding in dolphin surface behavioral patterns. *Complexity*, 14(5), pp. 23–25. https://doi.org/10.1002/cplx.20266

Ferrer-i-Cancho, R., Bentz, C. (2018). The evolution of optimized language in the light of standard information theory. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, T. Verhoef (Eds.), *The evolution of language: Proceedings of the 12th international conference (EVOLANG XII)*. NCU Press. https://doi.org/10.12775/3991-1.029

Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2021b). Dependency distance mininimization predicts compression. *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pp. 45–57.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, 23, pp. 389–407. https://doi.org/10.31234/osf.io/w5m38

Goldberg, A. E., Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, 26(4), pp. 300–311. https://doi.org/10.1016/j.tics.2022.01.005

Goodman, L. A., Kruskal, W. H. (1963). Measures of association for cross classifications III: Approximate sampling theory. *Journal of the American Statistical Association*, 58(302), pp. 310–364. https://doi.org/10.1080/01621459.1963.10500850

Gujarati, D., Porter, D. (2008). *Basic econometrics*. McGraw-Hill Education.

Gulordava, K., Merlo, P. (2015). Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and ancient Greek. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 121–130.

Gulordava, K., Merlo, P. (2016). Multi-lingual dependency parsing evaluation: A large-scale analysis of word order properties using artificial data. *Transactions of the Association for Computational Linguistics*, 4, pp. 343–356. https://doi.org/10.1162/tacl_a_00103

Hawkins, J. A. (1998). Some issues in a performance theory of word order. In A. Siewierska (Ed.), *Constituent order in the languages of Europe*. Mouton de Gruyter. https://doi.org/10.1515/9783110812206.729

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

Joyce, T., Hodošček, B., Nishina, K. (2012). Orthographic representation and variation within the Japanese writing system: Some corpus-based observations. *Written Language and Literacy*, 15(2), pp. 254–278. https://doi.org/10.1075/wll.15.2.07joy

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, pp. 45–52. https://doi.org/10.1016/j.cognition.2017.05.001

Kendall, M. G. (1970). *Rank correlation methods* (4th). Griffin.

Koplenig, A. (2021). Quantifying the efficiency of written language. *Linguistics Vanguard*, 7(s3), p. 20190057. https://doi.org/doi:10.1515/lingvan-2019-0057

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

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, p. 105527. https://doi.org/https://doi.org/10.1016/j.cognition.2023.105527

Levshina, N. (2022a). *Communicative efficiency: Language structure and use*. Cambridge University Press. https://doi.org/10.1017/9781108887809

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

Liu, H., Xu, C., Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21, pp. 171–193. https://doi.org/10.1016/j.plrev.2017.03.002

Meylan, S. C., Griffiths, T. L. (2021). The challenges of large-scale, web-based language datasets: Word length and predictability evisited. *Cognitive Science*, 45(6), e12983. https://doi.org/10.1111/cogs.12983

Milička, J. (2018). Average word length from the diachronic perspective: The case of Arabic. *Linguistic Frontiers*, 1(2), pp. 81–89. https://doi.org/doi:10.2478/lf-2018-0007

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

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

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 in languages. zip's law of abbreviation revisited. *Glottometrics*, 54, pp. 58–87. https://doi.org/10.53482/2023_54_407

Petrini, S., Casas-i-Muñoz, A., Cluet-i-Martinell, J., Wang, M., Bentz, C., Ferrer i Cancho, R. (2026, March). *IQL Research Project (21-22). The optimality of word lengths.* Zenodo. https://doi.org/10.5281/zenodo.18890298

Piantadosi, S. T., Tily, H., Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), pp. 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

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

Semple, S., Hsu, M. J., Agoramoorthy, G. (2010). Efficiency of coding in macaque vocal communication. *Biology Letters*, 6, pp. 469–471. https://doi.org/10.1098/rsbl.2009.1062

Sigurd, B., Eeg-Olofsson, M., van Weijer, J. (2004). Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica*, 58(1), pp. 37–52. https://doi.org/10.1111/j.0039-3193.2004.00109.x

Sneddon, J. N. (2010). *Indonesian reference grammar*. Allen & Unwin.

Strauss, U., Grzybek, P., Altmann, G. (2007). Word length and word frequency. In P. Grzybek (Ed.), *Contributions to the science of text and language* (pp. 277–294). Springer. https://doi.org/10.1007/978-1-4020-4068-9_13

Tily, H. J. (2010). The role of processing complexity in word order variation and change [Doctoral dissertation, Stanford University] [Chapter 3: Dependency lengths].

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), p. 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), pp. 44–52. https://doi.org/10.1080/00031305.2021.2004922

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

Appendices

Appendix A The optimality scores

A.1 The design of two new optimality scores

By adapting the new score for the optimality of dependency distances (Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022) to word lengths, we obtain Equation 3. Importantly, it can be shown that Ψ is proportional to the Pearson correlation r between frequency and length (Section C.4), i.e.

$$\Psi = -ar,$$

where

$$a = \frac{(n-1)s_p s_l}{L_r - L_{min}}$$

is a constant determined by the distribution of the l_i's and the distribution of the p_i's; s_p and s_l are the standard deviations of word probability and word length, respectively. Traditionally, r is seen as a measure of linear association between two variables (van den Heuvel and Zhan (2022) and references therein). Therefore, a potential limitation of Ψ is that it may be unable to capture the actual non-linear association between frequency and length (Sigurd et al., 2004; Strauss et al., 2007).

This takes us to a second attempt to define an optimality score for word lengths, that is based on Kendall τ correlation. Applying the same template that we used for the design of Ψ, we get the score

$$\Omega = \frac{\tau_r - \tau}{\tau_r - \tau_{min}}, \tag{6}$$

where τ is the actual Kendall τ correlation between frequency and length, τ_r is the expected value of τ under our null hypothesis (Petrini et al., 2023) and τ_{min} is defined by rank ordering as for L_{min} (Section 2.3.2). Since $\tau_r = \mathbb{E}[\tau] = 0$ (van den Heuvel and Zhan, 2022), Ω becomes simply

$$\Omega = \frac{\tau}{\tau_{min}}.$$

Since $L = L_{min}$ implies $n_c = 0$ (Ferrer-i-Cancho, Bentz, and Seguin, 2022), the definition of Kendall τ (Equation 7) yields

$$\begin{aligned} \Omega &= \frac{c(n_c - n_d)}{c(0 - n_{d,min})} \\ &= \frac{n_d - n_c}{n_{d,min}}, \end{aligned}$$

where $n_{d,min}$ is the number of discordant pairs when $L = L_{min}$.

Having said that, notice that one cannot expect Ψ to be unable to capture non-linear associations. It has been shown that there are non-linear associations that are better captured by Pearson correlation, against common belief (van den Heuvel and Zhan, 2022). Therefore, Ψ and Ω should be seen, for the time being, as distinct approaches to deal with monotonic associations.

A.2 The statistical significance of an optimality score

According to the mathematical analysis in Section C.4, testing if η or Ψ are significantly large (or L is significantly small), as expected by the principle of compression, is equivalent to testing if Pearson r is significantly small by means of a left-sided correlation test. In contrast, testing if Ω is significantly large is equivalent to testing if τ is significantly small. Therefore, testing if Ω is significantly large is also equivalent to testing the law of abbreviation by means of a left-sided Kendall τ correlation test. Similarly, testing if Ψ (or η) is significantly large is also equivalent to testing the law of abbreviation by means of a left-sided Pearson r correlation test. In light of these equivalences, previous research on the law of abbreviation by means of Pearson r or Kendall τ (Bentz and Ferrer-i-Cancho, 2016; Ferrer-i-Cancho and Lusseau, 2009; Petrini et al., 2023; Semple et al., 2010) has implicitly tested if η, Ψ or Ω are significantly large.

A.3 Overview of the theoretical properties of the scores

Here we examine the theoretical properties of distinct scores, including correlation coefficients between word frequency and word length (Pearson r and Kendall τ). We begin with a quick summary. Firstly, L is the only score that is not normalized. However, while η is singly normalized (it is normalized only with respect to the minimum baseline, L_{min}), Ψ and Ω are dually normalized because they are normalized with respect to the minimum baseline (L_{min} or τ_{min}) and the random baseline (L_r or $\tau_r = 0$). η, Ψ and Ω exhibit constancy under optimal coding (namely $\Omega = \Psi = \eta = 1$ when $L = L_{min}$). Ψ and Ω, as the correlation scores, exhibit stability under a random mapping of probabilities into length – namely their expected value in that random mapping is zero – while that of η is moving (depending on certain parameters of the communication system). Hence η lacks this property. Constancy under optimality and stability under the null hypothesis together make a critical difference. The scores that lack one of them or both, i.e. L, η and the correlation scores (r and τ) are less informative about the actual distance to optimality than Ψ and Ω when comparing distinct systems, e.g. distinct languages (Ferrer-i-Cancho and Bentz, 2018) or the same language at different evolutionary stages (Chen et al., 2015). Table A1 summarizes the desirable mathematical properties of these scores.

After this quick overview, we refer the reader to Section C.2 for further details on constancy under optimal coding and expand the other mathematical properties giving further details (in parenthesis, we indicate the sections of Section C where full detail is given)

- *Stability under the null hypothesis* (Section C.3). A score exhibits stability under the null hypothesis if its expectation takes a constant value under the null hypothesis (Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022). The optimality scores Ψ and Ω as well as the correlation coefficients r and τ, are stable under the null hypothesis, namely their expectation is zero when the mapping of word probabilities into word lengths is random (Petrini et al., 2023). In contrast, η lacks that property.

- *Invariance under linear transformation* (Section C.7). A score exhibits invariance under linear transformation when its value does not change if a linear transformation is applied to the target variable (Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022). The Pearson correlation between two variables is invariant under linear transformation only if (a) the linear transformation that is applied to only one of the variables is increasing or (b) the transformations that are applied to each of the variables are both increasing or both decreasing (Section C.5); the same applies to Kendall τ correlation. For simplicity, here we focus on increasing linear transformations of word length.
  Ψ and Ω are invariant (do not change) if length is transformed linearly, namely, each length l is replaced by a linear function of it, i.e. $g(l) = al + b$, where a and b are some constants, with $a > 0$.
  In contrast, η, is not invariant under a linear transformation but invariant under a proportional change of scale, which is obtained when $b = 0$.

- *Invariance under monotonic transformation* (Section C.8). A score exhibits invariance under strictly increasing transformation if its value does not change if a transformation of this sort is applied to the target variable. The Kendall τ correlation between two variables is invariant under strictly non-linear monotonic transformation only if (a) the transformation that is applied to only one of the variables is increasing or (b) the transformations that are applied to each of the variables are both increasing or both decreasing (Section C.6). In contrast, that invariance is only warranted for Pearson correlation when the transformations are linear.
  For simplicity, here we focus on increasing non-linear transformations of word length. Ω is invariant (does not change) if each length l is replaced by a strictly increasing function of it, $h(l)$. Notice that h can be non-linear. In contrast, Ψ and η are not invariant under that transformation.

Table A1: Summary of the mathematical properties of each of the word length scores: mean word length (L), the Pearson and the Kendall correlation between frequency and length (r and τ, respectively) and the optimality scores η, Ψ and Ω. Concerning invariance, we assume for simplicity that the transformation is applied to word length only.

Properties	L	η	r	Ψ	τ	Ω
Normalization	none	single	single	dual	single	dual
Constancy under optimal coding		✓		✓		✓
Stability under the null hypothesis			✓	✓	✓	✓
Invariance under translation			✓	✓	✓	✓
Invariance under proportional change of scale		✓	✓	✓	✓	✓
Invariance under increasing linear transformation			✓	✓	✓	✓
Invariance under strictly increasing non-linear transformation					✓	✓

Appendix B The minimum baselines

To calculate η, L_{min} has been computed making two kinds of assumptions borrowed from standard information theory and its extensions (Ferrer-i-Cancho, Bentz, and Seguin, 2022; Ferrer-i-Cancho and Bentz, 2018):

- Non-singular coding (NS), namely two word types should not be assigned the same string. We use L_{min}^{NS} to refer to the specific value of L_{min} in this setting.
- Unique decodability (UD), namely the assigned strings, in addition to being non-singular, should have a unique segmentation when concatenated forming a sequence of strings without blanks or other sorts of separators. We use L_{min}^{UD} to refer to the specific value of L_{min} in this setting.

Each of these assumptions is known as coding scheme in the language of information theory (Cover and Thomas, 2006). See Ferrer-i-Cancho, Bentz, and Seguin (2022) for further details on these two ways of defining the minimum baseline.

Previous research on the optimality of word lengths has used L_{min}^{NS} and L_{min}^{UD} (Ferrer-i-Cancho and Bentz, 2018; Moreno Fernández, 2021). Here we focus on a third way of computing L_{min}, that we call rank ordering (RO) (Moreno Fernández, 2021) and that is called "Zipfian coding" in related work (Pimentel et al., 2021). We use L_{min}^{RO} to refer to this specific value of L_{min}. This method follows from the generalized theory of compression (Ferrer-i-Cancho, Bentz, and Seguin, 2022) and has already been used in previous research on the optimality of word lengths (Moreno Fernández, 2021; Pimentel et al., 2021).

L_{min}^{RO} is obtained when the current values of l_i are reassigned to frequencies (or equivalently probabilities) so as to minimize L. L is minimized when probabilities are sorted decreasingly and lengths are sorted increasingly (Ferrer-i-Cancho, Bentz, and Seguin, 2022). Then the i-th most frequent type gets the i-th shortest length, hence the name *rank ordering*.

To see this, consider the matrix with two columns, f_i and l_i, that are used to compute L (Table B1). Suppose that we shuffle the column of l_i in Table B1. We use l_i' to refer to the new value of l_i after the

Table B1: Matrix indicating the frequency and length of three types. The mean token length is $L = \frac{235}{125} = 1.88$ and the mean type length is $L_r = \frac{6}{3} = 2$.

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

Table B2: All the $3! = 6$ permutations of the column l_i in Table B1 that can be produced. Each permutation is indicated with letters from A to F. A is the one minimizing L. L', the mean length of tokens in a given permutation, 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'	$\frac{155}{125} = 1.24$	$\frac{170}{125} = 1.36$	$\frac{235}{125} = 1.88$	$\frac{265}{125} = 2.12$	$\frac{330}{125} = 2.64$	$\frac{345}{125} = 2.76$

shuffling and L' to refer to the new value of L, that is

$$L' = \frac{1}{T} \sum_{i=1}^{n} f_i l'_i.$$

Table B2 shows all the possible shufflings (permutations) of the l_i column that can be produced and the corresponding values of L' for the matrix in Table B1. Then L_r is the average value of L' over all permutations of column l_i and L_{min}^{RO} is the value of L' of the permutation that minimizes the value of L'. According to Table B2, $L_{min}^{RO} = \frac{155}{125} = 1.24$ for the matrix in Table B1. By symmetry, L_{min}^{RO} is also the value of L' of the permutation of column f_i that minimizes the value of L', where now

$$L' = \frac{1}{T} \sum_{i=1}^{n} f'_i l_i$$

and f'_i is the new value of f_i after the shuffling.

The rationale of rank ordering is that it is sensitive to imperceptibility and distinguishability factors as well as to the phonotactic constraints shaping the length of words in a language (Ferrer-i-Cancho, Bentz, and Seguin, 2022, Section 2.2). In contrast, optimal non-singular coding and optimal uniquely decodable encoding completely neglect constraints that are language specific or specific to humans (e.g. their cognition, their anatomy, etc.), and abstract away from the cultural and biological evolution processes that lead to more efficient languages (Kanwal et al., 2017). Rank ordering is conservative with respect to optimality because it assumes that a language cannot produce word lengths that are better than the ones

that are being observed. In contrast, non-singular coding and uniquely decodable coding point towards the theoretical limits of optimal coding in languages.

The Kendall τ correlation can be defined as (Kendall, 1970)

$$\tau = c(n_c - n_d), \tag{7}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs and c is a normalization factor. The previous definition of τ (Equation 7) is the template for the two correlation scores that were proposed originally by Kendall, now known as τ-a and τ-b (Kendall, 1970). τ-a is the simplest and is obtained when

$$c = \frac{1}{\binom{n}{2}}.$$

τ-b is obtained by another c the corrects for ties. It is worth mentioning that τ-b is the one it implemented in the base R statistical programming language.

τ_{min}^{RO} is the smallest value that τ can achieve by shuffling the column of the p_i's or the column of the l_i's in the matrix. τ_{min}^{RO} coincides with the ordering of the l_i's in the matrix such that L is minimized, namely, $L = L_{min}^{RO}$ (Ferrer-i-Cancho, Bentz, and Seguin, 2022). $\tau_{min}^{RO} = -1$ if there are no ties both in the p_i's and the l_i's; otherwise (the common situation), $\tau_{min}^{RO} > -1$ (Section C, Property C.1).

τ_{min}^{RO} is unaffected if the lengths are replaced by new "lengths" that result from applying a monotonically increasing function of l as the following property states. That property will be used to analyze the mathematical properties of distinct scores in Section C.

Property B.1. *Let L_{min}^{RO} be the minimum value of L under rank ordering, defined as*

$$L_{min}^{RO} = \sum_{i=1}^{n} p_i l_{i,min}.$$

Let $h(l)$ be a strictly monotonically increasing function of l. We use a prime mark, $'$, to indicate the new value of a function after applying $h(l)$ to the l_i's. Accordingly,

$$L' = \sum_{i=1}^{n} p_i h(l_i)$$

and

$$\tau' = \tau(p, h(l)).$$

Then the minimum value that L' can achieve is

$$L'^{RO}_{min} = \sum_{i=1}^{n} p_i h(l_{i,min}) \tag{8}$$

while the minimum of τ remains unaltered, i.e.

$$\tau'^{RO}_{min} = \tau^{RO}_{min}.$$

Proof. Recall that L^{RO}_{min} is computed by sorting the lengths increasingly and the probabilities decreasingly (Ferrer-i-Cancho, Bentz, and Seguin, 2022) in the matrix that is used to calculate L. Applying the same method to calculate L'^{RO}_{min}, it turns out that, if the strictly monotonic transformation is increasing, the sorting by the new lengths that are obtained after replacing each original length l by $h(l)$ will preserve the position of the lengths in the original sorting, giving Equation 8. For the same reason, $\tau'_{min} = \tau_{min}$, because the ordering of the l_i's that minimizes L (yielding L^{RO}_{min}) coincides with the ordering of the $h(l_i)$'s that minimizes L'^{RO}_{min} since $h(l)$ is strictly increasing. □

In this article, we will investigate the degree of optimality of languages focusing on rank ordering as a lower bound of the true degree of optimality of word lengths in languages. The choice of rank ordering is for the sake of simplicity and to ensure that the optimality scores are properly normalized as explained in Section C.

Appendix C Theory

C.1 Technical remarks on normalization and coherence

η was designed to be normalized, namely, $0 \leq \eta \leq 1$ (Ferrer-i-Cancho, Bentz, and Seguin, 2022). However, a theoretical challenge in the calculation of η with the two versions of L_{min} from information theory (optimal non-singular coding and optimal uniquely decodable encoding in Section 2.3.2 and Section B), namely L^{NS}_{min} and L^{NS}_{min}, is to ensure that the assumptions of each coding scheme hold. If a language satisfies them, then $\eta \leq 1$ because $L \leq L_{min}$. If not, then $\eta > 1$ may be possible. The same problem may concern Ψ or Ω depending on the choice of the minimum baseline. This is another reason why we choose a minimum baseline based on rank ordering, i.e. L^{RO}_{min}, so that the random baseline and the minimum baseline are perfectly aligned: L^{RO}_{min} is the minimum average token length over all permutations in Table B2; L_r is the average token length over all these permutations (Petrini et al., 2023). Thus, our choice of the baselines warrants that $\eta, \Psi, \Omega \leq 1$ for any communication system. Hereafter, we use L_{min} and τ_{min} to refer to L^{RO}_{min} and τ^{RO}_{min}, respectively.

C.2 Constancy under optimal coding

A score exhibits constancy under optimality if it takes a constant value in case the system under consideration is organized optimally (Ferrer-i-Cancho, Gómez-Rodríguez, et al., 2022). η, Ψ and Ω exhibit constancy under optimality, namely they take a value of 1 when word lengths are minimum and actually their value never exceeds one. In case of optimal coding, $\tau \leq 0$ (Ferrer-i-Cancho, Bentz, and Seguin, 2022). Although r and τ are bounded below by –1, they do not necessary reach –1 when lengths are minimum (Ferrer-i-Cancho, Bentz, and Seguin, 2022). Obviously, L lacks any of these properties since $L \geq L_{min}$ and L_{min} does not need to be one. In the language of mathematics:

Property C.1.

1. $\Psi, \Omega, \eta \leq 1$, *with equality if and only if* $L_{min} = L$.
2. $-1 \leq r, \tau \leq 1$, *but* $L_{min} = L$ *does not imply* $r = -1$ *or* $\tau = -1$.

Proof.

1. Since $L_{min} \leq L$ by definition,

$$\Omega = \frac{\tau}{\tau_{min}} \leq 1$$
$$\eta = \frac{L_{min}}{L} \leq 1.$$

For the same reason,

$$-L \leq -L_{min}$$

and adding L_r to both sides of the equation, one obtains

$$L_r - L \leq L_r - L_{min}.$$

Dividing by $L_r - L_{min}$ on both sides of the inequality, one gets

$$\frac{L_r - L}{L_r - L_{min}} \leq \frac{L_r - L_{min}}{L_r - L_{min}}$$

because $L_r \geq L_{min}$ and hence $\Psi \leq 1$.

Finally, notice that $\tau_{min} \leq \tau$ by definition and τ_{min} is negative (Ferrer-i-Cancho, Bentz, and Seguin, 2022). Then, dividing by τ_{min} on both sides of $\tau \geq \tau_{min}$, we obtain

$$\Omega = \frac{\tau}{\tau_{min}} \leq \frac{\tau_{min}}{\tau_{min}} = 1.$$

2. It is well-known that $-1 \leq r, \tau \leq 1$ (Conover, 1999, Section 5.4). When $L_{min} = L$, $r = -1$ is only possible if the relationship between p and l is perfectly linear in the optimal shuffling of value l_i's in the matrix; $\tau = -1$ can only be achieved if there are not ties neither within the p_i's nor withing the l_i's (Ferrer-i-Cancho, Bentz, and Seguin, 2022). Notice that, Kendall proposed two rank correlation scores, i.e. τ-a and τ-b (Kendall, 1970). τ-b, which is the one that base R implements, incorporates a correction for ties but such correction does not ensure that τ-b yields a constant value (τ-b $= -1$) when there are no concordant pairs.

□

C.3 Stability under the null hypothesis

Property C.2.

$$\mathbb{E}[\Psi] = \mathbb{E}[\Omega] = \mathbb{E}[r] = \mathbb{E}[\tau] = 0$$

while

$$\mathbb{E}[\eta] \geq \frac{L_{min}}{L_r}. \tag{9}$$

Proof. It is well-known that $\mathbb{E}[r] = \mathbb{E}[\tau] = 0$ (see van den Heuvel and Zhan (2022) and references therein). On the one hand,

$$\begin{aligned} \mathbb{E}[\Omega] &= \mathbb{E}\left[\frac{\tau}{\tau_{min}}\right] \\ &= \frac{1}{\tau_{min}}\mathbb{E}[\tau] \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[\Psi] &= \mathbb{E}\left[\frac{L_r - L}{L_r - L_{min}}\right] \\ &= \frac{1}{L_r - L_{min}}(L_r - \mathbb{E}[L]) \\ &= \frac{1}{L_r - L_{min}}(L_r - L_r) \\ &= 0. \end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}[\eta] &= \mathbb{E}\left[\frac{L_{min}}{L}\right] \\ &= L_{min}\,\mathbb{E}\left[\frac{1}{L}\right].\end{aligned}$$

$\frac{1}{L}$ is a convex function of L because L is positive by definition. Jensen's inequality gives

$$\mathbb{E}\left[\frac{1}{L}\right] \geq \frac{1}{\mathbb{E}[L]}.$$

Finally

$$\mathbb{E}[\eta] \geq \frac{L_{min}}{L_r}.$$

□

Obviously, L lacks stability under the null hypothesis since $\mathbb{E}[L] = L_r$ and that is the mean length of types.

Property C.2 only indicates a lower bound for $\mathbb{E}[\eta]$. Then the true $\mathbb{E}[\eta]$ might be stable under the null hypothesis in the sense that $\mathbb{E}[\eta] = k$, where k is some constant. The fact that $L_{min}, L_r > 0$ and Equation 9 imply $k > 0$. In contrast, $\mathbb{E}[\Psi] = \mathbb{E}[\Omega] = k$ with $k = 0$. However, in Section D.3, we demonstrate that η is not stable under the null hypothesis with real data, indeed $\mathbb{E}[\eta] \approx \frac{L_{min}}{L_r}$, and confirm that Ψ and Ω are stable under the null hypothesis while L is not.

C.4 The relationship between L, η, Ψ and Pearson correlation

First, we show the relationship between Ψ and covariance or Pearson correlation through their estimators.

Given 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 we replace the random variables x and y by p (the probability of a type) and l (the length/duration of a type). Accordingly, the sample of n points becomes $\{(p_1, l_1), ..., (p_i, l_i), ..., (p_n, l_n)\}$, one point per type. Then the covariance between p and l 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 $\bar{l} = L_r$ (Equation 5), we finally obtain

$$s_{pl} = \frac{1}{n-1}(L - L_r). \tag{10}$$

Then Ψ turns out to be proportional to the sample covariance, i.e.

$$\Psi = \frac{L_r - L}{L_r - L_{min}} = -\frac{n-1}{L_r - L_{min}} s_{pl} \tag{11}$$

but with an opposite sign.

The sample Pearson correlation is

$$r = \frac{s_{xy}}{s_x s_y}, \tag{12}$$

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

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

Combining Equation 10 and Equation 12, we find that

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

Combining Equation 11 and Equation 12, we find that Ψ is negatively proportional to the sample Pearson correlation, i.e.

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

Similarly, it is easy to see that η and L are linear functions of r or s_{pl}. For instance,

$$\eta = \frac{1}{L_{min}} \left[(n-1) s_p s_l r + L_r \right].$$

Other linear relationships can be derived similarly.

C.5 Invariance of Pearson correlation under linear transformation

The Pearson correlation between two random variables X and Y is the covariance normalized by the product of the standard deviations, i.e.

$$r(X,Y) = \frac{COV(X,Y)}{\sigma(X)\sigma(Y)}.$$

It is often stated that Pearson correlation $r(X,Y)$ between two random variables X and Y (or its estimator) is invariant under linear transformation, i.e. $r(X,Y)$ does not change if is X is replaced by $f(X) = aX+b$ and y is replaced by $g(Y) = cY + d$ (Gujarati and Porter, 2008, Question 3.11). However, the statement is not accurate enough as the following property clarifies.

Property C.3. *$r(X,Y)$ is invariant under linear transformation.*

- *In a strict or strong sense, i.e.*

  $$r(aX + b, cY + d) = r(X,Y),$$

  if and only if $\mathrm{sgn}(a) = \mathrm{sgn}(c)$ *and*

  $$r(aX + b, cY + d) = -r(X,Y),$$

  if and only if $\mathrm{sgn}(a) \neq \mathrm{sgn}(c)$

- *In a weak sense, i.e.*

  $$|r(aX + b, cY + d)| = |r(X,Y)|,$$

  in any circumstance (namely independently of a, b, c and d).

Proof. Knowing that (DeGroot and Schervish, 2002, Sections 4.3 and 4.6)

$$\begin{aligned} COV(aX+b, cY+d) &= acCOV(X,Y) \\ \sigma(aX+b) &= |a|\sigma(X) \\ \sigma(cY+d) &= |c|\sigma(Y), \end{aligned}$$

it follows that

$$r(aX+b, cY+d) = \frac{ac}{|a||c|} r(X,Y).$$

Hence the statements of this property. □

C.6 Invariance of Kendall τ correlation under strictly monotonic transformation

The Kendall τ correlation between to random variables X and Y can be defined on the probability that two pairs of values (X_1, Y_1) and (X_2, Y_2) are concordant, i.e.

$$Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\},$$

and on the probability that they are discordant, i.e.

$$Pr\{(X_1 - X_2)(Y_1 - Y_2) < 0\},$$

as (van den Heuvel and Zhan, 2022)

$$\begin{aligned} \tau &= \tau(X,Y) \\ &= Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - Pr\{(X_1 - X_2)(Y_1 - Y_2) < 0\}. \end{aligned}$$

This definition corresponds to Kendall τ-a (Kendall, 1970).

Kendall τ correlation $r(X,Y)$ is invariant under non-linear transformation but under certain conditions. If X is replaced by applying some function h_x and Y is replaced by some function h_y, then τ is invariant when both $h_x(X)$ and $h_y(Y)$ are strictly monotonically increasing or both $h_x(X)$ and $h_y(Y)$ are strictly monotonically decreasing (van den Heuvel and Zhan (2022) and references therein). If only one variable is replaced, say X, then $h(X)$ must be strictly monotonically increasing. The argument is detailed in the following property and its proof.

Property C.4. *$\tau(X,Y)$ is invariant strictly monotonic linear transformation*

- *In a strict or strong sense, i.e.*

$$\tau(h_x(X), h_y(Y)) = \tau(X,Y)$$

if the trend of $h_x(X)$ and that of $h_y(Y)$ is the same (both increasing or both decreasing) and

$$\tau(h_x(X), h_y(Y)) = -\tau(X,Y)$$

if the trends differ (one is increasing and the other is decreasing) and

- *In a weak sense, i.e.*

$$|\tau(h_x(X), h_y(Y))| = |\tau(X,Y)|,$$

in any circumstance (namely independently of the trend of $h_x(X)$ and $h_y(Y)$).

Proof. Notice that

- τ is symmetric, namely, $\tau(X,Y) = \tau(Y,X)$.
- When $h_x(X)$ and $h_y(Y)$ are strictly monotonic (increasing or decreasing; the direction of $h_x(X)$ and $h_y(Y)$ does not need to be the same),

$$Pr\{(h_x(X_1) - h_x(X_2))(h_y(Y_1) - h_y(Y_2)) = 0\} = Pr\{(X_1 - X_2)(Y_1 - Y_2) = 0\}.$$

If we apply a strictly monotonically increasing (i) function h_i to one of the variables, say X, we get that $\tau(h_i(X), Y) = \tau(X,Y)$. To see it, notice that

$$Pr\{(h_i(X_1) - h_i(X_2))(Y_1 - Y_2) > 0\} = Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\}$$

and then

$$\begin{aligned} Pr\{(h_i(X_1) - h_i(X_2))(Y_1 - Y_2)) < 0\} &= 1 - Pr\{(h_i(X_1) - h_i(X_2))(Y_1 - Y_2) > 0\} \\ &\quad -Pr\{(h_i(X_1) - h_i(X_2))(Y_1 - Y_2) = 0\} \\ &= 1 - Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - Pr\{(X_1 - X_2)(Y_1 - Y_2) = 0\} \\ &= Pr\{(X_1 - X_2)(Y_1 - Y_2) < 0\}. \end{aligned}$$

By symmetry, we also have that $\tau(X, h_i(Y)) = \tau(X,Y)$.

If we apply a strictly monotonically decreasing (d) function h_d to one of the variables, say X, we get that $\tau(h_d(X), Y) = -\tau(X, Y)$. To see it, recall that

$$Pr\{(h_d(X_1) - h_d(X_2))(Y_1 - Y_2) = 0\} = Pr\{(X_1 - X_2)(Y_1 - Y_2) = 0\}$$

and notice that

$$Pr\{(h_d(X_1) - h_d(X_2))(Y_1 - Y_2) > 0\} = Pr\{(X_1 - X_2)(Y_1 - Y_2) < 0\}$$

and

$$Pr\{(h_d(X_1) - h_d(X_2))(Y_1 - Y_2) < 0\} = Pr\{(X_1 - X_2)(Y_1 - Y_2) > 0\}.$$

Hence $\tau(h_d(X), Y) = -\tau(X, Y)$. By symmetry, we also have that $\tau(X, h_d(Y)) = -\tau(X, Y)$.

Now suppose that we apply a strictly monotonic function h_x to X and a strictly monotonic function h_y to Y.

- If both functions are increasing, by iterating on the arguments above, we get

$$\tau(h_x(X), h_y(Y)) = \tau(X, h_y(Y)) = \tau(X, Y).$$

- If both functions are decreasing, by iterating on the arguments above, we get

$$\tau(h_x(X), h_y(Y)) = -\tau(X, h_y(Y)) = \tau(X, Y).$$

- If one function is decreasing and the other is increasing, say h_x is increasing and h_y is decreasing,

$$\tau(h_x(X), h_y(Y)) = \tau(X, h_y(Y)) = -\tau(X, Y).$$

  If h_x is decreasing and h_y is increasing,

$$\tau(h_x(X), h_y(Y)) = -\tau(X, Y)$$

  again by the symmetry of τ.

Hence the statements of this property. □

C.7 Invariance under linear transformation

Property C.5. *We use a prime, ′, to indicate the new value of a score after applying a linear transformation $g(l) = al + b$ to l with $a > 0$. Let η'' be the value of η after applying a proportional change of scale to l, namely $g(l)$ with $b = 0$. Then*

$$\Psi = \Psi' \quad \Omega = \Omega'$$
$$r = r' \quad \tau = \tau'$$
$$\eta = \eta''$$

while $L = L'$ and $\eta = \eta'$ do not generally hold.

Proof. r and τ are invariant under that linear transformation (Section C.5 and Section C.6), hence $r = r'$ and $\tau = \tau'$.

Knowing that

$$L = \sum_i p_i l_i,$$

it is easy to show that, after applying a linear transformation to word lengths ($g(l)$ can be increasing or decreasing), one obtains

$$L' = aL + b$$
$$L'_r = aL_r + b.$$

The condition $a > 0$ and Property B.1 give

$$L'_{min} = aL_{min} + b. \tag{13}$$

Hence

$$\begin{aligned} \Psi' &= \frac{L'_r - L'}{L'_r - L'_{min}} \\ &= \frac{aL_r + b - aL - b}{aL_r + b - aL_{min} - b} \\ &= \frac{a(L_r - L)}{a(L_r - L_{min})} \\ &= \Psi. \end{aligned}$$

In contrast,

$$\begin{aligned} \eta' &= \frac{L_{min}}{L'} \\ &= \frac{aL_{min} + b}{aL + b}. \end{aligned}$$

When $b = 0$,

$$\eta' = \eta'' = \eta.$$

Concerning Ω, recall that τ is invariant under strictly increasing transformation (Section C.6). When $a > 0$, the linear transformation is strictly increasing, hence the invariance of τ under that transformation gives

$$\tau' = \tau(p, g(l)) = \tau(p, l) = \tau.$$

Besides, $a > 0$ and Property B.1 give $\tau'_{min} = \tau_{min}$. □

C.8 Invariance under non-linear monotonic transformation

Property C.6. *We use a prime, ′, to indicate the new value of a score after applying a strictly increasing non-linear transformation $h(l)$ to l. Then*

$$\tau = \tau'$$

$$\Omega = \Omega'$$

while

$$L = L'$$

$$r = r'$$

$$\Psi = \Psi'$$

$$\eta = \eta'$$

do not generally hold.

Proof. Since $h(l)$ is strictly increasing, then τ is invariant (Section C.5 and Section C.6), i.e.

$$\tau(p, l) = \tau(p, h(l)).$$

As $h(l)$ is monotonically increasing, Property B.1 gives

$$\tau'_{min} = \tau_{min}(p, h(l)) = \tau_{min}(p, l).$$

Finally, combining the results above

$$\begin{aligned} \Omega' &= \frac{\tau(p, h(l))}{\tau_{min}(p, h(l))} \\ &= \frac{\tau(p, l)}{\tau_{min}(p, l)} \\ &= \Omega. \end{aligned}$$

$L' = L$ is trivially unwarranted. To see that $r' = r$ is not warranted, suppose that $r(p, l) = 1$, namely the association between both variables is perfectly linear and increasing. If we apply a strictly increasing non-linear transformation $h(l)$, then $r(p, h(l)) < 1$ because the association between p, $h(l)$ is not linear. Similarly, $\Psi' = \Psi$ is unwarranted. A counterexample suffices. Consider configuration C in Table B1. $L_r = 2$, $L = 1.88$ and $L_{min} = 1.14$ give $\Psi \approx 0.158$. Now suppose that we apply $h(l) = 2^l$ to C. Then $L'_r = \frac{14}{13}$, $L' = \frac{32}{25}$, and $L'_{min} = \frac{64}{75}$ give $\Psi' = 0.\overline{8}$. Hence $\Psi \neq \Psi'$. The finding is not surprising given that Ψ is a linear function of the Pearson correlation coefficient (Section C.4). η cannot be invariant under that transformation because it is not even invariant if the transformation is strictly increasing but linear with non-zero slope (Property C.6). □

Appendix D Analysis

Here we present complementary analyses, tables and plots.

D.1 Optimality scores

In Table D1, Table D2 and Table D3, we show the values of the optimality scores and their ingredients for PUD and for CV when length is measured in characters and also in duration, respectively.

Table D1: Word lengths and their optimality in PUD. Word length is measured in number of characters. Han script is used in its traditional variant.

Language	Family	Script	L_{min}	L	L_r	τ	τ_{min}	η	Ψ	Ω
Arabic	Afro-Asiatic	Arabic	3.40	4.16	5.56	-0.13	-0.72	0.82	0.65	0.18
Indonesian	Austronesian	Latin	4.03	5.89	7.18	-0.09	-0.80	0.68	0.41	0.11
Russian	Indo-European	Cyrillic	5.18	6.04	8.08	-0.19	-0.64	0.86	0.71	0.30
Hindi	Indo-European	Devanagari	3.11	4.09	5.85	-0.19	-0.78	0.76	0.64	0.24
Czech	Indo-European	Latin	4.70	5.44	7.27	-0.22	-0.67	0.86	0.71	0.34
English	Indo-European	Latin	3.77	4.86	6.99	-0.20	-0.76	0.78	0.66	0.26
French	Indo-European	Latin	3.71	4.85	7.55	-0.17	-0.76	0.77	0.70	0.22
German	Indo-European	Latin	4.55	5.79	8.55	-0.23	-0.68	0.79	0.69	0.33
Icelandic	Indo-European	Latin	4.44	5.32	7.91	-0.26	-0.66	0.84	0.75	0.40
Italian	Indo-European	Latin	3.89	4.89	7.72	-0.16	-0.76	0.80	0.74	0.22
Polish	Indo-European	Latin	5.16	6.05	7.98	-0.19	-0.64	0.85	0.68	0.29
Portuguese	Indo-European	Latin	3.68	4.68	7.50	-0.17	-0.74	0.79	0.74	0.24
Spanish	Indo-European	Latin	3.75	4.85	7.57	-0.16	-0.74	0.77	0.71	0.21
Swedish	Indo-European	Latin	4.30	5.41	7.99	-0.23	-0.68	0.80	0.70	0.34
Japanese	Japonic	Japanese	1.46	1.75	2.76	-0.22	-0.60	0.84	0.78	0.36
Japanese-strokes	Japonic	Japanese	5.15	7.70	13.91	-0.09	-0.78	0.67	0.71	0.12
Japanese-romaji	Japonic	Latin	2.68	3.84	6.02	-0.14	-0.80	0.70	0.65	0.18
Korean	Koreanic	Hangul	2.42	2.75	3.24	-0.24	-0.64	0.88	0.60	0.38
Thai	Kra-Dai	Thai	3.16	4.33	6.33	-0.23	-0.82	0.73	0.63	0.28
Chinese	Sino-Tibetan	Han (Traditional variant)	1.53	1.73	2.36	-0.26	-0.55	0.88	0.76	0.48
Chinese-strokes	Sino-Tibetan	Han (Traditional variant)	10.47	15.00	21.59	-0.18	-0.77	0.70	0.59	0.24
Chinese-pinyin	Sino-Tibetan	Latin	3.78	4.90	6.49	-0.16	-0.76	0.77	0.59	0.21
Turkish	Turkic	Latin	5.26	6.35	7.87	-0.24	-0.67	0.83	0.58	0.36
Finnish	Uralic	Latin	6.55	7.51	9.31	-0.23	-0.60	0.87	0.65	0.39

Table D2: Word lengths and their optimality in CV. Word length is measured in number of characters. '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_{min}	L	L_r	τ	τ_{min}	η	Ψ	Ω
Arabic	Afro-Asiatic	Arabic	3.26	4.10	5.06	-0.14	-0.87	0.80	0.53	0.16
Maltese	Afro-Asiatic	Latin	3.72	5.07	7.35	-0.20	-0.81	0.73	0.63	0.24
Vietnamese	Austroasiatic	Latin	2.75	3.24	3.47	-0.19	-0.73	0.85	0.33	0.26
Indonesian	Austronesian	Latin	3.70	5.37	7.24	-0.20	-0.91	0.69	0.53	0.22
Esperanto	Conlang	Latin	3.23	4.83	7.73	-0.18	-0.92	0.67	0.65	0.19
Interlingua	Conlang	Latin	3.36	4.43	7.43	-0.24	-0.79	0.76	0.74	0.30
Tamil	Dravidian	Tamil	4.65	5.68	7.08	-0.28	-0.88	0.82	0.58	0.32
Persian	Indo-European	Arabic	2.69	3.80	5.49	-0.21	-0.92	0.71	0.60	0.22
Assamese	Indo-European	Assamese	4.10	4.57	5.36	-0.31	-0.68	0.90	0.62	0.46
Russian	Indo-European	Cyrillic	4.05	6.31	9.00	-0.13	-0.94	0.64	0.54	0.14
Ukrainian	Indo-European	Cyrillic	4.52	5.52	7.67	-0.16	-0.86	0.82	0.68	0.19
Panjabi	Indo-European	Devanagari	3.55	3.68	3.88	-0.32	-0.48	0.96	0.60	0.68
Modern Greek	Indo-European	Greek	3.61	4.85	7.64	-0.24	-0.85	0.75	0.69	0.29
Breton	Indo-European	Latin	2.90	3.97	6.31	-0.24	-0.90	0.73	0.69	0.26
Catalan	Indo-European	Latin	2.99	4.90	8.58	-0.15	-0.92	0.61	0.66	0.17
Czech	Indo-European	Latin	3.81	4.83	7.17	-0.22	-0.92	0.79	0.69	0.24
Dutch	Indo-European	Latin	3.24	4.72	8.26	-0.28	-0.95	0.69	0.71	0.29
English	Indo-European	Latin	2.39	4.61	7.79	-0.07	-0.83	0.52	0.59	0.09
French	Indo-European	Latin	2.67	5.04	8.13	-0.04	-0.85	0.53	0.57	0.04
German	Indo-European	Latin	3.00	5.73	10.30	-0.12	-0.87	0.52	0.63	0.13
Irish	Indo-European	Latin	3.01	4.20	6.58	-0.21	-0.93	0.71	0.66	0.23
Italian	Indo-European	Latin	3.16	5.29	8.16	-0.06	-0.90	0.60	0.57	0.06
Latvian	Indo-European	Latin	3.95	4.79	7.09	-0.26	-0.73	0.82	0.73	0.35
Polish	Indo-European	Latin	3.82	5.27	7.87	-0.17	-0.94	0.72	0.64	0.18
Portuguese	Indo-European	Latin	3.14	4.53	7.49	-0.19	-0.91	0.69	0.68	0.21

Continued on next page.

Table D2: Word lengths and their optimality in CV (continued)

Language	Family	Script	L_{min}	L	L_r	τ	τ_{min}	η	Ψ	Ω
Romanian	Indo-European	Latin	3.61	5.03	7.67	-0.21	-0.78	0.72	0.65	0.27
Romansh	Indo-European	Latin	3.83	4.94	7.56	-0.24	-0.79	0.78	0.70	0.30
Slovenian	Indo-European	Latin	3.65	4.56	6.43	-0.21	-0.87	0.80	0.67	0.24
Spanish	Indo-European	Latin	2.74	5.01	7.92	-0.03	-0.91	0.55	0.56	0.04
Swedish	Indo-European	Latin	3.08	4.04	6.87	-0.28	-0.90	0.76	0.75	0.31
Welsh	Indo-European	Latin	2.79	4.17	7.05	-0.21	-0.91	0.67	0.68	0.23
Western Frisian	Indo-European	Latin	3.44	4.38	7.99	-0.29	-0.80	0.79	0.79	0.36
Oriya	Indo-European	Odia	3.68	4.21	5.35	-0.33	-0.73	0.87	0.68	0.46
Dhivehi	Indo-European	Thaana	1.97	3.32	7.61	-0.16	-0.84	0.59	0.76	0.19
Georgian	Kartvelian	Georgian	5.91	7.17	8.22	-0.12	-0.67	0.82	0.45	0.18
Basque	Language isolate	Latin	4.44	6.41	8.89	-0.16	-0.91	0.69	0.56	0.18
Mongolian	Mongolic	Mongolian	4.18	5.47	7.31	-0.23	-0.86	0.76	0.59	0.26
Kinyarwanda	Niger-Congo	Latin	3.72	6.13	9.20	-0.19	-0.90	0.61	0.56	0.21
Abkhazian	Northwest Caucasian	Cyrillic	5.59	5.94	6.42	-0.32	-0.55	0.94	0.58	0.59
Hakha Chin	Sino-Tibetan	Latin	2.56	3.29	5.29	-0.29	-0.81	0.78	0.73	0.35
Chuvash	Turkic	Cyrillic	4.79	6.00	7.35	-0.22	-0.83	0.80	0.53	0.27
Kirghiz	Turkic	Cyrillic	4.49	6.01	7.78	-0.19	-0.89	0.75	0.54	0.22
Tatar	Turkic	Cyrillic	4.04	5.41	7.45	-0.24	-0.89	0.75	0.60	0.27
Yakut	Turkic	Cyrillic	5.02	6.32	7.99	-0.26	-0.74	0.79	0.56	0.36
Turkish	Turkic	Latin	4.60	6.00	8.09	-0.22	-0.89	0.77	0.60	0.24
Estonian	Uralic	Latin	4.68	6.16	8.85	-0.24	-0.84	0.76	0.65	0.29

Table D3: Word lengths and their optimality in CV. Word length is measured in duration. The format is the same as in Table D2.

Language	Family	Script	L_{min}	L	L_r	τ	τ_{min}	η	Ψ	Ω
Arabic	Afro-Asiatic	Arabic	0.35	0.46	0.58	-0.12	-0.89	0.76	0.52	0.14
Maltese	Afro-Asiatic	Latin	0.26	0.35	0.54	-0.21	-0.81	0.74	0.68	0.26
Vietnamese	Austroasiatic	Latin	0.21	0.29	0.33	-0.07	-0.80	0.72	0.33	0.09
Indonesian	Austronesian	Latin	0.27	0.38	0.52	-0.22	-0.91	0.72	0.58	0.24
Esperanto	Conlang	Latin	0.35	0.49	0.81	-0.18	-0.91	0.71	0.69	0.20
Interlingua	Conlang	Latin	0.32	0.40	0.69	-0.24	-0.79	0.81	0.79	0.31
Tamil	Dravidian	Tamil	0.45	0.54	0.66	-0.31	-0.87	0.84	0.60	0.36
Persian	Indo-European	Arabic	0.26	0.36	0.54	-0.25	-0.99	0.73	0.65	0.25
Assamese	Indo-European	Assamese	0.36	0.43	0.50	-0.22	-0.68	0.84	0.52	0.32
Russian	Indo-European	Cyrillic	0.30	0.42	0.60	-0.15	-0.94	0.71	0.60	0.15
Ukrainian	Indo-European	Cyrillic	0.37	0.43	0.59	-0.18	-0.86	0.85	0.72	0.20
Panjabi	Indo-European	Devanagari	0.65	0.70	0.73	-0.18	-0.45	0.92	0.38	0.40
Modern Greek	Indo-European	Greek	0.29	0.38	0.63	-0.21	-0.84	0.76	0.72	0.26
Breton	Indo-European	Latin	0.22	0.31	0.51	-0.25	-0.88	0.72	0.71	0.28
Catalan	Indo-European	Latin	0.24	0.35	0.68	-0.21	-0.94	0.67	0.73	0.23
Czech	Indo-European	Latin	0.30	0.37	0.57	-0.21	-0.92	0.81	0.74	0.23
Dutch	Indo-European	Latin	0.22	0.29	0.55	-0.28	-0.95	0.75	0.78	0.29
English	Indo-European	Latin	0.19	0.33	0.67	-0.17	-0.83	0.59	0.72	0.21
French	Indo-European	Latin	0.21	0.32	0.63	-0.21	-0.86	0.64	0.73	0.24
German	Indo-European	Latin	0.25	0.37	0.76	-0.22	-0.87	0.67	0.76	0.25
Irish	Indo-European	Latin	0.22	0.30	0.47	-0.24	-0.93	0.76	0.71	0.26
Italian	Indo-European	Latin	0.26	0.38	0.65	-0.19	-0.90	0.70	0.71	0.21
Latvian	Indo-European	Latin	0.32	0.39	0.59	-0.23	-0.74	0.84	0.77	0.31
Polish	Indo-European	Latin	0.30	0.38	0.57	-0.17	-0.95	0.79	0.71	0.18
Portuguese	Indo-European	Latin	0.27	0.35	0.61	-0.22	-0.92	0.76	0.76	0.24

Continued on next page.

Table D3: Word lengths and their optimality in CV (duration) – continued

Language	Family	Script	L_{min}	L	L_r	τ	τ_{min}	η	Ψ	Ω
Romanian	Indo-European	Latin	0.27	0.36	0.57	-0.23	-0.77	0.74	0.69	0.30
Romansh	Indo-European	Latin	0.33	0.41	0.66	-0.26	-0.78	0.82	0.77	0.34
Slovenian	Indo-European	Latin	0.36	0.44	0.63	-0.25	-0.85	0.83	0.72	0.30
Spanish	Indo-European	Latin	0.22	0.36	0.62	-0.14	-0.91	0.63	0.67	0.15
Swedish	Indo-European	Latin	0.22	0.27	0.52	-0.29	-0.90	0.81	0.83	0.33
Welsh	Indo-European	Latin	0.22	0.32	0.58	-0.20	-0.93	0.70	0.73	0.22
Western Frisian	Indo-European	Latin	0.26	0.32	0.61	-0.31	-0.80	0.80	0.82	0.39
Oriya	Indo-European	Odia	0.35	0.39	0.49	-0.33	-0.72	0.89	0.70	0.45
Dhivehi	Indo-European	Thaana	0.14	0.32	0.71	-0.17	-0.82	0.46	0.70	0.21
Georgian	Kartvelian	Georgian	0.44	0.52	0.61	-0.15	-0.65	0.85	0.52	0.23
Basque	Language isolate	Latin	0.33	0.44	0.63	-0.21	-0.93	0.73	0.61	0.22
Mongolian	Mongolic	Mongolian	0.29	0.36	0.48	-0.25	-0.85	0.80	0.62	0.29
Kinyarwanda	Niger-Congo	Latin	0.27	0.44	0.72	-0.21	-0.89	0.61	0.62	0.24
Abkhazian	Northwest Caucasian	Cyrillic	0.67	0.74	0.81	-0.20	-0.55	0.90	0.47	0.37
Hakha Chin	Sino-Tibetan	Latin	0.22	0.29	0.44	-0.25	-0.83	0.76	0.69	0.30
Chuvash	Turkic	Cyrillic	0.36	0.44	0.54	-0.26	-0.82	0.82	0.57	0.32
Kirghiz	Turkic	Cyrillic	0.34	0.44	0.57	-0.20	-0.89	0.78	0.56	0.22
Tatar	Turkic	Cyrillic	0.31	0.38	0.52	-0.26	-0.88	0.80	0.64	0.29
Yakut	Turkic	Cyrillic	0.35	0.43	0.54	-0.25	-0.73	0.81	0.56	0.35
Turkish	Turkic	Latin	0.32	0.41	0.54	-0.21	-0.89	0.78	0.60	0.23
Estonian	Uralic	Latin	0.32	0.39	0.58	-0.23	-0.82	0.81	0.71	0.28

D.2 Substituting Kendall τ correlation with Spearman ρ correlation in the definition of Ω

We use Spearman's ρ correlation instead of Kendall's τ in the definition of Ω so as to understand if a different rank-correlation coefficient would lead to a change in the scale of variation of values of Ω, turning them closer to those of Ψ. In Table D4, we show Ω_τ, the original definition of Ω, Ω_ρ, the new definition of Ω that results from replacing τ by ρ, and all the correlations necessary to compute them. The

values of Ω_ρ increase with respect to those of Ω_τ, but by no more than 10%, suggesting that the rather low values of Ω compared to those of Ψ, are not due to the initial choice of τ to define Ω (Equation 4). We suspect that the low values of Ω_ρ and Ω_τ could originate from some similarity between τ and ρ (e.g. both are rank correlation scores and then both may yield low values of Ω) or the fact that the template of the definition of Ω_ρ and that of Ω_τ is the same (Ω_ρ and Ω_τ differ only in the choice of the correlation coefficient). However, that issue should be the subject of further research.

Table D4: Kendall τ versus Spearman ρ in the frequency-length correlation and also in Ω in the PUD collection. τ and ρ are the original correlations and τ_{min} and ρ_{min} are the minimum correlations according to the rank ordering minimum baseline. Ω_τ is the original value of Ω whereas Ω_ρ is the value of Ω that is obtained when Kendall τ is replaced by Spearman ρ in the definition of Ω.

Language	**Family**	**Script**	τ	ρ	τ_{min}	ρ_{min}	Ω_τ	Ω_ρ
Arabic	Afro-Asiatic	Arabic	-0.13	-0.16	-0.74	-0.82	0.18	0.19
Indonesian	Austronesian	Latin	-0.09	-0.11	-0.81	-0.89	0.11	0.12
Russian	Indo-European	Cyrillic	-0.19	-0.23	-0.64	-0.73	0.30	0.32
Hindi	Indo-European	Devanagari	-0.19	-0.23	-0.78	-0.86	0.24	0.27
Czech	Indo-European	Latin	-0.22	-0.27	-0.67	-0.75	0.34	0.36
English	Indo-European	Latin	-0.20	-0.24	-0.76	-0.85	0.26	0.28
French	Indo-European	Latin	-0.16	-0.20	-0.75	-0.83	0.21	0.23
German	Indo-European	Latin	-0.23	-0.28	-0.68	-0.77	0.33	0.36
Icelandic	Indo-European	Latin	-0.26	-0.32	-0.66	-0.75	0.40	0.42
Italian	Indo-European	Latin	-0.15	-0.18	-0.74	-0.83	0.20	0.22
Polish	Indo-European	Latin	-0.18	-0.22	-0.65	-0.73	0.29	0.30
Portuguese	Indo-European	Latin	-0.19	-0.24	-0.74	-0.83	0.26	0.28
Spanish	Indo-European	Latin	-0.16	-0.19	-0.74	-0.83	0.21	0.23
Swedish	Indo-European	Latin	-0.23	-0.28	-0.68	-0.77	0.34	0.36
Japanese	Japonic	Japanese	-0.21	-0.25	-0.60	-0.67	0.35	0.37
Japanese-strokes	Japonic	Japanese	-0.09	-0.12	-0.78	-0.87	0.12	0.13
Japanese-romaji	Japonic	Latin	-0.15	-0.19	-0.80	-0.87	0.19	0.21
Korean	Koreanic	Hangul	-0.24	-0.27	-0.64	-0.71	0.38	0.39
Thai	Kra-Dai	Thai	-0.23	-0.29	-0.82	-0.90	0.28	0.32
Chinese	Sino-Tibetan	Han (Traditional variant)	-0.24	-0.27	-0.55	-0.59	0.44	0.45
Chinese-strokes	Sino-Tibetan	Han (Traditional variant)	-0.18	-0.24	-0.77	-0.87	0.24	0.27
Chinese-pinyin	Sino-Tibetan	Latin	-0.18	-0.21	-0.77	-0.85	0.23	0.25
Turkish	Turkic	Latin	-0.24	-0.29	-0.67	-0.76	0.36	0.38
Finnish	Uralic	Latin	-0.23	-0.28	-0.60	-0.69	0.39	0.41

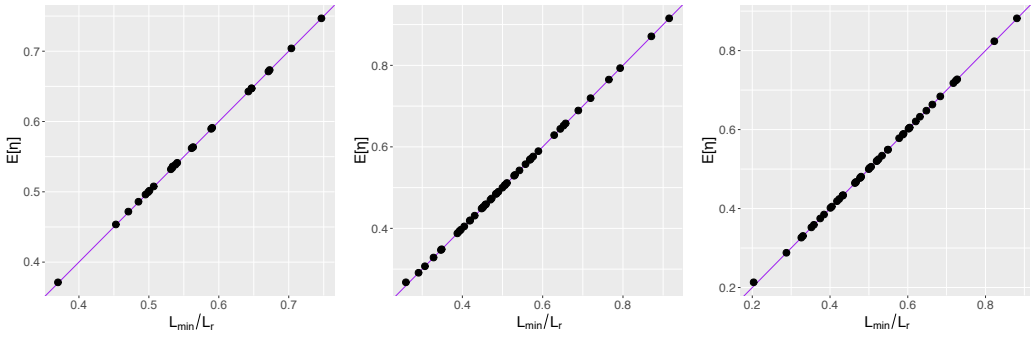
D.3 Stability under the null hypothesis

We estimate $\mathbb{E}[\eta]$, $\mathbb{E}[\Psi]$, and $\mathbb{E}[\Omega]$ under the null hypothesis of a random mapping of word frequencies into lengths by means of a Monte Carlo procedure over a maximum of 10^6 randomizations. The goal is to confirm that Ψ and Ω both yield values tending to zero (Table A1) while checking if η is eventually stable under that null hypothesis.

Table D5: Summary statistics of estimates of $\mathbb{E}[\eta]$, $\mathbb{E}[\Psi]$ and $\mathbb{E}[\Omega]$ under the null hypothesis, for languages in every collection and each definition of length in CV. In PUD, scores in strokes for Chinese and Japanese are excluded for the sake of homogeneity. 'sd' stands for standard deviation.

Score	Collection	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sd
$\mathbb{E}[\eta]$	PUD-characters	0.472	0.526	0.552	0.580	0.647	0.747	0.079
	CV-characters	0.268	0.423	0.495	0.518	0.575	0.915	0.145
	CV-duration	0.213	0.426	0.504	0.518	0.603	0.882	0.139
$\mathbb{E}[\Psi]$	PUD-characters	-0.000	-0.000	0.000	-0.000	0.000	0.000	0.000
	CV-characters	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000
	CV-duration	-0.000	-0.000	-0.000	0.000	0.000	0.001	0.000
$\mathbb{E}[\Omega]$	PUD-characters	-0.000	-0.000	-0.000	-0.000	0.000	0.000	0.000
	CV-characters	-0.000	-0.000	0.000	-0.000	0.000	0.000	0.000
	CV-duration	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000

The summary in Table D5 confirms the predictions in Section C.3, namely that $\mathbb{E}[\Psi]$ and $\mathbb{E}[\Omega]$ are close to 0, while $\mathbb{E}[\eta]$ takes values spread across its whole domain, ranging from a minimum of 0.21 in CV (duration) to a maximum of 0.91 in CV (characters). However, the values taken by $\mathbb{E}[\eta]$ are not arbitrary: the lower bound derived in Section C.3 is actually a very accurate estimate of the expectation itself (Figure D1).



(a) PUD (b) CV(characters) (c) CV(duration)

Figure D1: The Monte Carlo estimate of $\mathbb{E}[\eta]$ against its theoretical lower bound, namely $\frac{L_{min}}{L_r}$. The purple line indicates the identity function, $\mathbb{E}[\eta] = \frac{L_{min}}{L_r}$. (a) PUD collection with length measured in characters. (b) CV collection with length measured in characters. (c) CV collection with length measured in duration.

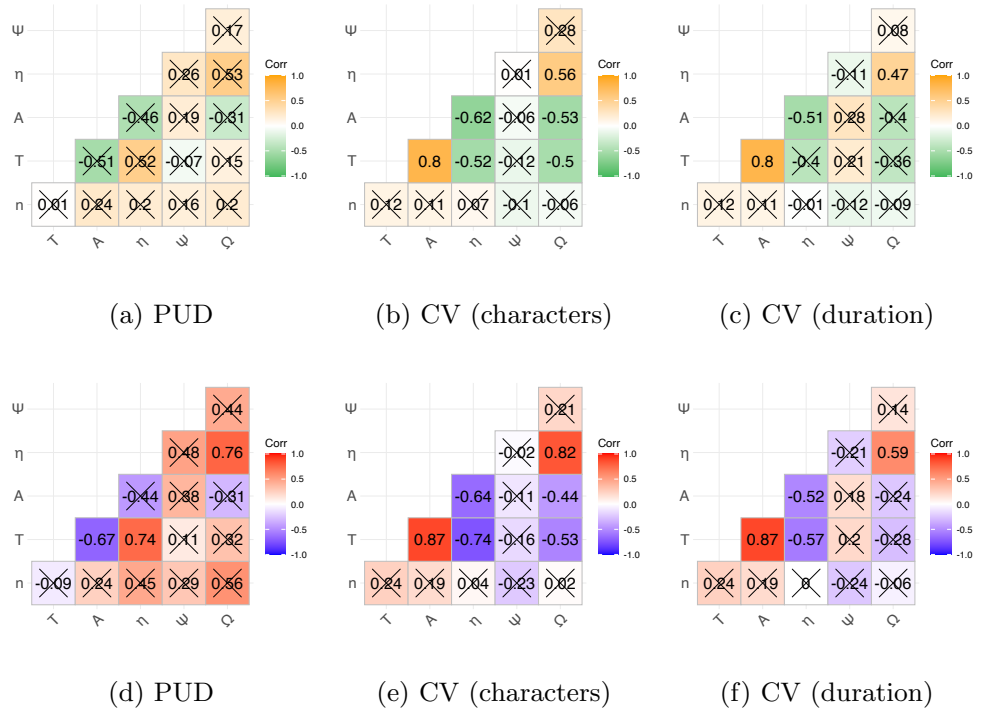
Appendix E Desirable properties of the scores

Here we investigate empirically the desirable properties of scores presented in Section 2.6.

E.1 Dependence of the scores on basic parameters

First, we explore the association between the optimality scores (η, Ψ, Ω) and basic language parameters: observed alphabet size A (number of distinct characters), observed vocabulary size n (number of types) and text length T (number of tokens). The alphabet and vocabulary size observed in the analyzed text samples are lower bounds of the true values in the language due to undersampling. However, the real parameters are likely to be a strictly monotonic function of the observed ones. For this reason, we measure the association with Kendall τ correlation, that is known to be invariant under strictly monotonically increasing transformations (Section C). Given the recent challenges to the common view about the limits of Pearson r correlation (van den Heuvel and Zhan, 2022), we also use r to verify the robustness of the results.

In the parallel corpus PUD, where there is less diversity in terms of basic parameters, the only significant association found for a score is that between η and T, when Pearson correlation is used (Figure E1 (a, d)). Concerning the CV collection, η and Ω show sensitivity to the alphabet and sample size of a language, especially when length is measured in characters (Figure E1 (b, e)). However, the association of the scores with alphabet size might be a reflection of the strong correlation existing between the amount of tokens and the subset of the real alphabet that can be observed in them. Interestingly, when length is measured in duration, the association with η remains while no significant association is detected between Ω and basic parameters, consistently for the two correlation coefficients (Figure E1 (c, f)). Thus, Ψ is the only score among the analysed ones that does not seem to be related to A, n, or T, even in a highly heterogeneous collection.



(a) PUD (b) CV (characters) (c) CV (duration)

(d) PUD (e) CV (characters) (f) CV (duration)

Figure E1: Correlations between optimality scores (η, Ψ and Ω) and basic language parameters (T, n and A). Recall that T is the number of tokens, n is the number of types, A is the alphabet size. Correlation tests are two-sided and p-values are corrected using Holm-Bonferroni correction. A crossed value indicates a non significant correlation coefficient at a 95% confidence level. For PUD, only immediate word constituents are used to measure word length in Chinese and Japanese. (a) Kendall τ correlation in PUD collection with word length measured in characters. (b) Kendall τ correlation in the CV collection with word length measured in characters. (c) Kendall τ correlation for CV collection with word length measured in duration. (d) Same as (a) using Pearson correlation. (e) Same as (b) using Pearson correlation. (f) Same as (c) using Pearson correlation.

E.2 Convergence speed

We investigate the dependence of η, Ψ, and Ω on text size t (in number of tokens) within each language. We wish to know whether they converge or not and their convergence speed.

Methods

For a given t, we sample t tokens uniformly at random from the whole text. We do not use a prefix of length t of the text as in related research on convergence of entropy estimators (Bentz et al., 2017) because both PUD and CV contain a series of disconnected groups of sentences. In PUD each group of sentences is taken from a distinct text source and groups consist of a few sentences, often just one sentence. By proceeding in this way, we are easing the convergence of the estimators of the scores that we use, that neglect word order in their definition. We explore t by increasing powers of 2. For each t, we perform 10^2 experiments and compute the average over the available values. Indeed, for small sample sizes, computations could result in divisions by 0, or the impossibility to compute τ (for Ω) given the low amount of distinct types, thus not yielding a value for that particular experiment.

Results

In Figure E2, we show the results for PUD; in Figure E3 for CV when length is measured in characters; in Figure E4 when length is measured in duration. In all cases Ω has the widest range of variation, and it keeps rapidly decreasing as sample size grows, making it hard to make inferences about its possible convergence. Both η and Ψ evolve within a smaller range, however – while the former mainly shows a slow but decreasing trend – Ψ seems to approach stability in some languages, especially when the sample is large enough. In fact, we can mainly observe this phenomenon in CV, which contains larger samples.

PUD is parallel but texts are rather short. Now we focus on CV because it has the largest samples and then is ideal for investigating convergence. The languages for which convergence is suggested visually are: English, Kinyarwanda, Persian, Tatar and Welsh when length is measured in characters, and French, Kinyarwanda, Persian, Romanian and Tatar when length is measured in duration. The observed trends also suggest that the scores computed in the under-sampled languages are likely to vary largely in a larger sample of the same language, turning comparisons of these scores potentially misleading in non-parallel sources.

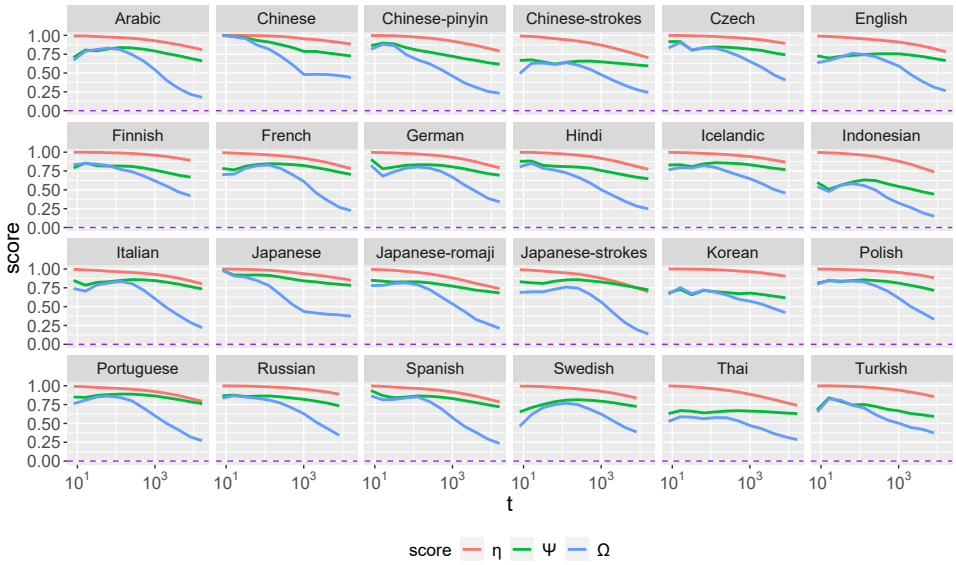


Figure E2: Convergence of the scores in the PUD collection. Values of the optimality scores (η, Ψ, and Ω) for increasing number of tokens t. For each value of t, we show the average value of the score over 10^2 random experiments (only those for which a value could be computed). The dashed line marks 0.

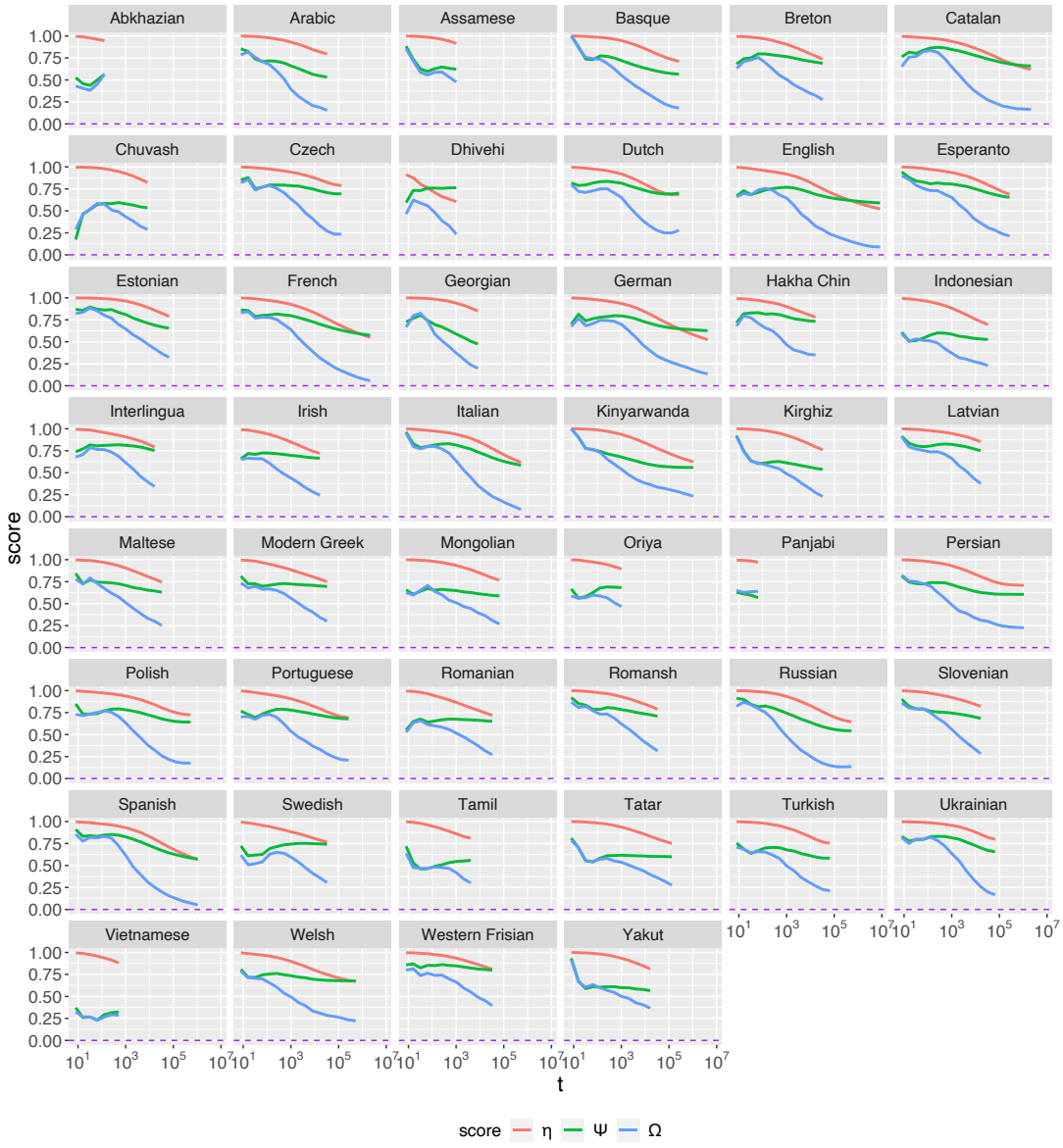


Figure E3: Convergence of the scores in the CV collection with word length measured in characters. The format is the same as in Figure E2.

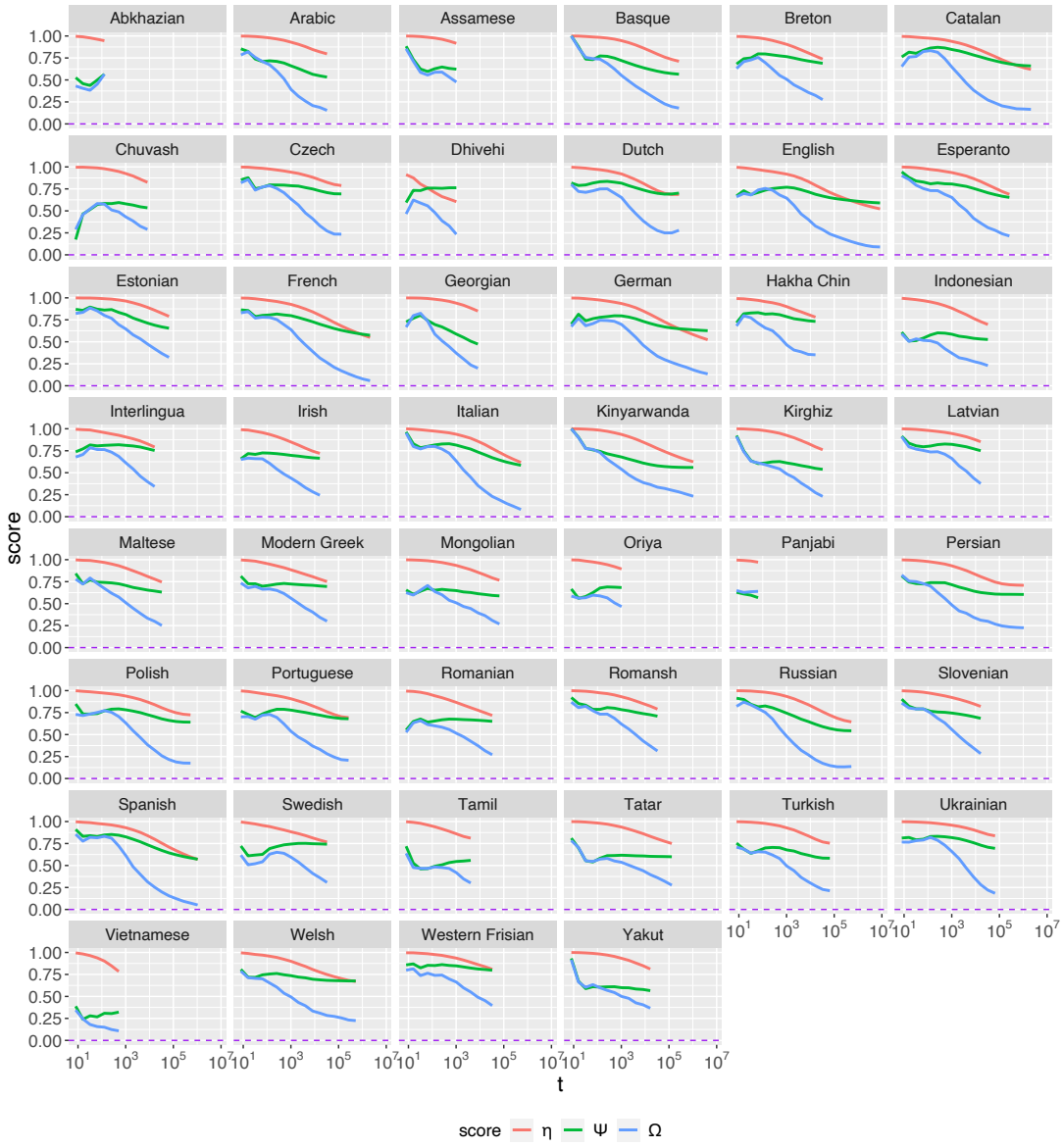
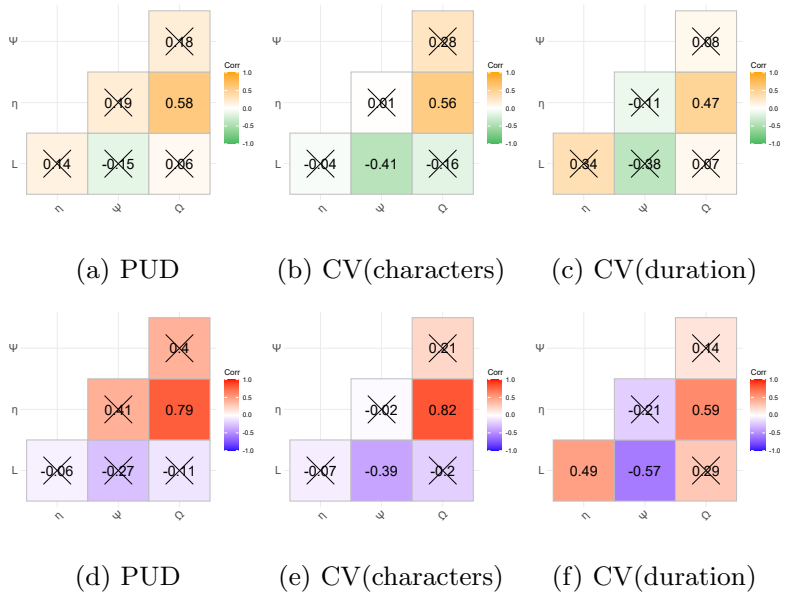


Figure E4: Convergence of the scores in the CV collection with word length measured in duration. The format is the same as in Figure E2.

E.3 Can a score be replaced by a simpler one?

To examine the replaceability of a score with a simpler one, we analyze the correlations between scores. In particular, L is considered the simplest score (it does not integrate any baseline), after which comes η (as it only integrates the minimum baseline). According to both Kendall and Pearson correlation, the only significant association in PUD is found between Ω and η (Figure E5 (a-d)). This relation is consistently found also for both definitions of length in CV (Figure E5 (b, c, e and f)), suggesting its reality even in non-parallel corpora. In the context of CV, Ψ turns out to be significantly associated with L, the simplest score, especially when length is measured in characters. Despite the significance, CV is not parallel and the associations are not particularly strong (the absolute value of the correlations does not exceed 0.6),

suggesting that the additional complexity introduced by Ψ still adds additional information on top of the information provided by L.



(a) PUD (b) CV(characters) (c) CV(duration)

(d) PUD (e) CV(characters) (f) CV(duration)

Figure E5: The correlation between scores. Correlation tests were two-sided the p-values were corrected with a Holm-Bonferroni correction. A crossed value signals a non significant correlation coefficient at a 95% confidence level. (a) Kendall τ correlation in PUD collection with word length measured in characters. (b) Kendall τ correlation in the CV collection with word length measured in characters. (c) Kendall τ correlation for CV collection with word length measured in duration. (d) Same as (a) using Pearson correlation. (e) Same as (b) using Pearson correlation. (f) Same as (c) using Pearson correlation.