Glottometrics

60/2026

International Quantitative Linguistics Association

Glottometrics

Vol. 60, 2026

Glottometrics is an open access scientific journal for the quantitative research of language and text published twice a year by International Quantitative Linguistics Association (IQLA). The journal was established in 2001 by a pioneer of Quantitative Linguistics Gabriel Altmann.

Manuscripts can be submitted by email to glottometrics@gmail.com.
Submission guideline is available at https://glottometrics.iqla.org/.

International Quantitative Linguistics Association (IQLA)

Friedmanngasse 50
1160 Vienna
Austria

eISSN 2625-8226

Contents

An Improved Karlin Model Fit Test: Application to English and Uzbek Texts and Challenges **1–18**
Shahzod Fayzullaev

From ‘said’ to ‘said differently’: modelling repetition versus lexical variation in English-to-Slovak translation of reporting verbs in literary novels **19–39**
Łukasz Grabowski, Filip Kalaš, Daniel Borysowski

Comparative analysis of linguistic features and machine learning methods in the task of assessing the complexity of texts **40–64**
Artem Zaikin, Valery Solovyev, Marina Solnyshkina

Syntactic Complexity Across Genres in Karel Čapek’s Writing **65–76**
Michaela Nogolová, Xinying Chen, Miroslav Kubát, Žaneta Stiborská

The optimality of word lengths. Theoretical foundation and an empirical study **77–149**
Sonia Petrini, Antoni Casas-i-Muñoz, Jordi Cluet-i-Martinell, Mengxue Wang, Christian Bentz, Ramon Ferrer-i-Cancho

An Improved Karlin Model Fit Test: Application to English and Uzbek Texts and Challenges

Shahzod Fayzullaev[1,2*]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Urgench State University, Urgench, Uzbekistan
[*] Corresponding author's email: shahzodjohn1@gmail.com

DOI: https://doi.org/10.53482/2026_60_430

ABSTRACT

Zipf's law and similar frequency laws have been studied in many languages, but their behavior in Uzbek has not been investigated. In this paper, we examine the fit of the generalized Karlin model of Zipf's law to Uzbek texts. For 386 texts consisting of three different genres (prose, poetry, and newspapers), we compute the T_n and H_n statistics and their analogs in the first half of the text and propose a new goodness-of-fit statistic Q_n based on their joint asymptotic behaviour. Our results show that model fit varies systematically with text length and genre: newspapers and poetry, which are typically shorter and more thematically compact, fit the model significantly better than long narrative prose. These results clarify how the Karlin model works in Uzbek texts and provide an empirical baseline for future comparative studies of understudied languages, including agglutinative ones. We also compare the new statistic with two previously proposed tests, based on type and hapax, on ten English reference texts. The results show that Q_n produces virtually identical p-values to the second test and leads to the same accept/reject decisions at standard significance levels, while the first test is systematically more conservative.

Keywords: Karlin model, vocabulary growth, hapax legomena, goodness-of-fit, quantitative linguistics, Uzbek texts

1 Introduction

Hapax legomena are words that appear in a text only once. This term comes from ancient Greek and means "said once". In linguistics, the number of hapax legomena is an important indicator that reflects the stylistic features of the author, the complexity of the text or the structure of the language. This term was first introduced into scientific circulation by Harrison (1921), and it was emphasized that it determines the style of the author. An infinite urn scheme is a probabilistic model in which n balls are thrown into an infinite box with probability p_k, independently and with the same distribution. This scheme, proposed by Karlin (1967), is applied in linguistics as follows: we have a text corpus of size n, consecutive words in the text are like balls, and they correspond to one of the words in an infinite dictionary with probability p_k, just as a ball falls into one of the infinite boxes. The number of words that occur only once in the text, i.e. hapaxes, is defined as the number of boxes containing only one ball after throwing n balls, and the number of non-empty boxes is defined as the number of different words in the text, which is usually called the number of types or simply types. Petrini et al. (2023) investigated the relationship between word frequency and word length in a wide range of languages. They analyzed the average word length L and showed that more frequent words tend to be shorter, in accordance with Zipf's law of contraction. They also introduced a random baseline L_r and showed that the relation $L < L_r$ holds consistently across languages, which they interpret as direct evidence of compression in human language. Galieva and Vavilova (2021) analyzed the syllable structure in Tatar fiction by comparing the first (main) and last (affixal zone) syllables of polysyllabic words. Using χ^2 tests, they found a highly significant discrepancy: simple CV syllables dominate at the beginning of words, while sonorant syllables (SV, SVS, etc.) predominate at the end of words. Tuzzi et al. (2009) statistically confirmed the validity of Zipf's law on a corpus of New Year's addresses of Italian presidents; the model worked well even when the same texts were sometimes written by different authors. Abebe et al. (2024) proposed a word change point method based on the urn model that counts forward and backward different numbers of words to determine the subject/author change point in a text, demonstrating theoretical consistency and performance with an error rate of $< 3\%$ for English and multilingual corpora. And in a recent study, Abebe (2025) developed a new algorithm to detect two change points in a linked text consisting of three different texts using the same probabilistic models as in the previously mentioned study. Kudryavtseva and Kovalevskii (2025) compared AI-generated texts with human-written texts based on word frequency, the Zipf distribution, and the hapax legomena coefficient. The study found that AI-generated texts had a significantly smaller vocabulary and a significantly lower proportion of rare words. Popescu and Altmann (2008) showed that in a large cross-linguistic sample the proportion of hapax legomena in a text relative to types fluctuates around a nearly fixed value. Another study of the Menzerath-Altmann law by Mačutek et al. (2026) found that while most languages show an inverse relationship between word length and the average syllable length of that word, some languages are exceptions to this rule. We refer you to a study conducted in the field of cognitive linguistics Bao et al. (2025), the results of which show that the level-frequency distribution of DM (discourse marker) satisfies the Zipf-Alexeev law at the corpus and text levels. The study reviewed in Milička (2009) studies the type-token relationship as a property of the text rather than as a property of a particular language, and proposes a model based on a combinatorial description of the distribution of different types in the text. Davis (2018) reexamined the relationship between Zipf's and Heaps' laws and proposed a completely new, logarithmic model for the type-token curve. The study showed that smaller texts do not conform to Zipf's law, and the law becomes more accurate as the text size increases. However, it becomes less accurate with larger text sizes, and it is possible to identify an ideal corpus that conforms to Zipf's law for larger texts. Uzbek language has an agglutinative structure, and words are formed by adding together successive morphemes. This leads to the formation of numerous variants for a single lemma. A large number of morphological variants increases the number of hapax legomena and the number of different words (types), which significantly affects the robustness of statistical models such as Zipf and Heaps. These statistical models have not yet been tested in Uzbek, and little research has been conducted in other agglutinative languages. In this study, the Karlin model is applied to Uzbek texts for the first time and analyze original, lemmatized, and stemmed versions of the corpus separately in order to examine how morphological normalization affects the model fit.

More specifically, this paper's contribution is threefold. First, we introduce a new fit statistic, Q_n, for the Karlin infinite urn model, derived from the joint asymptotic behavior of the number of types and hapax legomena at positions n and $n/2$.

Second, we apply this test to a large corpus of 386 Uzbek texts (prose, poetry, and newspapers) in three versions (original, lemmatized, and stemmed) and analyze how the fit of the model depends on genre and text length. Third, we compare Q_n with two previously proposed tests, type-based and hapax-based, on ten English reference texts and show that Q_n behaves very similarly to $Q_n^{(2)}$, while $Q_n^{(1)}$ acts as a more conservative criterion.

2 Material

This study analyzes 386 texts in three different genres:

1) prose (158 works and stories);

2) poetry (167 poems and poetry collections); and

3) various newspapers (61 issues).

2.1 Sources of texts

Uzbek prose and poetry were downloaded from the publicly available Ziyouz (https://www.ziyouz.com) website. Issues of the *Xalq so'zi* and *Yangi Uzbekistan* newspapers were downloaded from the official websites of https://xs.uz and https://yuz.uz, respectively, and issues of the https://press.natlib.uz/ru newspaper "O'zbekiston ovozi" were downloaded from the electronic archive of national publications of the National Library of Uzbekistan (https://press.natlib.uz/ru).

2.2 Normalization and transliteration

All texts were processed using a special program written in Python before lemmatization and stemming: These steps are performed as follows: All Unicode apostrophes (ʼ, ’, ‘, ʻ) were replaced with the "‘"

symbol. Although UzbekLemma is generally resistant to apostrophe changes, accurate normalization prevents an artificial increase in the number of hapaxes and types in the text. All punctuation marks, with the exception of Uzbek apostrophes, were removed. Page numbers and other layout-related numeric designations were removed, but numeric expressions that were part of the current text (e.g., "2025-yil," "16-bob") were retained as tokens. Proper nouns (personal names, toponyms, and organizational names) and loanwords were not filtered out and were counted as ordinary word types.

2.3 Evaluation of Lemmatization and Stemming Tools

Transliteration of Cyrillic texts into Latin was performed using the *UzTransliterator* Python library. This study additionally tested two existing tools – the *UzbekLemma* lemmatizer and the *uznltk* stemmer – to assess the impact of lemmatization and stemming on the model's results. For this purpose, a small test list of 500 Uzbek words from different categories was compiled, and the results of each tool were manually verified.

According to the evaluation results:

- Overall accuracy of *UzbekLemma*: 73.60 %
- Stemming accuracy of *uznltk*: 87.17 %

Lemmatization errors were observed primarily in the following cases:

- homonyms with similar verb and noun forms,
- verb forms with multiple suffixes (kelayotganimizni, borilmaydi, etc.).

Nevertheless, in many cases, the tools correctly normalize words. To illustrate how lemmatization and stemming behave in practice, Table 1 presents several representative examples from a list of tests.

Table 1: Examples of lemmatization and stemming outputs for Uzbek words. Lemmatization preserves the dictionary form, while stemming applies more aggressive affix removal.

Original word	Lemma (UzbekLemma)	Stem (uznltk)
yil	yil	yil
yashil	yashil	yashil
deputatlar	deputat	deputat
tadbirda	tadbir	tadbir
qiladigan	qilmoq	qil

These results show that the tools perform well with common nouns, adjectives, and regular affixed forms. Despite these errors, they do not materially change the main statistical indicators of the analysis. This is because an incorrectly lemmatized word is often transformed into the same normalized form as its other variants. Even if this is technically an incorrect labeling, it still combines several surface forms under a single token. In most cases, this reduces excessive dispersion between different word forms and stabilizes the number of types and hapaxes. Thus, although the above percentages indicate imperfections in the tools, the observed inaccuracies do not lead to statistically significant biases in the analysis.

3 Methodology

Let $\{X_n\}_{n=1}^{\infty}$ denote random variables representing consecutive words in a text that satisfy Karlin's elementary probability model, i.e.

$$p_i = \mathbb{P}(X_1 = i) = l(\theta, i)\, i^{-1/\theta}, \quad i \geq 1, \tag{1}$$

where $l(\alpha, i)$ is a slowly varying function, θ is an unknown parameter, and $0 < \theta < 1$.

We introduce the following definitions:

- n – the total number of words in the corpus under consideration, or, in other words, tokens.
- $T_n, T_{[n/2]}$ – types in the text and in the first half of the text, respectively.
- $H_n, H_{[n/2]}$ – hapaxes in the text and in the first half of the text, respectively.
- $\widehat{\theta} = H_n/T_n$ – an estimate of the parameter θ proposed by Chebunin and Kovalevskii (2019) and proven to be strongly consistent.
- τ_n and η_n – expected value of T_n and H_n, respectively

According to Karlin (1967), we have the law of large numbers for T_n and H_n, as well as the asymptotics for τ_n and η_n as follows:

$$\tau_n \sim \Gamma(1-\theta) l(\theta, n) n^{\theta}, \quad \eta_n \sim \theta\Gamma(1-\theta) l(\theta, n) n^{\theta}, \quad \text{as} \quad n \to \infty. \tag{2}$$

The functional central limit theorem, proven by Chebunin and Kovalevskii (2016), forms the basis of our next theorem. We will construct two processes as follows:

$$Y_n(t) = \left(T_{[nt]} - \tau_{[nt]}\right) / \sqrt{\tau_n}, \qquad Z_n(t) = \left(H_{[nt]} - \eta_{[nt]}\right) / \sqrt{\tau_n}.$$

where $t \in [0, 1]$.

In Karlin's model, both the number of types T_n and the number of hapax legomena H_n grow proportionally to n^θ, and the ratio H_n/T_n approaches θ Chebunin and Kovalevskii (2019) and Karlin (1967). Therefore, differences in their growth rates serve as a sensitive indicator of deviations from the model. Comparing the first half of the text $(T_{[n/2]}, H_{[n/2]})$ with the full text (T_n, H_n) provides a practical way to identify local asymmetries in vocabulary development.

Theorem. Suppose that the discrete probabilistic model given by (1) satisfies the condition

$$p_i = ci^{-1/\theta}(1 + o(i^{-1/2})) \quad \text{as } i \to \infty, \tag{3}$$

where $c > 0$ and $0 < \theta < 1$. Then, for the statistic

$$V_n = \frac{\begin{vmatrix} T_n & H_n \\ T_{[n/2]} & H_{[n/2]} \end{vmatrix}}{T_n^{3/2}},$$

there is weak convergence to a centered normal random variable:

$$V_n \xrightarrow{d} V = Y_1(1/2) - 2^{-\theta}Y_1(1) - \theta\, Y(1/2) + \theta\, 2^{-\theta} Y(1),$$

with zero expectation and variance is determined by the expression

$$\mathrm{Var}(V) = \Sigma^2(\theta) = v(\theta)\, G(\theta)\, v^\top(\theta),$$

where $v(\theta) = (\theta 2^{-\theta}, -\theta, -2^{-\theta}, 1)$ and $G(\theta)$ is a (4×4) matrix from Lemma 1 in Fayzullaev and Kovalevskii (2024).

The proof of the theorem is in the appendix.

Based on the above theorem, we construct statistics to test the elementary probability model in the form of a p-value $= 2\Phi^{-1}(-|Q_n|)$, where $Q_n = V_n/\Sigma(\widehat{\theta})$ and Φ^{-1} is the quantile function of the standard normal distribution. Since $\widehat{\theta}$ actually depends on n, we can redefine $\Sigma(\widehat{\theta})$ as $\Sigma(\widehat{\theta}) := \Sigma_n(\widehat{\theta})$, and in this case $Q_n = V_n/\Sigma_n(\widehat{\theta})$. If p-value $\geq \varepsilon$, we fail to reject the hypothesis that the elementary probability model corresponds to the text at the significance level ε; otherwise, we reject it.

4 Results

The corpus consists of 386 texts across three genres: 158 prose works (novels and short stories), 167 poetry collections and individual poems, and 61 newspaper articles. For each text, we calculated the values of n, T_n, $T_{[n/2]}$, H_n, $H_{[n/2]}$, the estimate $\widehat{\theta} = H_n/T_n$, the standardized statistic Q_n, and the corresponding p-value. A link to the full table is provided in the *Data Availability Statement* section, and the original texts are available via the links in the *Materials* section. In addition to analyzing the original texts, we repeated all calculations on two preprocessed versions of the corpus: lemmatized (UzbekLemma) and stemmed (uznltk). This allows us to test the robustness of the inferences to morphological normalization. Throughout Section 4, we use the 10% significance level ($\varepsilon = 0.1$) as a convenient reference threshold for descriptive summaries. For completeness, Figure 4 also indicates the more common levels 0.05 and 0.01; using these stricter thresholds does not change the qualitative conclusions.

4.1 Genre Patterns

At the level of the original texts, the proportions with p-value < 0.1 are as follows:

- newspapers: 8 out of 61 issues;
- poetry: 29 out of 167 texts;
- prose: 108 out of 158 texts.

Thus, most newspaper issues and poems fit Karlin's model well, while approximately 70% of prose works have p-value < 0.1 and therefore demonstrate a weaker fit.

A similar analysis of the lemmatized corpus yields only minor changes in these proportions: 8 newspaper issues, 41 poetry texts, and 99 prose texts have p-value < 0.1. For the stemmed text corpus, we again obtain 8 newspaper issues with p-values < 0.1, while for poetry and prose these figures become 27 and 64, respectively. In all three cases, newspapers and poetry remain the genres with the best overall fit, while prose systematically shows the worst fit.

For newspapers, the classification into "good fit" and "poor fit" is very stable across all three preprocessing options. Of the 61 issues, 49 have p-values ≥ 0.1 in all three versions (the original, lemmatized, and stemmed versions), and 5 issues have p-values < 0.1 in all three versions. Only a small number of borderline cases (with p-value close to 0.1) change their classification when lemmatization or stemming is applied. This shows that the observed differences between texts and genres are due to genuine structural properties of the texts and not to pre-processing stages.

4.2 Examples of poor and good fit in prose

Within the prose subcorpus, the set of texts that fit Karlin's model well, and those that clearly violate it, is remarkably stable across all three preprocessing options (original, lemmatized, and stemmed texts). In particular, several long works yield very low p-values across all three options, while a small group of shorter prose texts consistently exhibit relatively high p-values. Only a few borderline cases change their classification when lemmatization or stemming is applied. A typical example of a persistent poor match is Oybek's novel "Qutlug' qon". In this novel, p-value in the original version is practically zero (p-value $< 10^{-4}$), and remains very small after lemmatization (p-value ≈ 0.0005) and stemming (p-value ≈ 0.004). As shown in Figure 1, the curves T_k and H_k generally retain a power-law shape, but the number of hapax legomena increases systematically faster than the theoretical prediction $\widehat{\theta}T_k$, especially in the first half of the text. A similar pattern is observed in Risolat Khaidarova's novel "Javzo", where the p-value remains below 0.01 in all three text variants (approximately 0.0002 for the original, 0.0006 for the lemmatized version, and 0.0025 for the stemmed version). These works belong to a small group of seven long prose texts for which the test statistics systematically reject Karlin's model, regardless of whether we analyze the original tokens, lemmas, or stems. This indicates that the rejection is due to the texts' genuine structural properties (length, theme shifts, richness of descriptive fragments), rather than to the peculiarities of morphological preprocessing.

On the other hand, there are also prose works that conform reasonably well to the model in all three variants. For example, in the collections "Sobiq o'g'ri" by Said Ahmad and "Mezon" by Shomirza Turdimov, the p-value exceeds 0.5 for both the original texts and their lemmatized and stemmed versions (for Sobiq o'gri, the p-value is about 0.67 for the original text, 0.58 for the lemmatized version, and 0.75 for the stemmed version; for Mezon, the corresponding p-values are about 0.68, 0.67, and 0.52). In these texts, after a short initial phase, the growth in the number of hapax legomena becomes roughly proportional to the growth in the number of types. For the work "Sobiq o'g'ri", Figure 2 shows that H_k closely follows the theoretical curve $\widehat{\theta}T_k$, while Figure 3 shows that the process $D_k = H_k - \widehat{\theta}T_k$ oscillates around zero without a clear trend. This behavior is typical for works with relatively homogeneous vocabulary and a high degree of repetition, for example, due to frequent dialogue or repetitive narrative patterns.

However, there are several borderline cases where morphological normalization changes the test result. For example, in Hamid G'ulom's novel "Qoradaryo", the p-value for the original text is slightly below 0.1 (around 0.067), while for the lemmatized and stemmed versions it increases to p-values of around 0.77 and 0.84, respectively. Here, lemmatization and stemming remove some of the superficial morphological variability and make the trajectories of T_k and H_k more regular, so the text is classified as compatible with Karlin's model after normalization. However, such cases are rare compared to the highly consistent and highly inconsistent texts discussed above. Overall, these examples demonstrate that long prose works tend to present the greatest challenge to Karlin's elementary probabilistic model. If the text contains strong internal variation, shifting themes, abrupt topic shifts, and bursts of rare words, the test rejects the model in all three preprocessing modes. Conversely, if the prose text has a more homogeneous vocabulary and a high degree of repetition, the model can provide a good approximation, and this result remains stable under lemmatization and stemming.

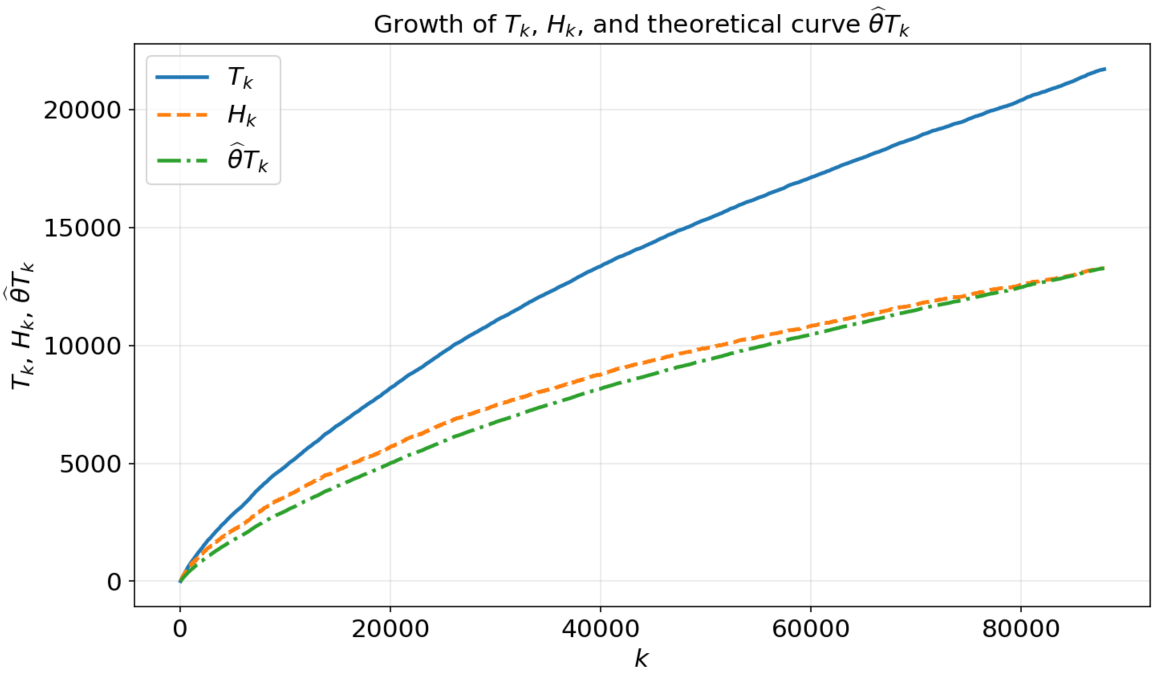


Figure 1: Growth of T_k, H_k, and the theoretical curve $\widehat{\theta}T_k$ in Oybek's novel *Qutlug' qon* (original text). The curve for H_k grows systematically faster than $\widehat{\theta}T_k$, which leads to a very small *p-value*.

4.3 Newspapers and Poetry

For newspapers and poetry, Karlin's model performs significantly better than for prose. In the original newspaper texts, 53 of 61 issues (about 87%) have a p-value ≥ 0.1, and 32 issues (more than half the sample) have a p-value ≥ 0.5. The average p-value in this subcorpus is about 0.52. For poetry, 138 of 167 texts (about 83%) have a p-value ≥ 0.1, and 39 poems achieve a p-value ≥ 0.5, with an average of about 0.34. Thus, shorter and more thematically compact texts tend to demonstrate significantly better matches than longer prose works.

Lemmatization and stemming do not change these findings. In all three versions of the newspaper corpus (original, lemmatized, and stemmed), exactly 8 issues have p-value < 0.1, while the remaining 53 issues remain above 0.1. Five issues are classified as "poorly conforming" in all three versions, and 49 are classified as "well conforming" in all three, so only a small group of seven edge cases change status depending on pre-processing. A similar pattern is observed for poetry: in the original texts, 29 of 167 poems have a p-value < 0.1, while in the lemmatized corpus, 41 poems fall below this threshold, and in the stemmed corpus, 27 do so. However, 15 poems consistently demonstrate a p-value < 0.1 in all three

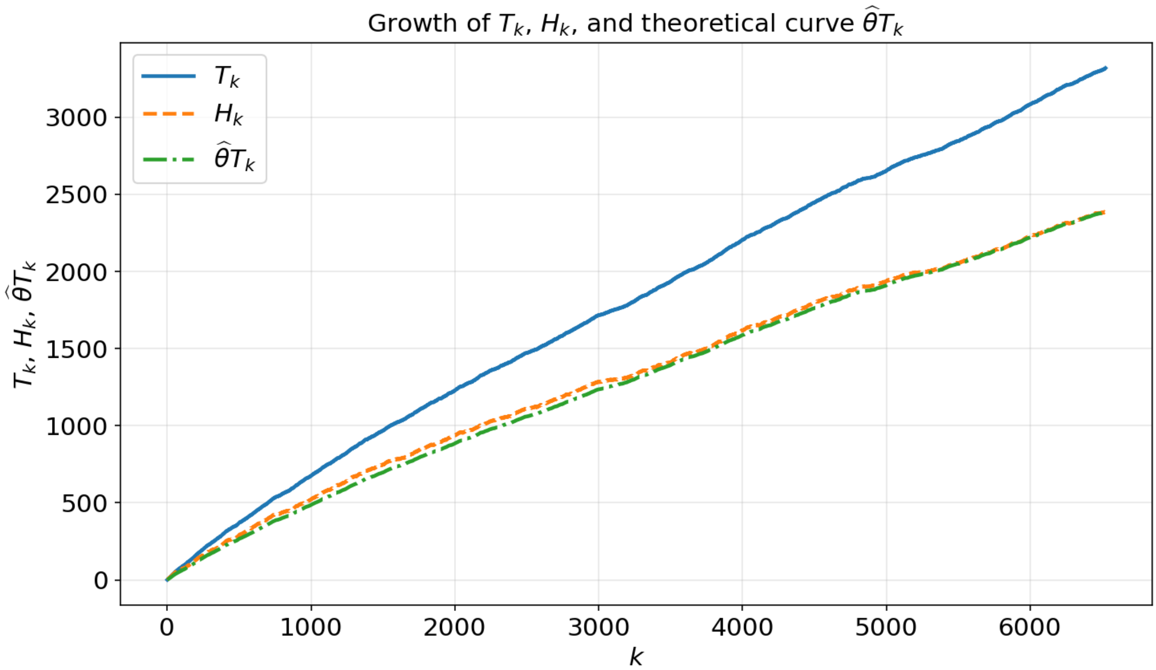


Figure 2: Growth of T_k, H_k, and $\widehat{\theta}T_k$ in Said Ahmad's *Sobiq o'g'ri* (original text). Here H_k closely follows the theoretical curve, indicating a good fit of Karlin's model.

variants, while 116 poems have a p-value ≥ 0.1 in all variants. This stability suggests that the observed differences in model fit are primarily due to the internal structure of the texts, rather than the details of morphological normalization.

Individual examples illustrate these trends. In "Xalq so'zi" issue 179, the p-value is approximately 0.60 for the original text, 0.95 for the lemmatized version, and 0.62 for the stemmed version. Analysis of the trajectories of T and H shows that despite local irregularities caused by shifts between different topics and article types, their growth remains roughly proportional, so Karlin's model provides a satisfactory approximation for this issue. In the poetry subcorpus, Jamal Sirojiddin's "Tanbur" collection represents a typical case of very good fit: the p-value is 0.75 for the original text and increases to 0.96 and 0.99 for the lemmatized and stemmed versions, respectively. Here, the number of hapax legomena closely follows the theoretical curve $\widehat{\theta}T_k$ throughout the text, and morphological normalization makes this proportionality even more regular.

Overall, newspapers and poetry support the interpretation that Karlin's elementary probabilistic model works best for relatively short and structurally homogeneous texts. In such cases, the results remain stable for both lemmatization and stemming, while long text with uneven internal structure continues to be the main source of systematic deviations from the model.

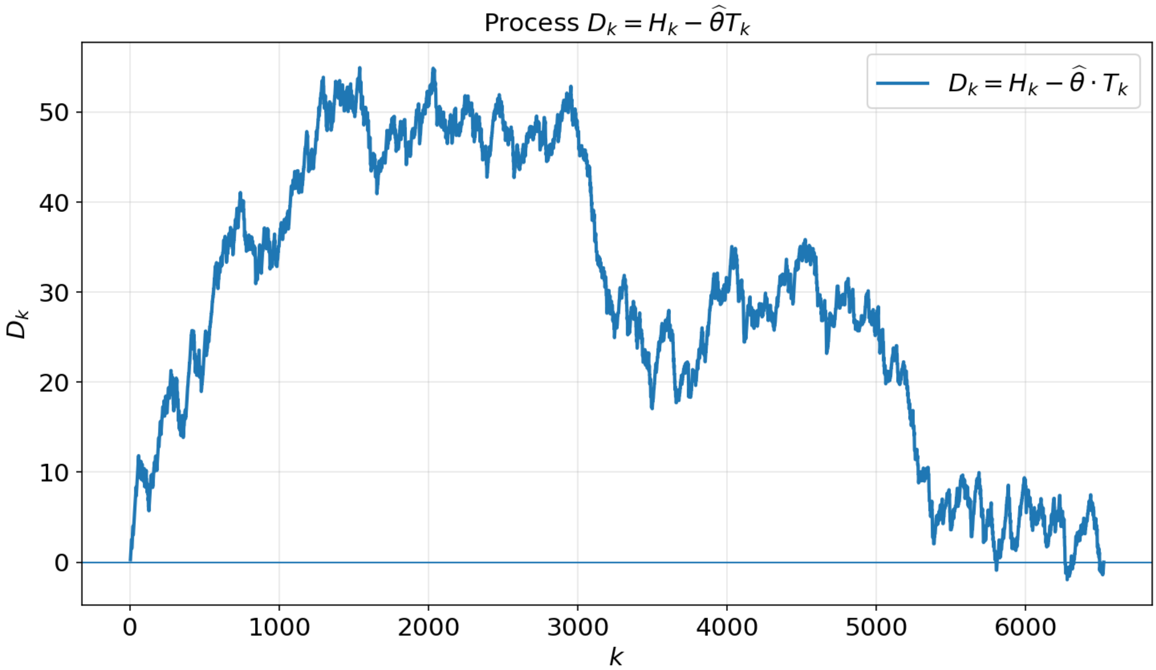


Figure 3: Behaviour of the process $D_k = H_k - \widehat{\theta} T_k$ for Said Ahmad's *Sobiq o'g'ri* (original text). The process fluctuates around zero without a clear trend, which is consistent with the model.

4.4 Dependence on Text Length

The relationship between the p-value and text length n is shown in Figure 4, where each point corresponds to one text from the corpus (based on the original, unnormalized version). The horizontal axis shows the number of tokens n, and the vertical axis the corresponding p-value. The dashed horizontal lines indicate levels p = 0.1, p = 0.05, and p = 0.01.

The figure shows a clear dependence of model fit on text length. For relatively short texts (up to approximately 10^4 words), the points are widely distributed across the interval (0,1), and many texts have p-values significantly higher than 0.1, sometimes close to 1. Poetry and newspapers are well represented in this domain. As n increases, the point cloud gradually shifts downward: for texts of medium length, the proportion of p-values ≥ 0.1 decreases, while for very long texts (with n on the order of 10^4–10^5 words), most p-values lie below 0.1, with many p-values being very small. Almost all of these very low p-values correspond to long prose works.

This pattern is consistent with the general behavior of statistical tests. In particular, for huge samples, classical goodness-of-fit tests may reject the null hypothesis even under very small departures from the model, because test power increases with n. This "large-sample" sensitivity has been discussed in quantitative linguistics; see, for example, Mačutek and Wimmer (2013). Therefore, part of the downward shift of p-values with increasing text length may reflect this general statistical effect. As sample size n increases, a test based on the Q_n statistic becomes more sensitive to systematic differences between the empirical trajectories T_k and H_k and the theoretical predictions of the Karlin model. For short texts, random fluctuations can mask such differences, and the model is rarely rejected. However, for very long texts, localized spikes in rare words (e.g., due to topic changes or extensive descriptive fragments) accumulate across the document, pushing p-values closer to zero.

To ensure that the dependence on n is not a preprocessing artifact, we repeated the same analysis for the lemmatized and stemmed versions of the corpus. In both cases, the distribution of points on the (n, p-value) plane exhibits the same qualitative trend: high p-values are more common among shorter texts, while longer prose works tend to yield low p-values. Morphological normalization slightly shifts some individual texts upward (especially in the newspaper and poetry subcorpora), but does not affect the overall decrease in p-values with increasing n. This suggests that the observed dependence on text length reflects genuine properties of the model and the data, rather than being a byproduct of the preprocessing procedure.

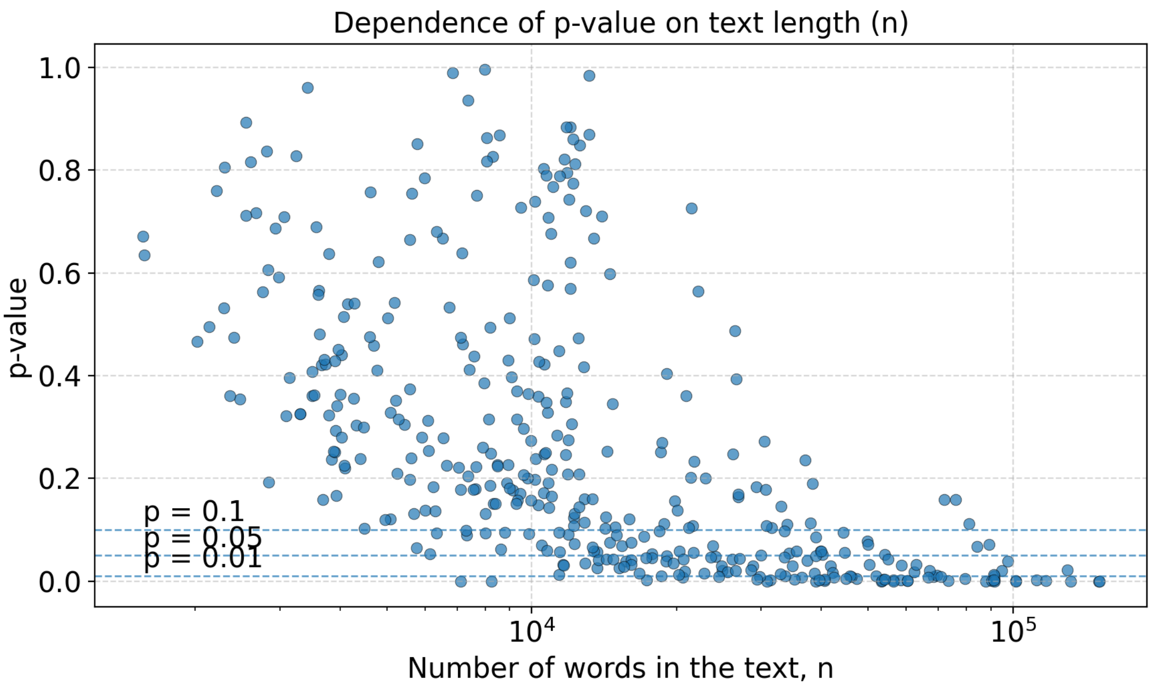


Figure 4: Plot of *p-value* versus text length n (original texts). Each point corresponds to one text from the corpus. Dashed lines indicate levels *p-value* = 0.1, 0.05, and 0.01.

4.5 English texts and comparison with existing tests

To verify that the new Q_n statistic behaves consistently with previously proposed hapax-type tests, we applied all three criteria to ten English texts analyzed in Fayzullaev and Kovalevskii (2024). For six long novels (Alice's Adventures in Wonderland, Dracula, Frankenstein, or the Modern Prometheus, Pride and

Prejudice, The Great Gatsby, and The Scarlet Letter), all three tests yield very low p-values ($p < 0.01$) and therefore confidently reject Karlin's model. In contrast, three shorter, more homogeneous texts (Metamorphosis, Romeo and Juliet, and The Picture of Dorian Gray) have p-values above 0.1 for all three tests, indicating acceptable model fit.

The only borderline case is the popular science text Simple Field Guide to Sabotage. Here, our Q_n and $Q_n^{(2)}$ statistics yield p-values around 0.10, while $Q_n^{(1)}$ yields a smaller p-value of approximately 0.02 and rejects the model at the 10% and 5% levels. Thus, $Q_n^{(1)}$ behaves as a more conservative test, while Q_n is numerically very close to $Q_n^{(2)}$ and leads to the same accept/reject decisions at all standard significance levels.

5 Conclusion

In this paper, we proposed a new fit statistic for the elementary probabilistic Karlin model, based on the joint behavior of numbers of type T_n and hapax legomena H_n at positions n and $[n/2]$, and applied it to a large corpus of Uzbek texts. The corpus includes prose, poetry, and newspapers. For each text, we analyzed three versions: the original token sequence, a lemmatized version, and a stemmed version. This allowed us to study both the theoretical properties of the test and its empirical behavior under different preprocessing conditions.

The main empirical result is that the Karlin model is significantly more suitable for short and structurally homogeneous texts than for long narrative prose. For newspapers and poetry, most texts have p-value ≥ 0.1, with a significant proportion reaching p-values above 0.5. In contrast, only about a third of prose works pass the test at the 0.1 level, and almost all the extremely low p-values in the corpus are generated by long novels and short story collections. The dependence on text length is particularly pronounced: for texts up to 10^4 words long, the points on the $(n, \text{p-value})$ plane are widely scattered, while for very long texts, most p-values cluster near zero. This suggests that Karlin's model captures general patterns of vocabulary growth but becomes sensitive to the long, uneven structure of prose texts.

Analysis of individual texts helps explain this behavior. In several novels with very small p-values, the number of hapax legomena increases much faster than the model predicts, especially in the first half of the text, and the process $D_k = H_k - \widehat{\theta} T_k$ moves away from zero. In other works, such as Said Ahmad's "Sobiq o'g'ri," the growth of H_k quickly becomes proportional to the growth of T_k, and D_k fluctuates around zero without a clear trend, leading to high p-values. These examples demonstrate that deviations from the Karlin model are closely related to localized spikes in rare vocabulary caused by topic shifts, long descriptive passages, or abrupt register changes.

A separate question was whether the observed effects are artifacts of morphological preprocessing. Our additional experiments show that this is not the case. For most texts, the classification into "good" and "bad" matches is robust across the original, lemmatized, and stemmed versions. Lemmatization and stemming slightly improve the match for some borderline cases (e.g., Hamid Gulom's "Qoradaryo"), but do not change the overall picture: newspapers and poems generally conform well to the model, while long prose remains the main source of systematic deviations.

Thus, the proposed statistics provide a convenient and informative tool for testing Karlin's model on real texts and identifying its failures. At the same time, our results show that a single global parameter is often insufficient to describe vocabulary growth in long human-generated texts. Future research will focus on extending the model to locally non-uniform or piecewise regimes (e.g., with parameters changing over the course of the text) and on constructing tests comparing $T_{[tn]}$ and $H_{[tn]}$ at multiple time points $t \in (0, 1)$. Such refinements should allow for a more accurate description of early irregularities while preserving the useful asymptotic structure of Karlin's approach.

In addition to these empirical observations, it is useful to summarize the main methodological properties of the proposed statistics.

In this paper, we introduced a new goodness-of-fit statistic

$$V_n = \frac{T_n H_{\lfloor n/2 \rfloor} - H_n T_{\lfloor n/2 \rfloor}}{T_n^{3/2}}, \qquad Q_n = \frac{V_n}{\Sigma_n(\hat{\theta})},$$

derived from the functional central limit theorem for Karlin's infinite urn scheme. The test exploits the joint behavior of the number of types T_n and the hapax legomena H_n at positions n and $n/2$, not just their finite values. This design makes the statistic particularly sensitive to local variations in vocabulary growth (e.g., topic shifts and spikes in rare words), while remaining robust to small random fluctuations in individual word frequencies.

From a practical standpoint, the test is simple to calculate: it requires only the aggregate values of T_n, H_n, $T_{\lfloor n/2 \rfloor}$, and $H_{\lfloor n/2 \rfloor}$, and it returns a single scalar value Q_n, which can be directly interpreted using the standard normal distribution. This allows for the comparison of a large number of texts and genres on a single scale and the clear identification of systematic deviations from Karlin's model.

For the ten English test texts from Fayzullaev and Kovalevskii (2024), our Q_n statistic produces p-values that are nearly indistinguishable from those of $Q_n^{(2)}$ and gives exactly the same accept/reject decisions at standard significance levels, whereas $Q_n^{(1)}$ systematically produces smaller p-values and hence acts as a more conservative test.

At the same time, we do not claim that the Q_n statistic is universally more powerful than classical goodness-of-fit procedures such as the Kolmogorov-Smirnov test, χ^2, or the likelihood ratio test. Our construction still relies on the simplifying assumptions of Karlin's model, in particular, the approximately independent sample of words. As discussed in the literature on Zipf-type laws (see, e.g., Altmann and Gerlach (2016)), this assumption is known to be violated in real texts due to long-range correlations and topic structure.

Furthermore, Q_n is based only on the number of types and hapax legomena. Consequently, two different texts may have very similar Q_n values even if their full word frequency distributions differ significantly. In this sense, the proposed test is best viewed as a convenient model-based diagnostic measure that complements, rather than replaces, the more general distribution-based tests used in quantitative linguistics.

Acknowledgments

I am grateful to my scientific supervisor, Doctor of Physical and Mathematical Sciences Artem Pavlovich Kovalevskii, for valuable guidance and important recommendations during the preparation of this work. I also acknowledge the support of the "El-Yurt Umidi" Foundation for the Training of Prospective Personnel under the President of the Republic of Uzbekistan, which provided a scholarship during my graduate studies.

I also thank the anonymous referee for careful reading of the manuscript and for constructive comments that helped improve the presentation.

Data availability statement

All datasets used in this study (original, lemmatized, stemmed statistics for 386 texts and the 500-word evaluation list) are publicly available on Zenodo: https://doi.org/10.5281/zenodo.17711166.

References

Abebe, B., Chebunin, M., Kovalevskii, A. (2024). Text segmentation via processes that count the number of different words forward and backward. *Journal of Quantitative Linguistics*, *31*(1), 1–18. https://doi.org/10.1080/09296174.2023.2275342

Abebe, B. (2025). A new method for detecting multiple text change points. *Glottotheory*, *16*(1), 1–15. https://doi.org/10.1515/glot-2025-2003

Altmann, E. G., Gerlach, M. (2016). Statistical laws in linguistics. In M. Degli Esposti, E. G. Altmann, F. Pachet (Eds.), *Creativity and universality in language* (pp. 7–26). Springer International Publishing. https://doi.org/10.1007/978-3-319-24403-7_2

Bao, M., Yan, J., Huang, D. (2025). Zipf's law for discourse markers in spoken Mongolian. *Journal of Quantitative Linguistics*, *32*(2), 166–180. https://doi.org/10.1080/09296174.2025.2463754

Chebunin, M., Kovalevskii, A. (2016). Functional central limit theorems for certain statistics in an infinite urn scheme. *Statistics and Probability Letters*, *119*, 344–348. https://doi.org/10.1016/j.spl.2016.08.019

Chebunin, M., Kovalevskii, A. (2019). Asymptotically normal estimators for Zipf's law. *Sankhya A*, *81*, 482–492. https://doi.org/10.1007/s13171-018-0135-9

Davis, V. (2018). Types, tokens, and hapaxes: A new Heap's law. *Glottotheory*, *9*(2), 113–129. https://doi.org/10.1515/glot-2018-0014

Fayzullaev, S., Kovalevskii, A. (2024). Hapax legomena via stochastic processes. *Glottometrics*, *56*, 22–39. https://doi.org/10.53482/2024_56_415

Galieva, A., Vavilova, Z. (2021). Initial and final syllables in tatar: From phonotactics to morphology. *Glottometrics*, *50*, 57–75.

Harrison, P. N. (1921). *The problem of the pastoral epistles*. Oxford University Press.

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, *17*(4), 373–401.

Kudryavtseva, A., Kovalevskii, A. (2025). Comparative statistical analysis of word frequencies in human-written and ai-generated texts. *Glottometrics*, *58*, 19–34. https://doi.org/10.53482/2025_58_423

Mačutek, J., Nogolová, M., Rovenchak, A., Čech, R. (2026). What does the Menzerath-Altmann law really say? *Journal of Quantitative Linguistics*, *33*(1), 28–43. https://doi.org/10.1080/09296174.2025.2545052

Mačutek, J., Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, *20*(3), 227–240. https://doi.org/10.1080/09296174.2013.799912

Milička, J. (2009). Type-token & hapax-token relation: A combinatorial model. *Glottotheory*, *2*(1), 99–110. https://doi.org/10.1515/glot-2009-0009

Petrini, S., Casas-i-Muñoz, A., Cluet-i-Martinell, J., Wang, M., Bentz, C., Ferrer-i-Cancho, R. (2023). Direct and indirect evidence of compression of word lengths. Zipf's law of abbreviation revisited. *Glottometrics*, *54*, 58–87. https://doi.org/10.53482/2023_54_407

Popescu, I.-I., Altmann, G. (2008). Hapax legomena and language typology. *Journal of Quantitative Linguistics*, *15*(4), 370–378. https://doi.org/10.1080/09296170802326699

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Zipf's laws in Italian texts. *Journal of Quantitative Linguistics*, *16*(4), 354–367. https://doi.org/10.1080/09296170903211519

Appendix

Proof of the theorem. First, consider the following expression:

$$B_n = \frac{T_n H_{[n/2]} - H_n T_{[n/2]}}{T_n^2} = \frac{H_{[n/2]}}{T_n} - \frac{H_n}{T_n}\frac{T_{[n/2]}}{T_n} = \frac{\frac{H_{[n/2]}-\eta_{[n/2]}}{\tau_n} + \frac{\eta_{[n/2]}}{\tau_n}}{\frac{T_n-\tau_n}{\tau_n}+1} -$$

$$-\frac{\frac{H_n-\eta_n}{\tau_n}+\frac{\eta_n}{\tau_n}}{\frac{T_n-\tau_n}{\tau_n}+1}\frac{\frac{T_{[n/2]}-\tau_{[n/2]}}{\tau_n}+\frac{\tau_{[n/2]}}{\tau_n}}{\frac{T_n-\tau_n}{\tau_n}+1} =$$

$$= \frac{\frac{Z_n(1/2)}{\sqrt{\tau_n}}+\frac{\eta_{[n/2]}}{\tau_n}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1} - \frac{\frac{Z_n(1)}{\sqrt{\tau_n}}+\frac{\eta_n}{\tau_n}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1}\frac{\frac{Y_n(1/2)}{\sqrt{\tau_n}}+\frac{\tau_{[n/2]}}{\tau_n}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1}$$

From (3) and Lemmas 1 and 2 from Chebunin and Kovalevskii (2019), we obtain the following expression for the mathematical expectations T_n and H_n.

$$\tau_n = \Gamma(1-\theta)c^\theta n^\theta + o(n^{\theta/2}),$$

$$\eta_n = \theta\Gamma(1-\theta)c^\theta n^\theta + o(n^{\theta/2}).$$

This gives:

$$\eta_n/\tau_n = \theta + o(n^{-\theta/2}),$$

$$\eta_{[nt]}/\eta_n = t^\theta + o(n^{-\theta/2}),$$

$$\tau_{[nt]}/\tau_n = t^\theta + o(n^{-\theta/2})$$

as $n \to \infty$, $t > 0$, and taking this into account, let us return to the original expression:

$$B_n = \frac{\frac{Z_n(1/2)}{\sqrt{\tau_n}}+\theta 2^{-\theta}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1} - \frac{\frac{Z_n(1)}{\sqrt{\tau_n}}+\theta}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1}\frac{\frac{Y_n(1/2)}{\sqrt{\tau_n}}+2^{-\theta}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1} + o(n^{-\theta/2})$$

If a random variable satisfies the condition $\xi_n \xrightarrow{p} 0$, then $\frac{1}{1+\xi_n} = 1 - \xi_n + o_p\,(\xi_n)$, and from this:

$$B_n = \left(\frac{Z_n(1/2)}{\sqrt{\tau_n}} + \theta 2^{-\theta}\right)\left(1 - \frac{Y_n(1)}{\sqrt{\tau_n}}\right) -$$

$$-\left(\frac{Z_n(1)}{\sqrt{\tau_n}}+\theta\right)\left(1-\frac{Y_n(1)}{\sqrt{\tau_n}}\right)\left(\frac{Y_n(1/2)}{\sqrt{\tau_n}}+2^{-\theta}\right)\left(1-\frac{Y_n(1)}{\sqrt{\tau_n}}\right) + o_p(n^{-\theta/2}) =$$

$$= \frac{1}{\sqrt{\tau_n}}\left(Z_n(1/2) - 2^{-\theta}Z_n(1) - \theta Y_n(1/2) + \theta 2^{-\theta}Y_n(1)\right) + o_p(n^{-\theta/2})$$

Considering that $V_n = \sqrt{T_n} B_n$, we have

$$V_n = \frac{T_n}{\sqrt{\tau_n}} \Big(Z_n(1/2) - 2^{-\theta} Z_n(1) - \theta Y_n(1/2) + \theta 2^{-\theta} Y_n(1) \Big) + o_p(n^{-\theta/2})$$

Considering the SLLN for T_n and the FCLT of $Z_n(t)$ and $Y_n(t)$ for the cases $t = 1$ and $t = 1/2$, V_n converges weakly to $V = Y_1(1/2) - 2^{-\theta} Y_1(1) - \theta Y(1/2) + \theta 2^{-\theta} Y(1)$ with a mean value of 0. Its variance is calculated based on the covariance matrix of the vector $(Y(1), Y(1/2), Y_1(1), Y_1(1/2))$. The proof is complete.

From ‘said’ to ‘said differently’: modelling repetition versus lexical variation in English-to-Slovak translation of reporting verbs in literary novels

Łukasz Grabowski[1*], Filip Kalaš[2], Daniel Borysowski[1]

[1] University of Opole, Opole, Poland
[2] Bratislava University of Economics and Business, Bratislava, Slovakia
[*] Corresponding author’s email: lukasz@uni.opole.pl

DOI: https://doi.org/10.53482/2026_60_431

ABSTRACT

This corpus-based multifactorial study aims to explore potential predictors of repetition versus lexical variety in translation of repeated reporting verbs from English into Slovak in literary novels. First, we provide a theoretical overview of research on repetition and reporting verbs in Slovak and Czech translation studies. Next, using a sample of 14 literary novels extracted from InterCorp corpus (v.15), we fit multiple negative binomial regression models with mixed effects to assess the effect that several predictor variables (frequency, semantic category, verb length, number of senses, translators) have on the response variable, i.e. the number of Slovak target-text reporting verbs an English source-text (ST) reporting verb is translated into. The findings revealed that factors such as frequency of use of ST reporting verbs, the semantic category of neutral ST reporting verbs, as well as the translators as a random effect, influence Slovak translators’ decisions of using a wide variety of Slovak reporting verbs instead of preserving the originals’ patterns of repetition. More precisely, the model allowed us to explain 70% of the variation (per conditional r-squared) in the response variable. Against the backdrop of prevailing stylistic norms in Slovak, the findings shed light on the translator’s choices in rendering recurrent reporting verbs introducing direct speech, a stylistically salient feature of literary texts.

Keywords: literary translation, parallel corpus, repetition, corpus-based analysis, regression modelling

1 Introduction

Repetition is a common literary device (Klinger, 2019), which affects coherence, or stylistic integrity of the text (Peprník, 1969). The poetic function of repetition (Pataková, 1987) further highlights its aesthetic and structural significance in literary texts: it is often linked to rhythm, emphasis, and emotional depth. All this means that handling repetition is not a trivial task in translation and its treatment varies depending on text type, cultural norms, and translator preferences, among others. For example, Ben-Ari (1998) asserts that translation rules frequently prohibit repetition, particularly in writing, where diversity is encouraged by target language stylistic traditions. Furthermore, the treatment of repetition in translation has implications for equivalence, a central concept in translation studies (Gromová, 2003; Pym, 2009).[1] The two have a complex relationship in that repetition may be either a problem or a strategy for achieving equivalence according to linguistic and cultural norms (Nida and Taber, 1982), e.g. literary translations may require adaptation of repetition to meet target language aesthetics while preserving the stylistic and emotional impact of the original. The notion of equivalence is intrinsically linked to the concept of shifts (cf. Catford, 1965; Zupan, 2006), emerging as a consequence of the translator's interpretive process.[2] These shifts contradict the notion of repetition, but it may happen that the avoidance of repeated lexical items contributes to the flow, readability, and naturalness of a text.

Repetition in translation has been also linked to so-called translation universals such as explicitation and simplification (Baker, 1993, 1996). Blum-Kulka's (1986) explicitation hypothesis implies that translators can manipulate repetitions to enhance textual clarity and render the target text more explicit than the original. Similarly, Hoey's (1991, 2005) lexical priming theory emphasises the cognitive function of repetition in the shaping of readers' expectations and text coherence, and implies that translators' choice of repetition might be conditioned by their own language exposure and stylistic preference. In addition, Slovak translation and stylistic studies, particularly by Miko (1970, 1978) and Popovič (1976), underscore the interaction between textual organisation and stylistic variation, which can possibly throw more light on how repetition is managed in Slovak translations.

The handling of repetition in translated texts has been thoroughly examined by Slovak and Czech translation scholars, who have identified a number of strategies, including synonymization, omission, and structural restructuring. While some contend that in order to maintain rhetorical function, repetition should be preserved in translation, others draw attention to the target language's stylistic conventions, which frequently value stylistic diversity over direct repetition. In other words, some scholars advocate for synonymization and reduction of source text repetitions to improve readability of translated texts (cf. Bečka, 1992; Levý et al., 2011; Nádvorníková, 2020) while others (cf. Augustinská, 1985; Grepl, 1967; Patáková, 1987) argue that repetition serves essential rhetorical, communicative and poetic functions that should be preserved in translation. Consequently, repetition is not only viewed as a linguistic feature but

[1] The Slovak school of translation is grounded in the principle of functional equivalence, which (Vilikovský, 1984, p. 39) defines as follows: "The identity of function serves as a common denominator across all levels of communication, irrespective of specific linguistic expressions. The task of translation is not to replicate the linguistic means themselves but rather to convey the function they fulfill within a broader context".

[2] (Gromová, 2003, p. 44) highlights that shifts in translation are inevitable due to linguistic, cultural, or literary differences between the source and target languages. As (Popovič, 1975, p. 112) asserts, the translator's primary responsibility is to comprehend and interpret the original text. However, interpretation is influenced not only by the source text itself but also by the translator's perspective and other contextual factors, accordingly leading to modification. The expressive structure of the source language cannot be translated to the target language without some degree of transformation or loss. Consequently, translation shifts must be acknowledged as an inherent aspect of the process.

also as a strategic tool in translation, where its preservation or omission can significantly alter the impact of the original text. For instance, the well-known repetition of *I have a dream* is occasionally reduced or altered in Czech translations of Martin Luther King’s *I Have a Dream* (*Mám sen, věřím* – ‘I have a dream, I believe’), which changes its rhythmic and persuasive power. The findings of the study on simplification hypothesis in translated Czech (Cvrček and Chlumská, 2015) suggest that repetition is often reduced or eliminated in translations to enhance clarity and readability. This is especially noticeable in scholarly and journalistic texts, where the intended audience may find excessive repetition to be unnecessary. In a similar vein, the findings of Čermáková’s (2015) corpus study confirm a tendency to neutralize repetition in Czech and Slovak translations. Gresty (2012) also looks at how Slovak writing rules allow for more repetition and verbosity than English: his study focuses on instances when Slovak expressions, such as *On sa naozaj naozaj veľmi snažil* (‘He really, really tried hard’), are translated into English (as *He made a great effort*), indicating a sacrifice of repetition in favour of conciseness. Timing and space limits are also important variables in reducing repetition, notably in audiovisual translation, which presents a major problem in preservation of source text repetitions (Gromová and Janecová, 2013). To preserve brevity, repeated lines, such as *No, no, no!*, are frequently shortened to a single *Nie*! in Slovak dubbing and subtitling, illustrating how technical constraints and translation mode influence translatorial decisions.

However, other scholars advocate preserving repetition as a rhetorical device in translation. Abdulla (2001) contends that repetition, especially in persuasive words, is intentional and significant rather than just redundant. Kundera (1998) argues that translators should maintain all repetitions of words and phrases in the original and strive to avoid using synonyms instead. In a similar vein, noting that rhetorical repetition in speeches and political documents tends to underline important points, Abdulla (2001) criticizes the propensity to misuse synonyms in translation.

Thus, repetition is an important stylistic phenomenon yet it remains relatively underexplored in English-to-Slovak translation. Most of the studies conducted so far are primarily descriptive in nature, which means that based on their findings it is difficult to precisely pinpoint those linguistic factors (predictors) that are responsible for either preservation of source-text repetition or its avoidance, that is, opting for more lexical variety in translation. To this end, our empirical explanatory research is positioned on the interface of corpus stylistics, translation studies, literary translation and multi-factorial statistics. According to Mahlberg (2018), corpus stylistics, which employs corpus linguistic methods (such as wordlist analysis, keyword analysis, and concordance analysis) to develop a more comprehensive description of language use in literary works, has strengthened methodological rigour and offered deeper insights into both frequent and rare features of literary texts. For example, Semino and Short (2004) demonstrated how corpus linguistic approaches can uncover stylistic patterns in literary dialogue, including the role of repetition in constructing narrative voice. As translation is inherently a complex and multifaceted phenomenon (i.e. no single factor, be it linguistic, social or cultural, is solely responsible for particular translatorial decisions), we can currently observe a growing popularity of inferential and multifactorial statistics as well as machine learning methods (e.g., regression models, decision trees, random forest) in research on potential predictors of translatorial choices and, thus, on language use in translation (e.g. De Baets and De Sutter, 2022; De Sutter and Lefer, 2019; De Sutter et al., 2023; Dupont and Zufferey, 2017; Grabowski and Borysowski, 2025; Kajzer-Wietrzny and Ivaska, 2020; Kang and Zhang, 2025; Kruger, 2019; Mastropierro, 2022; Wang and Xin, 2024; Zufferey, 2016). These methods have been successfully used in research on authorship attribution, translatorial attribution, and translator's style (e.g. Eder, 2011; Grieve, 2023; López-Escobedo et al., 2013; Rybicki and Heydel, 2013; Seroussi et al., 2014; Škorić et al., 2022), making them an appealing choice for studying how translators deal with repetitions in source texts.

Thus, our general goal is to identify those factors that impact the ways repeatedly used source-text English reporting verbs that introduce direct speech (i.e. following utterances) are translated into Slovak in selected literary novels. More precisely, we aim to identify the predictors of preservation or avoidance of repetition in translation (see Section 3 of this paper for a detailed description of methodology). Since reporting verbs serve as our unit of analysis, in what follows we present a more detailed overview of their treatment in Slovak and Czech linguistics, including their role and function in translation.

2 Reporting verbs and their treatment in translation: a view from the Slovak and Czech linguistics

In specialized literature, we may find two similar terms related to verbs related to the act of speaking, namely reporting verbs and verba dicendi. The relationship between them can be considered to be hierarchical: in short, all verba dicendi are reporting verbs but not all reporting verbs are verba dicendi. More precisely, reporting verbs are used to introduce reported discourse, whether direct or indirect speech or thought (e.g. *say, tell, report, remark, exclaim, think, believe, assume, admit, warn, beg, know, comment*, and others (Huddleston and Pullum, 2017; Quirk et al., 1985). A common feature of these is their position immediately preceding or following the direct speech. Based on their semantic-pragmatic characteristics and discourse functions, reporting verbs may be further classified into several subgroups. These include verbs of saying, e.g. *say, tell, report, remark, exclaim, comment*, verbs of thinking, e.g. *think, believe, assume, know*, and other verbs that introduce the speaker's stance, emotion, interpersonal positioning, or speech act, e.g. *admit, warn, beg*. However, this typology is not unified as various scholars categorize reporting verbs from different theoretical perspectives (cf. Caldas-Coulthard, 1987; Levin, 1993; Searle, 1976; Thompson and Yiyun, 1991). Among reporting verbs, *say* is regarded as the most frequent and neutral one (Quirk et al., 1985).

In Slovak and Czech linguistics, the Latinized term *verba dicendi* (verbs of saying) has become well established, although native equivalents such as *slovesá hovorenia* (Slovak) and *slovesa mluvení/hovoření* (Czech) are also frequently used. Nemcová (2012) characterizes verba dicendi as a group of verbs defined by a common semantic feature, namely, the description of speech activity.[3] Accordingly, any verb semantically associated with speech is typically categorised among verba dicendi. It should be noted that within the Czech and Slovak linguistic tradition, the distinction between reporting verbs and verba dicendi is often blurred. Research tends to focus predominantly on verbs of saying, while other types of reporting verbs are frequently overlooked. This tendency explains why numerous scholarly articles on verba dicendi are referenced in the present study, as they often encompass broader analysis of reporting verbs despite focusing primarily on speech-related verbs.

Research on verba dicendi, as the most frequently investigated group of reporting verbs, has been extensive in Slovak and Czech linguistics, highlighting the verbs' important roles in semantics (Daneš, 1973; Hirschová, 2017a, 2017b; Knápek, 2019; Nemcová, 2012; Preislerová, 2015; Svobodová, 2007), discourse (Hirschová, 1982, 1988; Šoltys, 1983), syntactic function (Bauer and Grepl, 1972), stylistic variation (Hoffmannová, 2024; Patáková, 1987; Pisárčiková, 1978; Samlerová, 2010) as well as in translation (Fárová, 2016; Fialová, 2020; Nádvorníková, 2017, 2020; Staroňová, 2023). In general, verba dicendi hold particular significance as they form an essential part of reported speech structures. Mistrík (1993, 2021) reinforced this by analysing their role in fictional prose, where verba dicendi bridge narrative elements with direct speech. Mistrík (2021) examined their placement in sentences introducing direct speech, revealing that such structures often contain verba dicendi, such as *povedal* 'said', *vravel* 'said', *odvetil* 'answered' and *spýtal sa* 'asked'. These verbs not only introduce direct speech but also act as synonymous paraphrases of the speech act itself. Thus, their syntactic and pragmatic function is deeply intertwined with situational context rather than explicit linguistic encoding. This feature of verba dicendi is particularly evident in what is referred to as authorial speech, which functions as a connective textual element within prose fiction (Mistrík, 1993). In this context, the most typical representative is the verb *hovoriť/povedať* 'say' in Slovak or *říkat/říci* 'say' in Czech. Additionally, Knápek (2019) noted that while verba dicendi primarily appear in authorial speech, they also frequently occur in direct speech from fictional characters. Knápek's (2019) study of verba dicendi in Karel Čapek's prose revealed that his detective fiction utilizes

[3]While verba dicendi primarily indicate information transmission, they may also overlap with verba sentiendi (e.g., *vidieť* 'see', *počuť* 'hear', *cítiť* 'feel' and verba cogitandi (e.g., *myslieť* 'think', *uvažovať* 'consider'), though these appear less frequently in introductory sentences (Šoltys, 1983). Hoffmannová (2024, p. 132) reaffirmed that verba dicendi emphasize the act of transmission rather than information acquisition. Beyond their primary reporting function, verba dicendi are also subject to metaphorical extensions. Drawing from the SYN2005 corpus of Czech, Samlerová (2010) identified 24 metaphorical extensions of verba dicendi. These include metaphors of animal sounds (*bručet* 'grumble', *kňučet* 'moan''), natural phenomena (*kvílet* 'howl', *zašumět* 'rustle'), speech mannerisms (*mumlat* 'mumble', *žvanit* 'chatter'), and evaluative speech (*kritizovat* 'criticize', *chvástat* se 'brag'). The findings underscore how verba dicendi go beyond their basic reporting function to convey additional semantic nuances through metaphorical associations, which highlights the interplay between semantics and pragmatics.

repetitive, straightforward reporting verbs for clarity, while his fairy tales incorporate archaic and stylized forms to evoke folkloric elements, which highlights how authors strategically deploy verba dicendi to shape narrative tone and reader perception.

Repetition in translation, particularly in the use of reporting verbs, reveals significant linguistic and stylistic differences across languages. One of the most significant differences in reporting verb usage concerns lexical diversity. Vilikovský (1984, p. 215) speaks at this point of the ingrained principle of replacing the stereotypical verbs of the original with a wider repertoire of variations in Slovak.[4] Nádvorníková (2017, 2020) compared the type-token ratio (TTR) of reporting verbs in English, French, and Czech fiction, demonstrating that Czech exhibits the highest degree of lexical variation. While French retains a relatively neutral verb *répondit* (‘answered’), Czech introduces a stylistically richer verb *zamumlal* (‘muttered’), which adds more information about the tone of speech. Similarly, Fialová (2020) analysed German-Czech translations and found that Czech prefers synonymous expansions over direct repetition. The German *sagen* (‘say’) is frequently rendered in Czech with context-sensitive verbs, such as *sdělit* ‘inform’, *upozornit* ‘warn’, *zašeptat* (‘whisper’). Pisárčiková (1978) notes that Slovak literature, even in modern texts, retains a strong preference for expressive reporting verbs, introducing synonyms or restructuring sentences to avoid repetition, e.g. *“To je strašné!” zvolal* – (‘That’s terrible! he exclaimed’). In a corpus-based analysis, Čermáková (2015) showed that translators usually neutralize repetition by replacing repeated words with synonyms or restructuring phrases. For example, the Czech translation of the English phrase *He said, he said, he said* is *“Řekl, poznamenal, dodal”* (‘He said, he remarked, he added’), which lessens the original’s stylistic impact. This result supports Levý’s (1963) theory that Czech and Slovak translators typically choose lexical variation over exact repetition to preserve fluidity. According to Levý’s analysis of Czech translations of Shakespeare, English conversations that contain repeated verbs like *speak, speak, speak* are frequently substituted with other phrases, e.g. *mluv, prav, řekni* ‘speak, say, say’ (Jettmarová, 2008; Schultze, 2015).

Translation studies have also provided valuable insights into the variability of verba dicendi across languages. Staroňová (2023), examining translations between Slovak and English, found that Slovak translators exhibit greater lexical diversity, using semantically rich verbs (i.e. with narrow meaning) to align with dominant literary norms. English translations, conversely, favour simplification, frequently reducing nuanced Slovak reporting verbs to verbs with broad meaning, such as *say*. These findings

[4] According to Vilikovský, the accumulation of interpretive variations can damage the author’s style and frequently have an unintended humorous effect, particularly if rare and emotional verbs are sought as equivalents. Interestingly, writers of the older generation use a wide repertoire of introductory verbs with a low repetition index; modern writers, on the other hand, use a smaller number of verbs with a high repetition index; in more than half of the cases, dialogue is introduced with the verb tell (Vilikovský, 1984, p. 218).

support the asymmetry hypothesis proposed by Klaudy and Károly (2005), which suggests explicitation is more common in translation of reporting verbs than implicitation.

All these findings make reporting verbs introducing direct speech a valuable unit of analysis for investigating broader translation patterns related to dealing with repetition (cf. Mastropierro, 2020, 2022; Mastropierro and Grabowski, 2024). Although we hypothesize that Slovak translators would rather avoid reproduction of repeatedly used reporting verbs introducing direct speech in the English originals (e.g. translating the verb *said* with a wide range of reporting verb equivalents in Slovak), we do not precisely know the rationale behind such decisions. What about reporting verbs with narrower meanings in the English originals, e.g. *answered, replied?* Would they be handled in the same way as the verbs with broader, neutral meanings, like *said or told*? Is the treatment of reporting verbs in translation conditioned by other factors (e.g., word length, the number of senses or the translator's idiolect)? Thus, in this corpus-based multifactorial study we attempt to provide answers to the following research questions: (i) What linguistic factors have a significant effect on the avoidance or reproduction of reporting verbs' repetition in the selected Slovak novels translated from English? (ii) What is the proportion of variance explained by fixed and/or random effects? We hope that our findings will contribute new perspectives to the discussion on repetition in translation into Slovak, enhancing our understanding of this phenomenon. In what follows, we describe the methodology in greater detail.

3 Methodology

The methods used in this study are grounded in corpus linguistics, as we use bilingual concordances extracted from the InterCorp parallel corpus (Čermák and Rosen, 2012) as a research material, and multifactorial statistics. More precisely, we conducted regression modelling in order to identify those linguistic predictors that impact repetition or lexical variety in the English-to-Slovak translation of reporting verbs signalling direct speech (i.e. character's utterances) in literary texts. Using custom-designed bilingual English-Slovak CQL queries, we extracted pairs of aligned reporting verbs from 14 novels (*Winnie the Pooh, The Jungle Book – other, The Jungle Book – Mowgli, the House at Pooh Corner, The Hobbit or There and Back Again, The Hitch Hiker's Guide to the Galaxy, The Fellowship of the Ring, The Da Vinci Code, Harry Potter and the Philosopher's Stone, For Whom the Bell Tolls, Dracula, Catch-22, Alice in Wonderland, 1984*). This selection was based primarily on the availability of texts translated from English into Slovak, found in the InterCorp corpus (ver. 15), which is a large annotated multilingual parallel corpus (Rosen et al., 2022).

As mentioned earlier, reporting verbs in source (ST) and target texts (TT) were selected as the unit of analysis. Therefore, using the CQL queries we searched for the following patterns in the STs: "closing quotation marks + he or she + past tense verb" and "closing quotation marks + past tense verb + he or she"

(Mastropierro and Grabowski, 2024), and these were matched by equivalent patterns in Slovak, taking into consideration specific orthographic conventions of recording dialogues in each language. From the obtained parallel concordances in each novel, we retrieved the translation of each ST verb into a Slovak reporting verb. For the analysis, we used only those ST verbs with a frequency of 2 or higher, as otherwise we would not deal with repetition. This way we retrieved 5,298 verb tokens and 130 verb types in the original novels, as well as all of their translations in Slovak as reporting verbs (530 unlemmatized and 411 lemmatized types). In further analyses, we used lemmatized types in order to ensure compatibility of the English and Slovak data, i.e., the Slovak 3rd person past tense feminine, neutral (in 1 novel) and masculine forms were lemmatized to the single base form. The lemmatization was conducted using lemmagen3, a Python wrapper developed by Podpečan (2024) for Lemmagen lemmatizer (ver. 2.2), which supports 19 languages, including Slovak (Juršic et al., 2010).[5]

All reporting verb types were annotated for six linguistic features that constitute potential predictors of reproduction or avoidance of repetition, starting with the frequency of each ST reporting verb type in each literary novel (e.g. *asked* was used 53 times in the English-original novel “Da Vinci Code” while *yelled* was used 3 times). Due to such considerable differences, we transformed the frequencies into a logarithmic scale and coded the variable as “logfreq”.

The next potential predictor, coded as “wnet_senses_com”, refers to the number of senses of each ST reporting verb, as recorded in the semantic-relational lexical database Princeton WordNet® 3.1 (Fellbaum, 1998), in the semantic domain of communication. This decision was guided by the fact that we focus on reporting verbs following dialogues, that is, introducing direct speech. For example, the verb *asked* has 7 distinct senses in the Princeton WordNet, and 5 of them belong to the domain of communication; in the case of the verb *snapped*, it has 13 distinct senses, and only 2 of them belong to the domain of communication.

The third variable indicates semantic-functional category of the English ST reporting verb based on Caldas-Coulthard’s (1987) categorisation, who distinguished between neutral verbs (e.g. *say, tell*), structuring verbs (e.g. *ask, reply*), metapropositional (e.g. *exclaim, instruct, swear*), metalinguistic (e.g. *narrate, quote*), prosodic (e.g. *cry, shout*), signalling discourse (e.g. *repeat, add*) and paralinguistic verbs, which include voice qualifiers (e.g. *whisper, murmur*) and voice quantification (e.g. *laugh, groan*). Hence, the nominal variable “sem_verb_type” has several levels coded as “N”, “Str”, “Mprop”, “Mlin”, “Pros”, “Sdis”, “Vier”, “Vion” respectively. Consequently, the potential predictors “sem_verb_type” represents the factor with 8 levels. The reason for selecting this typology was, first, that we used the

[5] It is available at the following link: https://pypi.org/project/lemmagen3/.

English reporting verbs as the unit of analysis and, second, that the typology is well-suited for reporting verbs found in literary texts rather than more formulaic genres or text types.

As it is reasonable to assume that repetition of longer words is more likely to be avoided in translation (as the translator is more likely to notice their repetition in a ST), the fourth potential predictor is “verb_length”, measured in characters.

The last factor, which is “translator_id”, indicates a Slovak translator of each novel coded as using acronyms (e.g. “PF” stands for P. Frank, who translated *The Hitch Hiker’s Guide to the Galaxy* into Slovak. All in all, 14 novels were translated by 12 individual translators (J. Samcová translated two texts, namely *The Jungle Book – other* and *The Jungle Book – Mowgli*) and 1 novel, namely *Alice in Wonderland*, was translated by a pair of translators, J. Vojtek and V. Vojtková, coded as JV_VV). This information was retrieved from metafiles available in the InterCorp corpus ver. 15. As we argue that individual translators’ choices impacted the ways in which reporting verbs were translated, and that we analysed only 14 literary novels out of a potentially infinite pool of texts, we decided to include “translator_id” as a random effect in our regression model (cf. Gries, 2015). This also allowed us to meet one of the assumptions for using mixed regression models, where random effects should have at least 5 or 6 levels (Bentz and Winter, 2013), and we have as many as 13 levels.

Hence, the predictors employed in the study include three numerical variables (“log_freq”, “wnet_senses_com” and “verb_length”) and two nominal variables (“sem_verb_type” and “translator_id”) treated as factors with multiple levels, which include both fixed and random effects. The dependent variable is called “type_count”, which is a count variable representing the number of Slovak TT reporting verb types (lemmas) used as translation equivalents of English ST reporting verbs. For instance, the value of “types_count” of 7 indicates that the ST reporting verb was translated into 7 different reporting verbs in Slovak. As an illustration, the verb *cried*, a prosodic verb according to the Caldas-Coulthard (1987) typology, was translated into Slovak as *zvolal* ‘exclaimed’, *kričal* ‘shouted’, *skríkol* ‘screamed’, *zavolal* ‘called’, *ukázal* ‘pointed’, *volal* ‘called’, *vykríkol* ‘cried out’ in the novel *The Hobbit or There and Back Again*. Thus, a value of 1 shows that the verb repetition in the original (ST) was preserved in translation (TT), while a value greater than 1 indicates lexical variety: the higher the value, the wider the range of distinct reporting verbs used in the TT as equivalents of a ST reporting verb. The process of data preparation was conducted using custom-designed Python scripts, and the final data was stored in 14 comma delimited files (csv), corresponding to each novel. These files have been made publicly accessible in an open data repository to ensure replicability and reproducibility of the study.[6]

[6]All the data and metadata, including CQL queries, used in this study is made available in an open data repository under the following link: https://osf.io/t2c6h/.

In order to identify statistically significant predictors of the number of Slovak translation equivalents (reporting verb types) of ST English reporting verbs as well as the predictors’ contribution to explaining variance, we initially intended to use Poisson regression, which is a type of a generalized linear model (GLM) typically employed to model count data and contingency tables, as recommended in specialized literature (Coxe et al., 2009; Scherber, 2017, 2019; Kabacoff, 2015, p. 312; Winter, 2019, p. 247; Winter and Bürkner, 2021, p. 1). The dependent variable “types_count” represents count data measured in non-negative integers, and the predictors in Poisson regression can be a mixture of numerical/continuous and nominal/categorical variables (Kabacoff, 2015, p. 312), which – as previously explained – is the case in this study. However, as we observed overdispersion in the data, that is, the variance for “type” is found to be considerably higher than the mean (mean of 4.65 and variance of 122.50, cf. Bentz and Winter, 2013), we used negative binomial regression instead of Poisson regression, as recommended by Scherber (2017, 2019) and Winter (2019) or Hair et al. (2009). This was later confirmed using the Likelihood Ratio Test (see Section 4). As a rule, the best fitting model is the one with the lowest AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), that is, the one that reaches significance with the fewest variables through their “backward selection” (Winter, 2019, p. 310). This is the type of stepwise regression: we start with a full model with all potential predictor variables and iteratively remove those that are not statistically significant, i.e. have p-values higher than 0.05 (ibid.). As such, our approach and selected methods helped us ensure comparability of our findings with research conducted in English-to-Polish and English-to-Russian language pairs (Grabowski and Borysowski, 2025; Grabowski et al., 2026).

The analyses were conducted in the R environment using the following packages: car (Fox and Weisberg, 2019), MASS (Venables and Ripley, 2002), MuMIn (Bartoń, 2024) and glmmTMB (Brooks et al., 2017). In what follows, we present our findings.

4 Results

In order to assess the influence of all predictor variables, including a random intercept (“translator_id”), on the number of different Slovak reporting verbs used as translation equivalents (variable “types”), we fitted a series of negative binomial regression models using “backward selection” (Winter, 2019, p. 310), which means that statistically insignificant variables (with the p-value higher than 0.05) were iteratively removed. The best fitting model is the one with the lowest AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), that is, the one that reaches significance with as few variables as possible through their “backward selection” (Winter, 2019, p. 310). In order to double check whether we should proceed with a negative binomial regression model, which takes into consideration an additional dispersion parameter, we fitted both Poisson and negative binomial models with all the predictors and run

Likelihood Ratio Test (LRT)[7], which confirmed that the latter one is significantly better (p-value < 0.05, Figure 1). Both AIC and BIC values are lower for the negative binomial regression model (glm1nb), and this improvement in model fit is statistically significant (p-value < 0.05). This finding further confirms overdispersion in the data.

```
Models:
glm1p: types_count ~ logfreq + sem_verb_type + wnet_senses_com + verb_length + , zi=~0, disp=~1
glm1p:     (1 | translator_id), zi=~0, disp=~1
glm1nb: types_count ~ logfreq + sem_verb_type + wnet_senses_com + verb_length + , zi=~0, disp=~1
glm1nb:     (1 | translator_id), zi=~0, disp=~1
       Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
glm1p  13 1083.7 1131.1 -528.86   1057.7
glm1nb 14 1067.0 1118.0 -519.51   1039.0 18.702      1  1.528e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1: Likelihood Ratio Test with nested models: results

Our experiments revealed that an optimal model with the lowest AIC value (1064.7) and BIC value (1108.5) was the following: types_count ~ logfreq + sem_verb_type + $(1 \mid$ translator_id) (Figure 2). It shows that the frequency of a ST reporting verb, semantic category of neutral reporting verbs (e.g. *said*, *told*) as well as the translator as a random effect make the largest individual contributions to explaining the proportion of variation in the response variable “types_count” in the Slovak translations. The model allows us to explain almost 70% of the variation (per conditional r-squared) in the response variable, that is, the number of different Slovak verb types an English ST verb is translated into. This was computed with the help of r.squaredGLLM function in R using a delta method (Bartoń, 2024; Nakagawa et al., 2017). Without taking into consideration “translator_id” treated as a random effect, we would have explained only around 60% of variation, which is 10% less.

As for the predictors’ individual contributions to explaining variance in the response variable “types”, a summary (at the bottom of Figure 2, with p-value and positive or negative estimates indicating the direction and effect size of each predictor, including all levels of variables) reveals that the frequency of an ST reporting verb significantly influences its likelihood of being translated into multiple Slovak TT reporting verbs. For example, with a one-unit change (increase) in “logfreq”, the log of the expected number of translation equivalents increases by 0.56, while holding the rest of the predictor variables constant. For this count, this corresponds to a multiplicative increase of $\exp(0.56) \approx 1.75$ times (or 75% increase), which means that higher frequency has a strong positive effect on the count of “types_count”. This finding confirms our intuition: if a ST reporting verb is frequently used, then translators notice its

[7] We ran anova() function with two nested models in R. LRT is described, for example in Lewis et al. (2011), and it can only be applied in cases where one model is a special case of another (then we deal with so-called nested models), which is the case in this study (negative binomial regression has an additional dispersion parameter).

```
 Family: nbinom2  ( log )
Formula:          types_count ~ logfreq + sem_verb_type + (1 | translator_id)
Data: data

     AIC      BIC   logLik deviance df.resid
  1064.7   1108.5   -520.4   1040.7      271

Random effects:

Conditional model:
 Groups        Name        Variance Std.Dev.
 translator_id (Intercept) 0.07934  0.2817
Number of obs: 283, groups:  translator_id, 11

Dispersion parameter for nbinom2 family (): 24.1

Conditional model:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.08661    0.12813   0.676 0.499112
logfreq              0.56172    0.03662  15.341  < 2e-16 ***
sem_verb_typeN       0.49798    0.14852   3.353 0.000799 ***
sem_verb_typeSdis   -0.16616    0.12950  -1.283 0.199457
sem_verb_typeStr    -0.24816    0.16105  -1.541 0.123339
sem_verb_typeVion   -0.06464    0.13692  -0.472 0.636864
sem_verb_typePros    0.16834    0.12370   1.361 0.173563
sem_verb_typeVier   -0.17722    0.16213  -1.093 0.274377
sem_verb_typeMProp   0.16916    0.47710   0.355 0.722922
sem_verb_typeMlin    0.00878    0.40322   0.022 0.982628
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: Negative binomial regression: final model summary

repetition and in order to avoid reproducing this repetition in translation they resort to using a wide array of TT reporting verbs. In the same vein, neutral verbs (e.g. *said*, *told*), which are broad in meaning and are frequently used as reporting verbs following dialogues, are also consistently translated with a wide range of Slovak translation equivalents, which translates into more lexical variety in translation as compared with the original novels. For example, the verb *said* occurs 131 times following dialogues in the novel *Alice in Wonderland* by Lewis Carroll (1865) and it has 33 lemmatized reporting verb equivalents in the Slovak translation performed by J. Vojtek and V. Vojtková (2004), and 37 unlemmatized types[8]. These findings also accord with a dominant stylistic convention in Slovak (see Section 1), where repetition and lack of lexical variety are synonymous with poor style, as in:

‘As wet as ever,’ ***said*** *Alice in a melancholy tone: ‘it doesn’t seem to dry me at all.’* → *"Taká mokrá ako predtým," smutno* ***odpovedala*** *Alica. "Tie tvoje reči ma akosi nesušia."*

[8]These equivalents are as follows: *povedala* ’she said’, *spýtala sa* ’she asked’, *odpovedala* ’she answered’, *povedal* ’he said’, *namietla* ’she objected’, *odsekla* ’she retorted’, *ozvala sa* ’she responded’, *súhlasila* ’she agreed’, čudovala sa ’she wondered’, bránila sa ’she resisted’, hrešila ’she scolded’, maznala sa ’she caressed’, nepočula ’she didn’t hear’, *oborila sa* ’she snapped’, *obrátil sa* ’he turned around’, *obrátila sa* ’she turned around’, *obzerala sa* ’she looked around’, *ohradila sa* ’she protested’, *opakovala* ’she repeated’, *opýtala sa* ’she asked’, *otriasol sa* ’he shivered’, *ozval sa* ’he responded’, *pokračovala* ’she continued’, *počkala* ’she waited’, *preľakla sa* ’she got scared’, *prisvedčila* ’she nodded’, *prosila* ’she begged’, *riekla* ’she said’, *spýtal sa* ’he asked’, *tíšila* ’she calmed’, *vravela* ’she said’, *vydýchla si* ’she had a rest’, *vyhŕkla* ’she burst out’, *vyčítala si* ’she reproached’, *zahundral* ’he grumbled’, *začala* ’she began’, *zvolala* ’she shouted’. 31 variants are feminine forms and 6 variants are masculine forms.

‘I beg your pardon,’ ***said*** *Alice very humbly: ‘you had got to the fifth bend, I think?’* → *"Prepáč, "****povedala*** *Alica skrúšene. "Ak sa nemýlim, máš už za sebou štyri zákruty."*

‘I wish I hadn’t cried so much!’ ***said*** *Alice, as she swam about, trying to find her way out.* → *"Nemala som toľko plakať,"****vyčítala si*** *Alica, ako plávala a usilovala sa dostať z kaluže.*

‘Poor little thing!’ ***said*** *Alice, in a coaxing tone, and she tried hard to whistle to it.* → *"Drobček môj!"* ***maznala sa*** *s ním, ba chcela naň aj zapískať.*

‘What can all that green stuff be?’ ***said*** *Alice* → *"Čo je to za zeleň?"* ***čudovala*** *sa Alica.*

‘I couldn’t help it,’ ***said*** *Five, in a sulky tone."* → *"Ja za to nemôžem,"* ***zahundral*** *Pätorka.*

Finally, we reported a low level of variance (0.079) and standard deviation (0.28) for the entire model, the values that indicate the spread of random effect. Although the 14 novels used in the study were translated by 12 different translators and one translation team of 2 translators (J. Vojtek & V. Vojtková), the tendencies that the model shows remain independent from individual translator’s habits and choices with respect to rendition of reporting verbs. In other words, there is some variability between the translators but it is rather small (e.g. unlike in the case when variability is in the region of $0.2 - 0.3$): no single translator significantly influenced the observed patterns and the effect was spread across the translators.

5 Discussion and conclusion

The study findings revealed that that factors such as frequency of use as well as a select semantic category (neutral verbs) of ST reporting verbs influence Slovak translators’ decisions of using a wide variety of Slovak reporting verbs, thus avoiding repetition, in the 14 translations of English-original literary novels under scrutiny. These findings, which address the research question (i), partly align with the findings obtained for the English-to-Polish language pair (Grabowski et al., 2026), where broad-meaning neutral reporting verbs (e.g. *said*, *told*), among others, were also found to have been primed for being translated with the whole variety of Polish equivalents, which confirms our intuition. Moreover, our findings indicate that both fixed and random effects should be accounted for when attempting to identify the predictors of translatorial decisions with respect to the translation of reporting verbs. Without including the random intercept, i.e. the translators of the studied English novels into Slovak, we would have explained 10% less variation in the dependent variable. More precisely, the full model explained almost 70% of the variation, which provides an answer to the research question (ii). In an additional experiment, we used individual novels (text_id) instead of the translators as a random intercept and the results did not change. Overall, understanding the factors that lead to either the avoidance or preservation of repetition in translations can be valuable for translator training, for example the findings may inform pedagogical approaches by helping translators recognize patterns in lexical variation and develop strategies for dealing with repetitions in original texts. These insights could also open up opportunities for reflecting upon these findings within the broader context of stylistic conventions in both source and target languages.

This descriptive and explanatory study has a number of limitations, though. As the findings suggest that the ways translators deal with repetition are influenced by the features of repeated items in the original, more predictors could be taken into consideration, e.g. gender and sociolinguistic background of the referent (i.e. literary protagonist), the number of translation equivalents in lexical databases or dictionaries, as well as the span of repetition. Furthermore, a narrative perspective and point of view (first-person vs. third-person narration) could be explored even further, as subjective perspectives may involve using more expressive or evaluative reporting verbs. Other aspects worth considering include cognitive load and translator expertise, notably if the study is extended to interpreting. For example, it might reveal whether the translator's expertise influences the likelihood of lexical diversification or adherence to repetition. On top of that, the emotional intensity of the surrounding discourse could be assessed by applying sentiment analysis, which could help determine whether stronger emotions correlate with greater lexical variation in translation. In this exploratory study, we used a formal equivalence approach as we focused on translating reporting verbs as reporting verbs, but it may also happen that ST reporting verbs are rendered through alternative stylistic means (nominalizations, omissions, periphrastic constructions etc.), which also needs to be addressed in the future. This would provide a broader picture of how lexical repetition is managed beyond verb-to-verb equivalence perspective adopted in this study.

Furthermore, in this research, we focused on selected English-to-Slovak translations of literary novels only, yet comparisons across other text types, genres or modalities could provide a more comprehensive picture of what impacts translatorial decisions with respect to repeated reporting verbs or verba dicendi. For example, Preislerová (2015) examined verba dicendi in Czech fiction and journalism, noting stylistic differences between the two genres.[9] Hoffmannová (2024), who explored the frequency of verba dicendi in contemporary spoken Czech, revealed broader linguistic shifts favouring simplified, frequently used reporting verbs in informal speech.[10] Although both studies were conducted on the Czech language material, it may be tempting to verify if they also apply to Slovak texts. For instance, a diachronic study comparing older and contemporary Slovak translations could reveal whether stylistic norms around repetition have shifted over time. We therefore assume that comparisons across various genres and text types, which could be treated as additional independent variables, could offer more fine-grained insights into the predictors of lexical variety or repetition in translation. Expanding the analysis to other Slavic

[9] Fictional texts use these verbs to enhance narrative depth and provide character insights, often employing evaluative and expressive reporting verbs. In contrast, journalism prioritizes neutral, standardized verba dicendi to maintain factuality and clarity.

[10] Hoffmannová's (2024) findings indicate that Czech verbs *říct/říkat* 'say' dominate spoken usage, while archaic and literary lexemes (e.g., *blahořečit* 'bless', *vyřknout* 'utter') have become largely obsolete.

languages (e.g., South Slavic languages) could additionally determine whether the tendencies in lexical diversification are language-specific or broadly shared across the Slavic language family.

Another future avenue that would allow further verification of the obtained results would be conducting a similar study in the opposite direction, namely in the Slovak-to-English translation. This way we could see whether the tendency of normalization applies in the case of translating Slovak reporting verbs into a narrower range of English reporting verbs, notably neutral ones. From a methodological perspective, integrating neural machine translation or AI-assisted translation (i.e. conducted using selected large language models) into the study could tell whether automated translation systems replicate or diverge from human translator tendencies in handling repetition.

Summing up, irrespective of the aforementioned limitations, this is one of the first multifactorial corpus-based study on translation conducted in the English-to-Slovak language pair with the use of multifactorial methods. As a first step, we hope that this research will inspire broader future investigations into the factors influencing translatorial decisions used when dealing with repeated linguistic items or constructions (lexical items, multi-word units, syntactic structures etc.) in source texts, extending beyond reporting verbs as the unit of analysis, beyond literary texts as a research material, and beyond the language pair or translation direction under scrutiny in this paper.

Acknowledgments

This research was funded by the National Science Centre (NCN), Poland, grant number: 2023/51/B/HS2/00697 (Repetition in translation: a multifactorial descriptive and explanatory study, 2024–2026, grant holder: Łukasz Grabowski).

Data availability statement

The data associated with this research are available in a data repository under the following link: https://osf.io/t2c6h/.

References

Abdulla, A. (2001). Rhetorical repetition in literary translation. *Babel*, 47, pp. 289–303. https://doi.org/10.1075/babel.47.4.02abd

Augustinská, D. (1985). Opakovanie ako výrazový prostriedok vo vecnom texte. *Slovenská reč*, 50, pp. 286–291.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, E. Tognini-Bonelli (Eds.), *Text and Technology: In honour of John Sinclair* (pp. 233–250). John Benjamins.

Baker, M. (1996). Corpus-based translation studies: The Challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and Translation in Language Engineering in Honour of Juan C. Sager* (pp. 175–186). John Benjamins.

Bartoń, K. (2024). *MuMIn: Multi-Model Inference. R package version 1.48.4*. Retrieved January 10, 2025, from https://CRAN.R-project.org/package=MuMIn

Bauer, J., Grepl, M. (1972). *Skladba spisovné češtiny*. Státní pedagogické nakladatelství.

Bečka, J. (1992). *Česká Stylistika*. Academia.

Ben-Ari, N. (1998). The ambivalent case of repetitions in literary translation. *Target*, 10(2), pp. 381–409.

Bentz, C., Winter, B. (2013). Languages with More Second Language Learners Tend to Lose Nominal Case. In S. Wichmann, J. Good (Eds.), *Quantifying Language Dynamics: On the Cutting edge of Areal and Phylogenetic Linguistics* (pp. 96–124). Brill.

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In S. Blum-Kulka, J. House (Eds.), *Interlingual and intercultural communication* (pp. 17–35). Gunter Narr Verlag.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., Bolker, B. M. (2017). Glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2), pp. 378–400.

Caldas-Coulthard, C. R. (1987). Reported speech in written narrative texts. In M. Coulthard (Ed.), *Discussing Discourse* (pp. 149–167). University of Birmingham.

Catford, C., John. (1965). *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. Oxford University Press.

Čermák, F., Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3), pp. 411–427.

Čermáková, A. (2015). Repetition in John Irving's novel a widow for one year: A corpus stylistics approach to literary translation. *International Journal of Corpus Linguistics*, 20(3), pp. 355–377.

Coxe, S., West, S. G., Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, 91(2), pp. 121–136.

Cvrček, V., Chlumská, L. (2015). Simplification in translated Czech: A new approach to type-token ratio. *Russian Linguistics*, 39(3), pp. 309–325.

Daneš, F. (1973). Verba dicendi a výpovědní funkce. *Studia Slavica Pragensia*, pp. 115–124.

De Baets, P., De Sutter, G. (2022). How do translators select among competing (near-) synonyms in translation? A corpus-based approach using random forest modelling. *Target*, 35(1), pp. 1–33.

De Sutter, G., Lefer, M.-A. (2019). On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*, 28(1), pp. 1–23.

De Sutter, G., Lefer, M.-A., Vanroy, B. (2023). Is linguistic decision-making constrained by the same cognitive factors in student and in professional translation? Evidence from subject placement in French-to-Dutch news translation. *International Journal of Learner Corpus Research*, 9(1), pp. 61–96.

Dupont, M., Zufferey, S. (2017). Methodological issues in the use of directional parallel corpora: A case study of English and French concessive connectives. *International journal of corpus linguistics*, 22(2), pp. 270–297.

Eder, M. (2011). Style-markers in authorship attribution: A cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6, pp. 99–114.

Fárová, L. (2016). Uvozovací slovesa v překladech třech různých jazyků. *Jazykové paralely*, pp. 145–161.

Fellbaum, C. (1998). *Wordnet: An electronic lexical database*. MIT press.

Fialová, E. (2020). Deutsche Verba dicendi und ihre tschechischen Entsprechungen. Eine korpusbasierte translatologische Analyse [Master’s thesis, Masaryk University].

Fox, J., Weisberg, S. (2019). *An R companion to applied regression* (3rd). Sage.

Grabowski, L., Borysowski, D. (2025). Between repetition and lexical variety: How do English-to-Russian translators of literary texts deal with recurrent reporting verbs? *Finnish Journal of Linguistics*, 38, pp. 195–210.

Grabowski, L., Olejniczak, J., Mastropierro, L. (2026). A multifactorial study of English-to-Polish translation of reporting verbs in literary novels: A negative binomial regression with mixed effects [In press]. *Research in Corpus Linguistics*, 15(1).

Grepl, M. (1967). O funkci záměrného opakování částí výpovědi ve výstavbě promluvy. *Naše řeč*, 50(2), pp. 77–87.

Gresty, J. (2012). On wordiness in written production of Slovak users of English. In *English matters III* (pp. 25–28). Prešovská univerzita v Prešove.

Gries, S. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), pp. 95–125.

Grieve, J. (2023). Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1), pp. 47–77.

Gromová, E. (2003). *Teória a didaktika prekladu*. Univerzita Konštantína Filozofa.

Gromová, E., Janecová, E. (2013). Audiovisual translation-dubbing and subtitling in Slovakia. *World Literature Studies*, 5(4), pp. 61–71.

Hair, J., Black, W., Babin, B., Anderson, R. (2009). *Multivariate data analysis* (9th). Pearson.

Hirschová, M. (1982). K některým otázkám reprodukování cizích výpovědí. *Acta Universitatis Palackianae, Studia Bohemica II, Facultas Philosophica Philologica*, 46, pp. 97–102.

Hirschová, M. (1988). *Česká verba dicendi v performativním užití: Příspěvek ke zkoumání komunikativních funkcí výpovědi*. Univerzita Palackého.

Hirschová, M. (2017a). *Verbum dicendi*. Retrieved April 21, 2026, from https://www.czechency.org/slovnik/VERBUM%20DICENDI

Hirschová, M. (2017b). *Verbum sentiendi*. Retrieved April 21, 2026, from https://www.czechency.org/slovnik/VERBUM%20SENTIENDI

Hoey, M. (1991). *Patterns of lexis in text*. Oxford University Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.

Hoffmannová, J. (2024). Verba dicendi v současné mluvené češtině. *Prace Filologiczne*, (1), pp. 131–143.

Huddleston, R., Pullum, G., K. (2017). *The Cambridge Grammar of the English Language*. Cambridge University Press.

Jettmarová, Z. (2008). Czech and Slovak translation theories: The lesser-known tradition. In J. Králová, Z. Jettmarová (Eds.), *Tradition versus Modernity. From the Classic Period of the Prague School to Translation Studies at the Beginning of the 21st Century* (pp. 15–46). Charles University in Prague.

Juršic, M., Mozetic, I., Erjavec, T., Lavrac, N. (2010). Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9), pp. 1190–1214.

Kabacoff, R. (2015). *R in action: Data analysis and graphics with R*. Manning Publications Co.

Kajzer-Wietrzny, M., Ivaska, I. (2020). A multivariate approach to lexical diversity in constrained language. *Across Languages and Cultures*, 21(2), pp. 169–194.

Kang, H., Zhang, Y. (2025). Socio-cognitive influences on translating forward causal connectives: A multivariate analysis of English translations of Tao Te Ching. *Lingua*, 314, p. 103872.

Klaudy, K., Károly, K. (2005). Implicitation in translation: Empirical evidence for operational asymmetry in translation. *Across Languages and Cultures*, 6(1), pp. 13–28.

Klinger, S. (2019). Repetition: Translating the interplay between its linguistic form and its literary function. *Babel*, 65(2), pp. 316–332.

Knápek, D. (2019). Funkce a frekvence verb dicendi ve vybraných prózách Karla Čapka [Bachelor’s thesis]. Palacky University in Olomouc.

Kruger, H. (2019). That again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English. *Across Languages and Cultures*, 20(1), pp. 1–33.

Kundera, M. (1998). *The Art of the Novel*. HarperCollins.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Levý, J. (1963). *Uměni překladu*. Československý spisovatel.

Levý, J., Corness, P., Jettmarová, Z. (2011). *The Art of Translation*. John Benjamins.

Lewis, F., Butler, A., Gilbert, L. (2011). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, 2(2), pp. 155–162.

López-Escobedo, F., Méndez-Cruz, C.-F., Sierra, G., Solórzano-Soto, J. (2013). Analysis of stylometric variables in long and short texts. *Procedia-Social and Behavioral Sciences*, 95, pp. 604–611.

Mahlberg, M. (2018). Corpus Stylistics. In V. Sotirova (Ed.), *The Bloomsbury Companion to Stylistics* (pp. 139–156). Bloomsbury.

Mastropierro, L. (2020). The translation of reporting verbs in Italian: The case of the Harry Potter series. *International Journal of Corpus Linguistics*, 25(3), pp. 241–269.

Mastropierro, L. (2022). The avoidance of repetition in translation: A multifactorial study of repeated reporting verbs in the Italian translation of the Harry Potter series. In *Advances in corpus applications in literary and translation studies* (pp. 138–157). Routledge.

Mastropierro, L., Grabowski, Ł. (2024). Repeated reporting verbs in English novels and their Italian and Polish translations: A preliminary multifactorial study. *Across Languages and Cultures*, 25(2), pp. 310–330.

Miko, F. (1970). Textová výstavba literárneho diela. *Slovenská literatúra*, 17(1), pp. 3–17.

Miko, F. (1978). *Štylistika a textová lingvistika*. SPN.

Mistrík, J. (1993). *Encyklopédia jazykovedy*. Obzor.

Mistrík, J. (2021). *Štylistika*. VEDA.

Nádvorníková, O. (2017). Les proportions des verbes say/dire/říci dans les propositions incises et leurs équivalents en traduction: Étude sur corpus parallèle. *Linguistica Pragensia*, 28(2), pp. 35–57.

Nádvorníková, O. (2020). Differences in the lexical variation of reporting verbs in French, English and Czech fiction and their impact on translation. *Languages in Contrast*, 20(2), pp. 209–234.

Nakagawa, S., Johnson, P. C., Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134), p. 20170213.

Nemcová, E. (2012). Niekol'ko poznámok k spracovaniu verb dicendi. In K. Buzássyová, B. Chocholová, N. Janočková (Eds.), *Slovo v slovníku Aspekty lexikálnej sémantiky – gramatika – štylistika (pragmatika) Na počesť Alexandry Jarošove* (pp. 192–208). VEDA.

Nida, E., Taber, C. (1982). *The theory and practice of translation*. Brill.

Patáková, M. (1987). Poetizačná a kontaktová funkcia opakovania v autorskej reči. *Slovenská reč*, 52(6), pp. 321–334.

Peprník, J. (1969). Reporting phrases in English prose. *Brno Studies in English*, 8(1), pp. 145–151.

Pisárčiková, M. (1978). Synonymia slovies v uvádzacích vetách. *Slovenská reč*, 43, pp. 210–216.

Podpečan, V. (2024). *Lemmagen3. Python library.* Retrieved December 3, 2024, from https://pypi.org/project/lemmagen3/

Popovič, A. (1975). *Teória umeleckého prekladu: Aspekty textu a literárnej metakomunikácie*. Tatran.

Popovič, A. (1976). *A dictionary for the analysis of literary translation*. University of Alberta Press.

Preislerová, A. (2015). Verba dicendi jako metapragmatický komentář. Slovesa mluvení v uvozovacích větách v beletrii a publicistice na materiálu SYN2010 [Master's thesis, Palacky University].

Pym, A. (2009). *Exploring translation theories*. Routledge.

Quirk, R., Greenbauch, S., Leech, G., Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Rosen, A., Vavřín, M., Zasina, A. J. (2022). *The InterCorp Corpus – Czech, version 15 of 11 November 2022*. https://kontext.korpus.cz/

Rybicki, J., Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's night and day in polish. *Literary and Linguistic Computing*, 28(4), pp. 708–717.

Samlerová, L. (2010). Metaforický charakter verb dicendi [Master's thesis, Technická univerzita v Liberci.].

Scherber, C. (2017). Using R to interpret interaction effects in statistical models. *Software Developer's Journal*. https://www.researchgate.net/profile/Christoph_Scherber/publication/312093784_Using_R_to_Interpret_Interaction_Effects_in_Statistical_Models/links/586f67ad08ae329d6215fc4c/Using-R-to-Interpret-Interaction-Effects-in-Statistical-Models.pdf

Scherber, C. (2019). *An introduction to generalized linear models*. http://www.christoph-scherber.de/content/PDF%20Files/Generalized%20linear%20models.pdf

Schultze, B. (2015). Jiří Levý's contribution to translation studies as represented in the De Gruyter Encyclopedia Übersetzung, Translation, Traduction. *Acta Universitatis Carolinae Philologica*, 3, pp. 105–112.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society*, 5(1), pp. 1–23.

Semino, E., Short, M. (2004). *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. Routledge.

Seroussi, Y., Zukerman, I., Bohnert, F. (2014). Authorship attribution with topic models. *Computational Linguistics*, 40(2), pp. 269–310.

Škorić, M., Stanković, R., Ikonić Nešić, M., Byszuk, J., Eder, M. (2022). Parallel stylometric document embeddings with deep learning based language models in literary authorship attribution. *Mathematics*, 10(5), p. 838.

Šoltys, O. (1983). *Verba dicendi a metajazyková informace*. Československá akademie věd, Ústav pro jazyk český.

Staroňová, K. (2023). Operational Asymmetry in Translation of English and Slovak Reporting Verbs [Master's thesis, Masaryk University].

Svobodová, J. (2007). Sémantika sloves mluvení v uvozovacích větách uměleckého textu. *Studia Bohemica*, 10, pp. 153–160.

Thompson, G., Yiyun, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied linguistics*, 12(4), pp. 365–382.

Venables, B., Ripley, B. (2002). *Modern Applied Statistics with S* (4th). Springer.

Vilikovský, J. (1984). *Preklad ako tvorba*. Slovenský spisovateľ.

Wang, G., Xin, Y. (2024). An analytical framework for corpus-based translation studies. *Humanities and Social Sciences Communications*, 11(1), p. 1709.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Winter, B., Bürkner, P.-C. (2021). Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass*, 15(11), e12439.

Zufferey, S. (2016). Discourse connectives across languages: Factors influencing their explicit or implicit translation. *Languages in Contrast*, 16(2), pp. 264–279.

Zupan, S. (2006). Repetition and translation shifts. *ELOPE: English Language Overseas Perspectives and Enquiries*, 3(1–2), pp. 257–268.

Comparative analysis of linguistic features and machine learning methods in the task of assessing the complexity of texts

Artem Zaikin[1*], Valery Solovyev[1], Marina Solnyshkina[1]

[1] Kazan Federal University, Kazan, Russian Federation
[*] Corresponding author's email: kaskrin@gmail.com

DOI: https://doi.org/10.53482/2026_60_432

ABSTRACT

Text complexity assessment is an important applied problem that lacks a comprehensive solution. Studies vary in the text corpora used, the features analyzed, the analysis algorithms applied, and the assessment methods employed. We present a new methodology for complexity assessment, developed based on a representative collection of Russian-language school textbooks compiled by the authors. Textual complexity is represented numerically by a textbook's target class grade. First, we evaluate the average number of words per sentence and the average number of syllables per word for their ability to predict text complexity and implement machine learning methods, including linear regression and neural networks. We confirm the linear Flesch-Kincaid complexity formula in the sense that it cannot be improved upon. The influence of text segmentation on the assessment results is also examined, and several approaches to utilizing such information are presented in the paper. This approach gave us approximately 10% improvement in the R^2 metric compared to the baseline model. We also examine a total of 47 linguistic parameters as predictors of complexity and evaluate their significance.

Keywords: text complexity, linguistic features, regression, machine learning, school textbooks

1 Introduction

The need for an objective assessment of text complexity to ensure adequate reader understanding was recognized over 80 years ago. The first mathematical readability formula, known as the Flesch Readability Formula, was proposed for English in Flesch (1948). It was later refined into the Flesch–Kincaid Grade Level formula (Kincaid et al., 1975). The formula is expressed as $FKG(ASL, ASW) = 0.39ASL + 11.8ASW - 15.59$, where ASL denotes the average sentence length (in words) and ASW represents the average word length (in syllables). The resulting numerical score approximately corresponds to the school grade level required to understand the text. The parameters ASL and ASW were not chosen arbitrarily; they are significant predictors of complexity and have since been utilized in various formulas and approaches. Subsequently, similar formulas were proposed for other languages, incorporating additional linguistic parameters. For example, Mikk (1974) analyzed Estonian texts and introduced a new parameter: the degree of word abstractness.

Naturally, each language requires its own readability formulas. In this article, we study the complexity of texts in Russian. For Russian, the first formula was proposed by Matskovskiy (1976): $X_1 = 0.62X_2 + 0.123X_3 + 0.051$, where X_2 is the average sentence length (in words) and X_3 is the percentage of words with more than three syllables. Specialized formulas for different genres and subject areas were also proposed. Shpakovskiy et al. (2007) proposed a formula for chemical texts that accounts for specialized terms and symbols. Oborneva (2006) proposed the formula $FKG_{Rus}(ASL, ASW) = 0.5ASL + 8.4ASW - 15.59$ using the same variables. This formula is well-suited for fiction, as it was derived from the English Flesch–Kincaid formula by adapting the coefficients to account for linguistic differences, based on a corpus of fiction texts. Furthermore, Solovyev et al. (2018) demonstrated that this formula yields unreasonably high scores for school textbooks and proposed a genre-specific formula: $FKG_{sis}(ASL, ASW) = 0.36ASL + 5.76ASW - 11.97$. The mean squared error is approximately 1.02, corresponding to one grade level. A limitation of this work was that the formula was derived from a small collection of only 14 textbooks for grades 5–11, all within a single subject area: social studies. Later attempts were made to improve this formula Solovyev et al. (2022) while maintaining a linear model. Solnyshkina et al. (2018) made the only known attempt to derive a quadratic formula. These works achieved a slight increase in accuracy compared to the simple FKG_{sis} formula, though the improvement was insignificant. Morozov et al. (2022) considered a coarser classification into five grade groups: 1–2, 3–4, 5–7, 8–9, and 10–11. An F-measure of 80% was achieved for fiction texts.

One independent area of research focuses on assessing the complexity of individual words. This field is known as Complex Word Identification or lexical complexity prediction (Yimam et al., 2018). The complexity of word combinations has also been studied (Feng and Yu, 2024). For Russian, several studies have employed neural network approaches in this domain (Abramov and Ivanov, 2022; Abramov et al., 2023). The results are comparable to those for English, though slightly lower.

A number of studies have focused on texts for teaching Russian as a foreign language. Using the generally accepted division of instructional texts into six difficulty levels, researchers have achieved 60% accuracy in complexity assessment (Aleksandrovich, 2022). Corlatescu et al. (2022) propose limiting the classification to only two difficulty levels; with this binary approach, accuracy reaches 92%. Almost all studies employ either linear models with multiple features or neural architectures such as BERT. Significant variations across studies in terms of text collections, feature sets, and analysis methods make it difficult to directly compare results. A comprehensive recent review of research on text complexity can be found in Solnyshkina et al. (2022).

Thus, questions regarding how complexity should be represented, and whether it depends linearly on average sentence length and average word length, remain relevant.

This study aims to analyze a new corpus of school textbook texts for grades 2–11. We evaluate both established methods and explore the application of novel approaches. The dataset is described in Section 2.

The following questions were considered within the framework of the work:

1. How well does linear regression based on two "classical" features – average sentence length and average word length – assess textbook complexity?
2. Is it possible to improve the assessment of textbook complexity based on the "classical" characteristics using more advanced models?
3. Is it possible to improve the assessment of the complexity of the entire textbook based on knowledge of its parts?
4. To what extent can the assessment of textbook complexity be improved knowing all the other characteristics?
5. Which characteristics are the most significant, and how does complexity depend on them?

2 Data

We compiled a corpus of texts from Russian school textbooks for grades 2 through 11. The Russian school system comprises 11 grades; first grade was excluded because the texts are too simple for meaningful complexity analysis. Third-grade textbooks were also excluded due to technical constraints.

The corpus covers four subject areas: humanities (history, social studies), natural sciences (physics, biology, "The world around us", ecology, geography), mathematical sciences (mathematics, computer science), and language arts (Russian language, literature). Chemistry textbooks were excluded because chemical formulas present significant preprocessing challenges.

All textbooks underwent a preprocessing stage prior to analysis: illustrations and ancillary content (e.g., page headers, publisher information, exercise labels) were removed, and obvious typographical errors were corrected.

We identified 47 linguistic features, listed in Appendix 1. The feature set encompasses descriptive, morphological, syntactic, lexical, and discursive categories. Most features previously shown to be informative for readability assessment are included in this list.

To compute feature values, we developed the RuLingva software package (https://rulingva.kpfu.ru/), which is described in Solnyshkina et al. (2024). RuLingva employs the Natasha morphological analyzer (https://github.com/natasha/natasha) for lemma extraction, part-of-speech tagging, and other preprocessing tasks.

We note the following regarding data structure. Textbooks are segmented into chunks of approximately equal length, and features are computed for each chunk. The feature representation of a textbook is defined as the average of its chunk-level values. In this hierarchical structure, the complexity label is assigned to the entire textbook rather than to individual chunks. However, in some subsequent analyses, complexity labels will also be associated with chunks for methodological purposes.

The initial dataset comprised 16,596 text chunks extracted from 206 textbooks, each described by 47 features (excluding token count). During preprocessing, chunks containing fewer than 500 tokens were removed, as typical chunks contain 700–800 tokens; such short chunks were treated as statistical outliers. After filtering, 16,467 chunks remained for analysis.

3 Applied methods for comparing models

Here we present the model evaluation metrics. Let y_n denote the observed complexity class for the n-th observation, and $\widehat{y}_n$ the predicted complexity class. Let N be the total number of observations. For each prediction method, we calculate the following metrics:

- **Exact-match accuracy**

  (1) $$R_0 = \frac{1}{N}\sum_n I(y_n = \widehat{y}_n),$$

  where $I(\cdot)$ is the indicator function.

- **Normalized absolute deviation score**

  (2) $$R_1 = 1 - \frac{\sum_n |y_n - \widehat{y}_n|}{\sum_n |y_n - y_{\text{med}}|},$$

  where y_{med} is the sample median of $y_1, \ldots, y_N$.

- **Coefficient of determination (R^2)**

  (3) $$R_2 = 1 - \frac{\sum_n (y_n - \widehat{y}_n)^2}{\sum_n (y_n - \overline{y})^2},$$

  where $\overline{y}$ is the sample mean.

The closer each score is to one, the better the model performance. To reduce variance and mitigate overfitting, all metrics are computed using the same cross-validation splits.

Although we report all metrics for each model, we consider R_1 to be the primary evaluation criterion, as it is best suited to our task. The R_0 metric is overly sensitive to small estimation errors (treating a prediction of grade 5 as equally wrong whether the true label is 4 or 10), while R_2 relies heavily on the assumption that the numerical encoding of complexity classes reflects meaningful ordinal distances between them.

Note that R_1 (analogous to the Nash–Sutcliffe efficiency coefficient) is scale-invariant and interpretable as the proportion of variance explained relative to a median baseline, making it particularly suitable for ordinal regression tasks where absolute error magnitude matters more than exact class matching.

There is no fundamental difference between the R_2 score and the mean squared error, $\mathrm{MSE} = N^{-1} \sum_n (y_n - \widehat{y}_n)^2$, nor between R_1 and the mean absolute error, $\mathrm{MAE} = N^{-1} \sum_n |y_n - \widehat{y}_n|$. Naturally, the model with the highest R_1 will exhibit the lowest MAE among a given set of models. The R_1 score expresses prediction error as a proportion of the total variation in the data, whereas MAE reports the absolute magnitude of prediction error. Since both interpretations are informative, we include MSE and MAE values for selected models.

4 Classical linear model

Here we investigate the linear relationship between complexity class and two numerical features: average sentence length (in words) and average word length (in syllables). The distributions of these features across complexity classes are shown in Figure 1. The relationship between average sentence length and complexity class appears more clearly linear than that for average word length. Moreover, average word length does not exhibit a monotonic relationship with class. Consequently, the regression coefficient for word length is smaller than those reported in previous works (e.g., Matskovskiy, 1976; Oborneva, 2006; Solovyev et al., 2018).

We model this relationship using both simple linear regression and ordinal regression approaches. Simple linear regression is fitted by least squares with complexity class as the target variable. Both regression coefficients are significantly positive, as summarized in Table 1.[1]

During prediction, we round the continuous regression output to the nearest valid class value. To assess model performance robustly, we apply 8-fold cross-validation. Results are presented as a contingency table in Table 2. From this table, we compute the following metrics: $R_0 \approx 0.276$, $R_1 \approx 0.4$, $R_2 \approx 0.554$, corresponding to $\mathrm{MSE} \approx 2.3$ and $\mathrm{MAE} \approx 1.15$.

[1] Thus, the final dependence could be represented via formula:

$$\begin{aligned}\text{Complexity class} = &-5.299 + 0.512 \cdot \text{Average number of words per sentence}\\ &+ 2.466 \cdot \text{Average number of syllables per word.}\end{aligned}$$

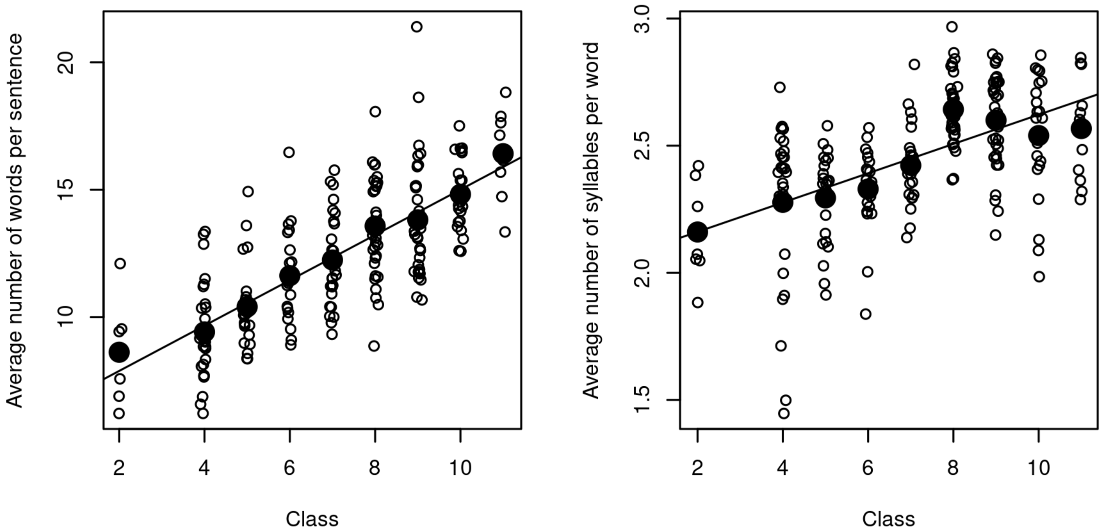


Figure 1: Average sentence length and average word length by complexity class. Each point represents a single textbook; class labels have been jittered slightly to reduce overplotting. Solid points indicate class-wise means. Lines represent least-squares fits with complexity class as the explanatory variable. Note that our analysis models the inverse relationship (complexity class as a function of linguistic features), so these lines should not be interpreted as the final predictive model.

Table 1: Simple linear regression coefficients fit by least squares.

Variable	Estimate	Std. err.	t value	P(> |t|)
(Intercept)	-5.299	0.996	-5.318	0
Average number of words per sentence	0.512	0.04	12.843	0
Average number of syllables per word	2.466	0.439	5.624	0

Table 2: Contingency table for simple linear regression. Rows represent true values, and columns represent estimates.

	2	4	5	6	7	8	9	10	11
2	1	3	2	1	0	0	0	0	0
4	3	2	12	5	3	1	0	0	0
5	0	0	12	9	1	2	0	0	0
6	0	0	2	9	6	4	0	0	0
7	0	0	1	8	7	8	2	0	0
8	0	0	1	1	5	13	13	1	1
9	0	0	0	4	7	11	7	3	2
10	0	0	0	2	1	9	5	5	0
11	0	0	0	0	1	2	2	5	1

Since complexity class is an ordered variable, ordinal regression is a natural modeling choice (Winship and Mare, 1984). Ordinal regression accounts for the ordered nature of complexity classes by modeling cumulative probabilities, rather than treating classes as nominal categories or continuous values.

We apply ordinal logistic regression (proportional odds model). Let X denote a vector of explanatory variables, θ the corresponding vector of regression coefficients, and K the maximum number of classes.

The model is defined as

$$\mathbf{P}(y \le k \mid X) = \sigma\left(\eta_k - X^\top \theta\right), \quad \sigma(z) = \frac{1}{1+\exp(-z)}, \quad k = 1, \dots, K-1,$$

where $\eta_1, \dots, \eta_{K-1}$ are cut-point parameters that, together with θ, are estimated via maximum likelihood. The predicted class is the one with the highest predicted probability.

Results for ordinal regression are presented in Table 3. The metric values are $R_0 \approx 0.247$, $R_1 \approx 0.382$, $R_2 \approx 0.546$. In terms of predictive accuracy, ordinal regression performs nearly identically to simple linear regression; however, it additionally provides class-probability estimates for each textbook. An example of such probabilistic predictions is shown in Figure 2. This figure also suggests that the complexity class estimation task is inherently uncertain, with substantial probability mass often spread across adjacent grades.

Table 3: Contingency table for ordinal linear regression. Rows represent true values, and columns represent estimates.

	2	4	5	6	7	8	9	10	11
2	0	6	1	0	0	0	0	0	0
4	3	12	7	0	2	2	0	0	0
5	0	12	4	2	4	1	1	0	0
6	0	2	7	0	7	5	0	0	0
7	0	1	7	0	6	8	4	0	0
8	0	1	0	0	5	12	14	2	1
9	0	0	1	1	5	10	9	6	2
10	0	0	0	2	1	6	7	5	1
11	0	0	0	0	0	2	2	4	3

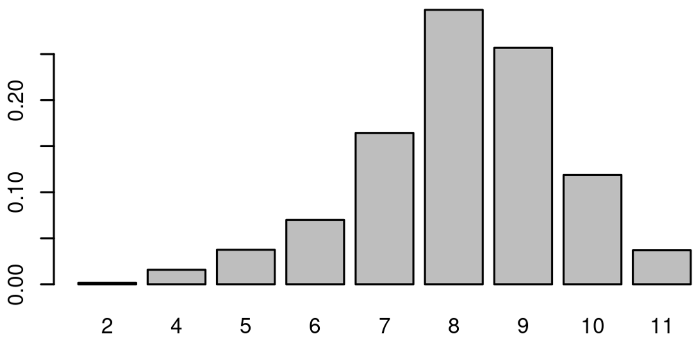


Figure 2: Predicted class probabilities for a randomly selected textbook, as estimated by ordinal regression. The true complexity class is 10 in this example.

5 Model refinement

We now evaluate additional regression methods on the same data and compare them to the baseline models. Let x_1 and x_2 denote the explanatory variables (average sentence length and average word length, respectively). All models are assessed using the R_0, R_1, and R_2 metrics computed via cross-validation. Additionally, each model is trained in two variants: on the textbook-level dataset and on the chunk-level dataset. The models considered are:

1. **Simple linear regression.**
2. **Ordinal logistic regression.**
3. **Additive regression** (Wood, 2011). This model extends simple linear regression by replacing the linear predictor $\theta_1 x_1 + \theta_2 x_2$ with an additive nonlinear specification $f_1(x_1) + f_2(x_2)$, where f_1 and f_2 are smooth functions represented by cubic splines. The model is fitted via penalized least squares, with smoothing parameters selected by generalized cross-validation (GCV).
4. **Ordinal additive regression.** This model combines the ordinal regression framework with the additive spline specification described above.
5. **Tensor-product spline regression.** This model extends additive regression by including a bivariate interaction term $f_{1,2}(x_1, x_2)$, parameterized as a cubic spline surface. The function space for such surfaces is constructed via the tensor product of univariate cubic spline bases.
6. **Ordinal tensor-product spline regression.**
7. **Gradient boosting for regression.** We use the XGBoost implementation (Chen and Guestrin, 2016).
8. **Gradient boosting for classification.** We use the XGBoost implementation (Chen and Guestrin, 2016).
9. **Feedforward neural network for regression.** Hyperparameters were selected via cross-validation. The final architecture comprises two hidden layers with 32 units each, ReLU activation, no dropout, and the Adam optimizer with a learning rate of 0.003. Training was terminated via early stopping on a validation set comprising approximately 20% of the data.
10. **Feedforward neural network for classification.** This model is nearly identical to the regression variant, but employs a softmax output layer for class-probability estimation.

To ensure consistent evaluation across models, all continuous predictions are rounded to the nearest valid complexity class.

Model comparison results are presented in Table 4. We conclude that none of the tested methods substantially outperform simple linear regression. The best-performing model was additive regression. Figure 3 shows the estimated component functions f_1 and f_2 from this model.

A further observation is that models trained on chunk-level data consistently yield poorer performance than those trained on textbook-level aggregates. More complex models, such as gradient boosting and neural networks, performed considerably worse than the baseline.

Table 4: Comparison of modeling approaches using classical features (average sentence length and average word length). All metrics are computed using the same book-level cross-validation splits.

Method	R_0	R_1	R_2
Simple linear regression	0.272	0.408	0.569
Additive regression	0.277	**0.418**	**0.574**
Ordinal linear regression	0.272	0.405	0.566
Ordinal additive regression	**0.296**	0.405	0.538
Simple linear regression, chunks	0.189	0.086	0.069
Additive regression, chunks	0.209	0.106	0.074
Ordinal linear regression, chunks	0.136	-0.086	-0.366
Ordinal additive regression, chunks	0.131	-0.104	-0.414
Tensor-product spline regression	0.282	0.413	0.572
Ordinal tensor-product spline regression	0.248	0.375	0.526
Tensor-product spline regression, chunks	0.204	0.094	0.049
Ordinal tensor-product spline regression, chunks	0.136	-0.124	-0.461
Neural network	0.034	-1.686	-5.125
Classification neural network	0.262	0.347	0.452
Neural network, chunks	0.034	-1.686	-5.125
Classification neural network, chunks	0.117	-0.197	-0.584
Gradient boosting	0.262	0.344	0.466
Classification gradient boosting	0.286	0.316	0.375
Gradient boosting, chunks	0.286	0.316	0.375
Classification gradient boosting, chunks	0.286	0.316	0.375

6 Leveraging the hierarchical data structure

Here we explore methods that exploit the hierarchical structure of our data. In this setting, predictions for a textbook can incorporate information from all its constituent chunks. We continue to restrict our analysis to the classical feature set: average sentence length and average word length.

We evaluate the following approaches:

1. **Cascade modeling.** This approach requires a base regression method (such as linear or additive regression) and operates as follows. First, the base model is trained on textbook-level data. This model is then used to generate predictions for each chunk within a textbook. Finally, chunk-level

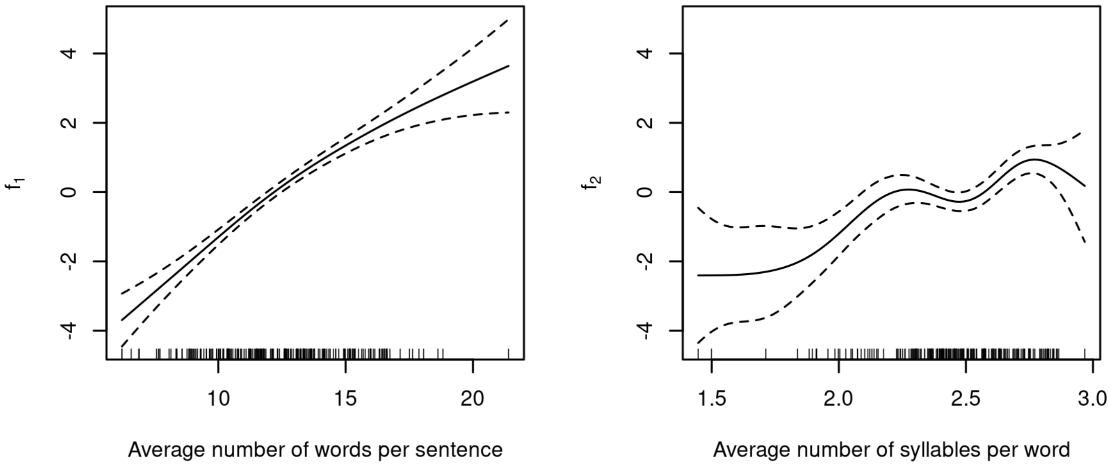


Figure 3: Estimated component functions f_1 (average sentence length) and f_2 (average word length) from the additive regression model predicting complexity class. Dashed lines indicate approximate 95% confidence intervals.

predictions are aggregated to produce a textbook-level prediction. Aggregation is implemented in two variants:

- *Simple averaging:* the arithmetic mean of chunk-level point predictions.
- *Probabilistic averaging:* applicable when the base model outputs class probabilities; probabilities are averaged across chunks before selecting the final class.

Base methods are some of those listed in the previous section.

2. **Covariance augmentation.** For each textbook, we compute the sample covariance matrix from its chunk-level features. The variance and covariance terms are then appended to the feature vector. With two classical features, this adds only three variables. Regression models from the previous section are then applied to this augmented feature set.

3. **Generative hierarchical model.** Let X denote the two-dimensional feature vector for a single chunk. We assume that, conditional on the complexity class y, X follows a bivariate normal distribution with mean θ and covariance matrix Λ. The mean parameter θ is itself a random vector, shared across all chunks from the same textbook, with distribution $\theta \mid y \sim \mathcal{N}(\mu(y), \kappa(y)^{-1}\Lambda)$. The covariance matrix Λ follows an inverse Wishart distribution with parameters $\nu(y)$ and $\Sigma(y)$. Class priors are specified as $\mathbf{P}(y = k) = \pi_k$. Prediction for a textbook is based on the posterior probabilities $\mathbf{P}(y = k \mid X_1, \ldots, X_n)$, where n is the number of chunks. We consider several model specifications: a uniform prior over classes and a homoscedastic variant where $\nu(y)$ and $\Sigma(y)$ are constant across classes. Parameter estimation, predictive inference, and derivations are provided in the Appendix 2.

The normality assumption underlying the generative model is motivated by visual inspection of chunk-level scatter plots. Figure 4 suggests this assumption is plausible. A further consequence of normality is that the mean vector and covariance matrix are sufficient statistics for the chunk distribution, rendering additional summary statistics redundant. However, formal multivariate normality tests (specifically, Royston's test; Royston, 1983) indicate that chunk-level data deviate from normality for approximately 70% of textbooks at the 0.1 significance level. Nevertheless, we retain the normality assumption to obtain a simple, interpretable model with closed-form solutions for parameter estimation.

Results are presented in Table 5. We find that cascade methods substantially improve the predictive performance of base models. Incorporating covariance information yields further gains. The generative model performs comparably to cascade approaches. Figure 5 displays the component functions from an additive regression model fitted on textbook-level data augmented with variance and covariance features. Although the generative model also leverages covariance information, it appears to do so in a less flexible manner, resulting in comparatively lower metric scores.

Table 5: Comparison of modeling approaches leveraging the hierarchical data structure and classical features. All metrics are computed using the same textbook-level cross-validation splits.

Method	R_0	R_1	R_2
Cascade, linear regression	0.272	0.403	0.559
Cascade, additive regression	0.248	0.403	0.574
Cascade, ordinal regression	0.228	0.357	0.516
Cascade, ordinal additive regression	0.267	0.377	0.499
Cascade, neural network	0.296	0.42	0.56
Cascade, classification neural network	0.272	0.314	0.373
Cascade, gradient boosting	0.267	0.347	0.475
Cascade, classification gradient boosting	0.262	0.311	0.387
Generative model, homoscedasticity, flat prior	0.282	0.311	0.407
Generative model, heteroscedasticity, flat prior	0.296	0.309	0.35
Generative model, homoscedasticity, variate prior	0.252	0.365	0.508
Generative model, heteroscedasticity, variate prior	0.282	0.39	0.53
Covariance augmentation, linear regression	0.286	0.433	0.604
Covariance augmentation, additive regression	**0.325**	**0.461**	**0.622**
Covariance augmentation, ordinal regression	0.272	0.441	0.618
Covariance augmentation, ordinal additive regression	0.301	0.446	0.613
Covariance augmentation, neural network	0.034	-1.686	-5.125
Covariance augmentation, gradient boosting	0.301	0.451	0.581
Covariance augmentation, classification gradient boosting	0.316	0.334	0.376

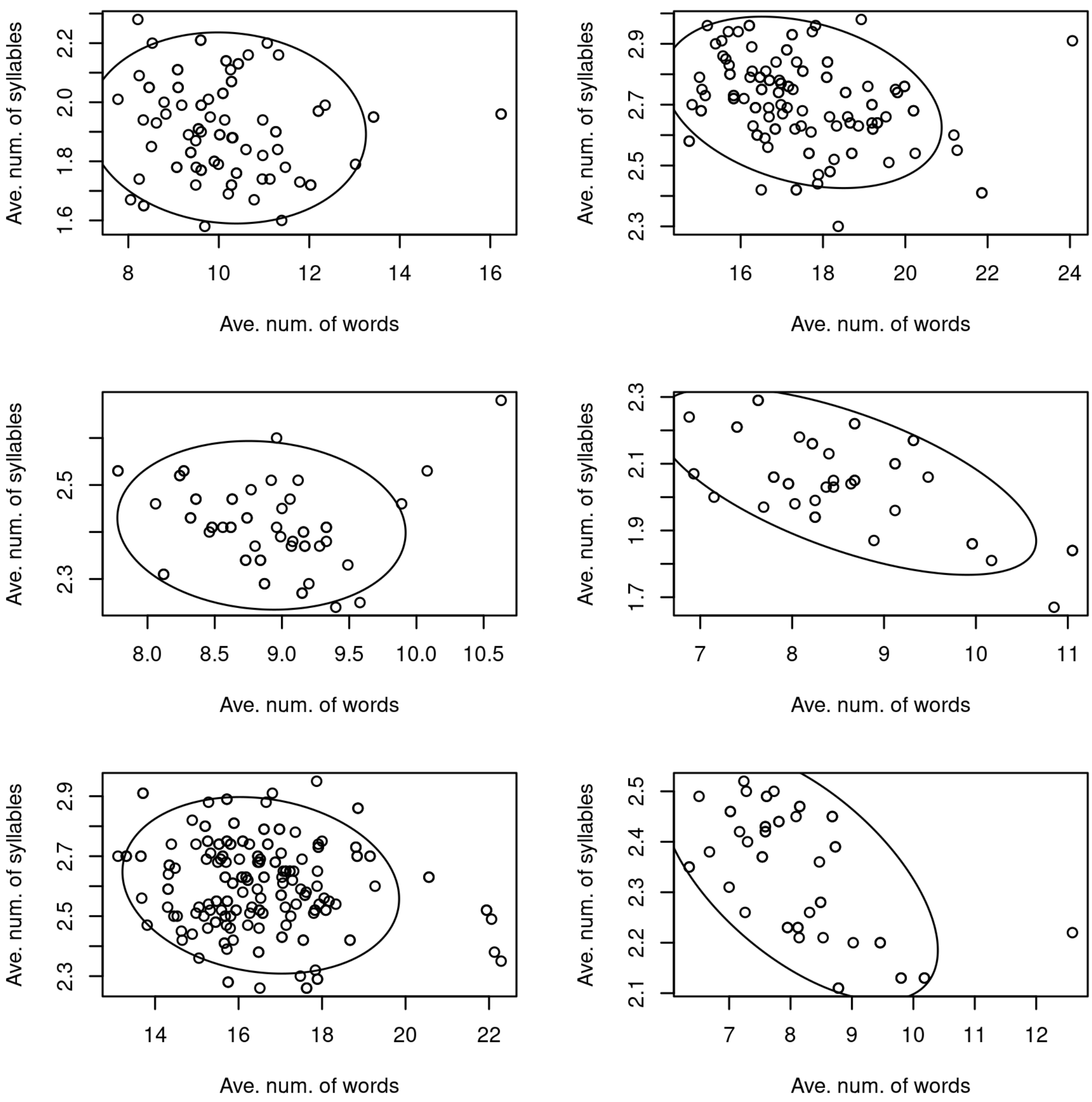


Figure 4: Chunk-level scatter plots with dispersion ellipses for several randomly selected textbooks.

7 Using the full feature set

As noted earlier, the dataset comprises 47 numerical features. In principle, all features could be leveraged to improve complexity class prediction. We apply the same modeling approaches as in previous sections, with some modifications to accommodate the higher dimensionality. Specifically, we consider:

1. The models from Section 5, excluding additive regression and tensor-product spline models.

2. Stepwise variable selection applied to linear regression. We evaluate three variants: forward selection, backward elimination, and bidirectional stepwise selection (Hastie et al., 2009). Model selection is based on leave-one-out cross-validation (LOOCV), which can be computed efficiently using the PRESS statistic from Allen (1974) under squared-error loss.

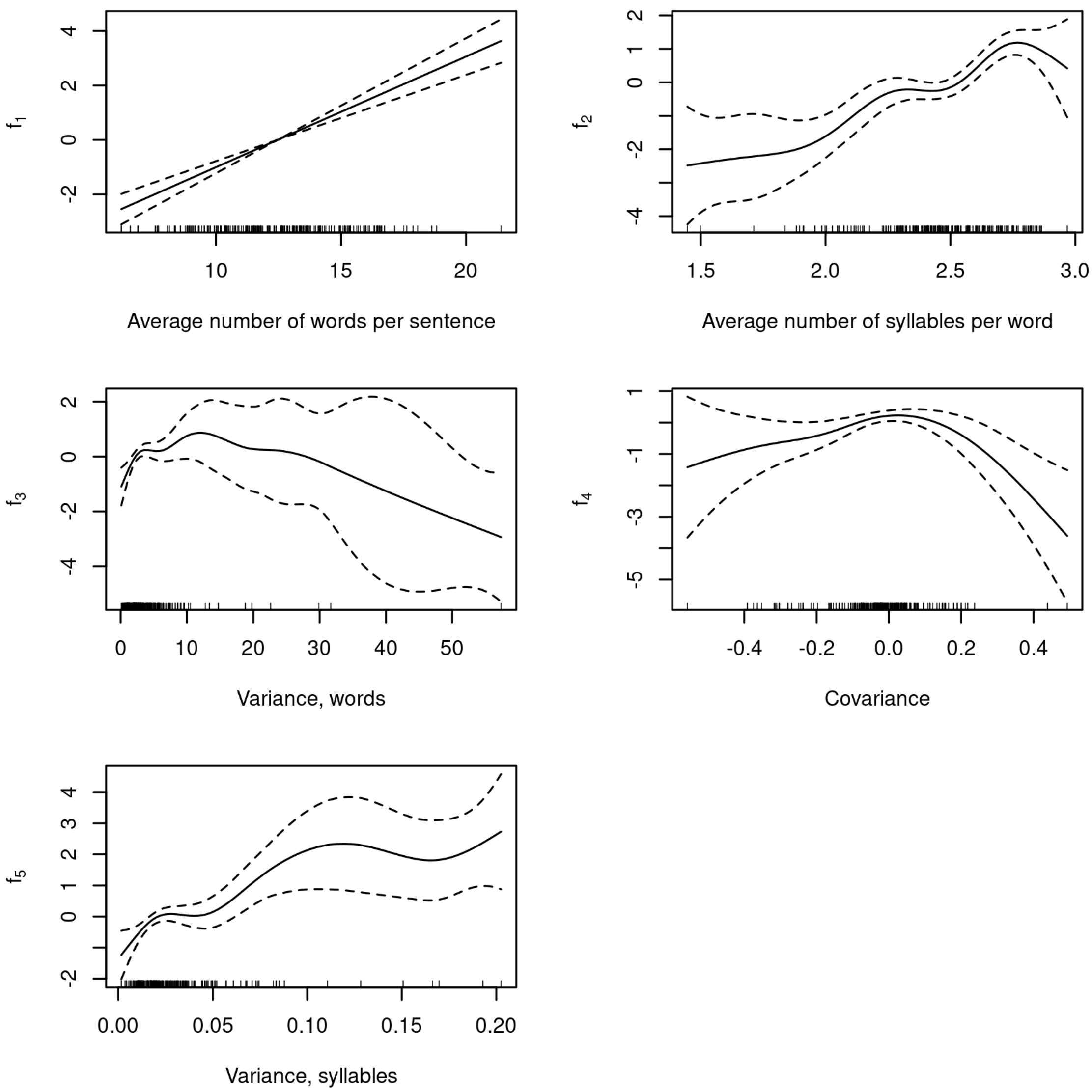


Figure 5: Component functions from the additive regression model predicting complexity class, fitted on textbook-level data augmented with variance and covariance features.

3. Selected models from Section 5 applied to the reduced feature set identified by stepwise selection.

Results are presented in Table 6. The best-performing models achieved $\mathrm{MSE} = 1.04$, $\mathrm{MAE} = 0.69$ for full set and $\mathrm{MSE} = 0.84$, $\mathrm{MAE} = 0.61$ for reduced feature set. Notably, stepwise selection degraded predictive performance, likely due to its instability when optimizing cross-validation metrics. Applying other models to the reduced feature set yielded better results; however, because feature selection was performed on the full dataset prior to cross-validation, these metrics are likely optimistically biased due to data leakage. To obtain unbiased estimates, feature selection should be nested within each cross-validation fold; we acknowledge this limitation and report the current results as upper-bound estimates. Additive models were not attempted with the full feature set, as the feature-to-sample ratio is too high for such methods to perform reliably.

Table 6: Comparison of modeling approaches using the full feature set. All metrics are computed using the same textbook-level cross-validation splits. Best results are highlighted separately for the full feature set and the reduced feature set.

Method	R_0	R_1	R_2
Linear regression	0.471	**0.641**	**0.799**
Ordinal regression	0.461	0.633	0.795
Gradient boosting	0.442	0.63	0.796
Classification gradient boosting	0.476	0.635	0.781
Neural network	0.476	0.635	0.781
Classification neural network	0.466	0.613	0.742
Backward stepwise regression	**0.515**	0.557	0.543
Forward stepwise regression	0.034	-1.686	-5.125
Two-way stepwise regression	0.495	0.597	0.676
Linear regression, reduced columns	0.485	**0.678**	**0.838**
Ordinal regression, reduced columns	0.49	0.661	0.816
Gradient boosting, reduced columns	0.447	0.557	0.66
Gradient boosting, classification, reduced columns	**0.495**	0.542	0.543
Neural network, reduced columns	0.034	-1.686	-5.125
Classification neural network, reduced columns	0.495	0.625	0.717

Leveraging the hierarchical data structure with the full feature set may yield further improvements; we leave this exploration to future work.

We also examine the relationship between complexity class and the selected features. Coefficients from the linear regression model fitted on the reduced feature set are shown in Table 7. This reduced set comprises 25 features. As is standard, the t-statistic for each coefficient indicates both the direction and relative strength of the predictor's association with the outcome. Notably, the two classical features (ASL and AWL) were not retained in the reduced model. This is likely due to their high correlation with other selected features; thus, information about sentence and word length is captured indirectly through correlated predictors.

8 Conclusions

The classical readability model – which expresses text complexity as a linear combination of average sentence length (ASL, in words) and average word length (ASW, in syllables) – was validated on our corpus. Researchers may substitute an ordinal regression framework if the ordered nature of complexity classes warrants it, though predictive performance remains comparable. While the classical model exhibits substantial error when complexity is defined by textbook target audience (i.e., grade level), more flexible models using the same two features do not outperform it, supporting the adequacy of the linear specification.

Parameter estimates for the classical model on our data align with prior work. The ASL coefficient is close to values reported previously (e.g., Matskovskiy, 1976; Oborneva, 2006; Solovyev et al., 2018), while the ASW coefficient is smaller in magnitude but remains positive. Predictive performance, compared to

Table 7: Linear regression coefficients estimated via least squares on the reduced feature set.

Variable	Estimate	Std. Error	t value	P$(>\|t\|)$
(Intercept)	-15.75562	6.98833	-2.25	0.02537
Average number of characters per word	2.67601	0.66429	4.03	8e-05
Nouns	0.21809	0.03297	6.62	0
Verbs	0.12258	0.02952	4.15	5e-05
Average number of adjectives per sentence	-2.51834	0.49561	-5.08	0
Numericals	-0.02619	0.00813	-3.22	0.00152
Average rank by Sharov dictionary	0.00029	0.00011	2.66	0.00849
Frequency by Sharov dictionary	0.00228	0.00063	3.61	0.00039
Abstractness score	3.37084	1.24693	2.7	0.00752
Local noun overlap	4.61296	2.64968	1.74	0.0834
Global noun overlap	14.71276	5.50109	2.67	0.00817
Local argument overlap	-5.70477	1.57202	-3.63	0.00037
Type/Token Ratio absolute	-474.42970	481.91975	-0.98	0.32621
Type/Token Ratio average	478.34127	482.14596	0.99	0.32248
Nominative case Noun	-0.02837	0.00918	-3.09	0.00233
Dative case Noun	-0.05885	0.03385	-1.74	0.08385
Present tense Verb	0.06762	0.01380	4.9	0
Past tense Verb	0.06839	0.01374	4.98	0
Adjective/Noun ratio	54.26626	9.66085	5.62	0
Monosyllabic words	-0.02601	0.00820	-3.17	0.00177
Two-syllable words	-0.02409	0.00660	-3.65	0.00034
Three-syllable words	-0.02989	0.00763	-3.92	0.00013
Four-syllable words	-0.03105	0.00941	-3.3	0.00117
Average number of adverbs per sentence	13.06065	1.46875	8.89	0
Content words	-0.12834	0.02310	-5.56	0
Lexical density	-28.94135	7.55847	-3.83	0.00018

earlier results on a textbook corpus (Solovyev et al., 2018), is somewhat lower; this difference is likely attributable to the greater subject-area diversity in our dataset.

Exploiting the hierarchical structure of the data – by segmenting textbooks into chunks and aggregating chunk-level information – improves textbook-level predictions. The most effective approach augments classical features with covariance statistics computed across chunks within each textbook. This method yielded approximately 10% improvement across the R_0, R_1, and R_2 metrics. A generative model based on a multivariate normality assumption achieved performance comparable to linear regression, suggesting that the added complexity did not translate into predictive gains.

Incorporating all 47 available features further improved predictive performance, though prediction errors were not eliminated. Linear regression emerged as the best-performing model among those evaluated. Relative to the classical two-feature linear model, the full-feature model achieved a 50% increase in prediction score: $R_1 = 0.641$ versus $R_1 = 0.418$. The strong performance of linear regression, relative to more flexible methods, likely reflects the limited sample size (206 textbooks) relative to feature dimensionality (47 predictors), which predisposes complex models to overfitting.

References

Abramov, A., Ivanov, V., Solovyev, V. (2023). Estimating Lexical Complexity in Multi-Domain Settings for the Russian Language. *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pp. 516–526.

Abramov, A., Ivanov, V. V. (2022). Collection and evaluation of lexical complexity data for Russian language using crowdsourcing. *Russian Journal of Linguistics*, 26(2), pp. 409–425. https://doi.org/10.22363/2687-0088-30118

Aleksandrovich, S. S. (2022). What neural networks know about linguistic complexity. *Russian Journal of Linguistics*, 26(2), pp. 371–390. https://doi.org/10.22363/2687-0088-30178

Allen, D. M. (1974). The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction. *Technometrics*, 16(1), pp. 125–127. https://doi.org/10.1080/00401706.1974.10489157

Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Corlatescu, D., Ruseti, S., Dascalu, M. (2022). ReaderBench: Multilevel analysis of Russian text characteristics. *Russian Journal of Linguistics*, 26(2), pp. 342–370. https://doi.org/10.22363/2687-0088-30145

Feng, J., Yu, S. (2024). A Method for Measuring Word Sequence Complexity of Text. *Journal of Quantitative Linguistics*, pp. 1–21. https://doi.org/10.1080/09296174.2024.2417448

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3), pp. 221–233. https://doi.org/10.1037/h0057532

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (1995). *Bayesian data analysis*. Chapman; Hall/CRC.

Hastie, T., Tibshirani, R., Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Technical report, Naval Technical Training Command Millington TN Research Branch*, pp. 1–51.

Matskovskiy, M. (1976). Problemy chitabelnosti pechatnogo teksta [The Problems of Typed Text Readability]. *Smyslovoe vospriyatie rechevogo soobshcheniya (v usloviyakh massovoy kommunikatsii)*, pp. 126–142.

McLachlan, G. J., Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons. https://doi.org/10.1002/9780470191613

Mikk, Y. (1974). Methodology for developing readability formulas. *Sovetskaya pedagogika i shkola*, (9), pp. 78–163.

Morozov, D. A., Glazkova, A. V., Iomdin, B. L. (2022). Text complexity and linguistic features: Their correlation in English and Russian. *Russian Journal of Linguistics*, 26(2), pp. 426–448. https://doi.org/10.22363/2687-0088-30132

Oborneva, I. V. (2006). Avtomatizirovannaja ocenka složnosti ucebnyx tekstov na osnove statisticeskix parametrov [Automatic evaluation of the complexity of educational texts on the basis of statistical parameters] [Doctoral dissertation, Ph. D. thesis].

Royston, J. P. (1983). Some techniques for assessing multivarate normality based on the Shapiro-Wilk W. *Journal of the Royal Statistical Society (C)*, 32(2), pp. 121–133. https://doi.org/10.2307/2347291

Shpakovskiy, Y., Et al. (2007). Otsenka trudnosti vospriyatiya i optimizatsiya slozhnosti uchebnogo teksta [Evaluation of the difficulty of perception and optimization of the text complexity] [Doctoral dissertation, PhD thesis].

Solnyshkina, M., Ivanov, V., Solovyev, V. (2018). Readability formula for Russian texts: A modified version. *Advances in Computational Intelligence: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part II 17*, pp. 132–145.

Solnyshkina, M., Solovyev, V., Danilov, A., Zamaletdinov, R., Akhtyamova, S. (2024). Multilevel Analyses of Russian Texts with RuLingva: A Case Study. *Mexican International Conference on Artificial Intelligence*, pp. 234–246.

Solnyshkina, M., Solovyev, V., Gafiyatova, E., Martynova, E. (2022). Text complexity as interdisciplinary problem. *Voprosy Kognitivnoy Lingvistiki*, pp. 18–39.

Solovyev, V., Ivanov, V., Solnyshkina, M. (2018). Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of intelligent & fuzzy systems*, 34(5), pp. 3049–3058. https://doi.org/10.21236/ADA006655

Solovyev, V., Solnyshkina, M. I., McNamara, D. S., Et al. (2022). Computational linguistics and discourse complexology: Paradigms and research methods. *Russian Journal of Linguistics*, 26(2), pp. 275–316. https://doi.org/10.22363/2687-0088-31326

Winship, C., Mare, R. D. (1984). Regression Models with Ordinal Variables. *American Sociological Review*, 49(4), pp. 512–525. https://doi.org/10.2307/2095465

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1), pp. 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A., Zampieri, M. (2018). A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*, pp. 1–13. https://doi.org/10.48550/arXiv.1804.09132

Appendix 1. List of RuLingva parameters

- Tokens (words only)
- Types (words only)
- Number of syllables
- Number of sentences
- Content words per sentence
- Average number of tokens in a sentence
- Average number of syllables in a word
- Average number of characters in a word
- Nouns
- Average number of nouns in a sentence
- Verbs
- Average number of verbs in a sentence
- Adjectives
- Average number of adjectives in a sentence
- Adverbs
- Pronouns
- Numerals
- Average frequency rank (by Sharoff)
- Frequency (by Sharoff)
- FKGL (SISmod)
- FLGL (Onorneva)
- Abstractness
- Local noun overlap

- Global noun overlap
- Local argument overlap
- Global argument overlap
- Type/Token ratio (absolute)
- Type/Token ratio (average)
- Nominative case (noun)
- Genitive case (noun)
- Accusative case (noun)
- Instrumental case (noun)
- Prepositional case (noun)
- Present tense (verb)
- Future tense (verb)
- Past tense (verb)
- Verb/Noun ratio
- Adjective/Noun ratio
- Social science terms
- One syllable words
- Two-syllable words
- Three-syllable words
- Four-syllable words
- Average number of adverbs in a sentence
- Hapax legomena
- Content words
- Lexical density

Appendix 2. Generative model parameters estimation

Let us define the generative model from Section 6 more formally. Suppose N books with complexity classes $y_1, \ldots, y_N$ are observed. Each class y_n takes values from the set $1, \ldots, K$. Each book consists of m_n chunks, $n = 1, \ldots, N$, and each chunk is represented as a vector of M real-valued characteristics. All chunks of a single book indexed by n are denoted by $\mathbf{X}_n = (X_{n,1}, \ldots, X_{n,m_n})$ and $X_{n,i} \in \mathbb{R}^M, i = 1, \ldots, m_n$.

The model assumes the existence of M-dimensional unobserved random vectors $\theta_1, \ldots, \theta_N$ and $M \times M$ positive definite random matrices $\Lambda_n, n = 1, \ldots, N$. The random variables $y_n, \mathbf{X}_n, \theta_n, \Lambda_n$ are assumed independent across n. Another assumption is that chunks $X_{n,i}$ of a single book n are conditionally independent given y_n, θ_n, Λ_n. Now, for any $n = 1, \ldots, N$ let

$$\mathbf{P}(y_n = k) = \pi_k, k = 1, \ldots, K;$$

$$\Lambda_n \mid y_n \sim \mathcal{W}^{-1}\big(\nu(y_n), \Sigma(y_n)\big);$$

$$\theta_n \mid \Lambda_n, y_n \sim \mathcal{N}\left(\mu(y_n), \frac{1}{\kappa(y_n)}\Lambda_n\right);$$

$$X_{n,i} \mid \theta_n, \Lambda_n, y_n \sim \mathcal{N}\big(\theta_n, \Lambda_n\big), \quad i = 1, \ldots, m_n.$$

In the last three expressions, a conditional distribution is implied. Parameters $\mu(y), y = 1, \ldots, K$ are vectors in $\mathbb{R}^M$, and $\Sigma(y)$ is an $M \times M$ positive definite matrix. The notation $\mathcal{W}^{-1}$ denotes the inverse Wishart distribution. The above probabilistic formulation of the model is a variation of the standard conjugate model of normally distributed observations, see Gelman et al. (1995, paragraph 3.6).

Let $\varphi(X \mid \mu, \Lambda)$ denote the density of the normal distribution with mean vector μ and covariance matrix Λ evaluated at X, and $p_{\mathcal{W}^{-1}}(\Lambda \mid \nu, \Sigma)$ denote the density of the inverse Wishart distribution with parameters ν, Σ evaluated at Λ. Their respective formulas for M-dimensional random variables are given by

$$\varphi(X \mid \mu, \Lambda) = (2\pi)^{-\frac{M}{2}} |\Lambda|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (X - \mu)^T \Lambda^{-1} (X - \mu)\right),$$

$$p_{\mathcal{W}^{-1}}(\Lambda \mid \nu, \Sigma) = |\Sigma|^{\frac{\nu}{2}} 2^{-\frac{\nu M}{2}} \left(\Gamma_M\left(\frac{\nu}{2}\right)\right)^{-1} |\Lambda|^{-\frac{\nu+M+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Sigma\Lambda^{-1}\right)\right),$$

where Γ_M is the multivariate gamma function

$$\Gamma_M(x) = \pi^{\frac{M(M-1)}{4}} \prod_{i=1}^{M} \Gamma\left(x + \frac{1-i}{2}\right).$$

We also define

$$N_y = \sum_{n=1}^{N} I(y_n = y),$$

$$\overline{\mathbf{X}}_n = \frac{1}{m_n} \sum_{i=1}^{m_n} X_{n,i}.$$

A standard result (see Gelman et al., 1995, paragraph 3.6) states that the posterior distribution is also Normal-Inverse-Wishart:

$$(4) \qquad \Lambda_n \mid y_n, \mathbf{X}_n \sim \mathcal{W}^{-1}\left(\nu(y_n) + m_n, \Sigma(y_n) + S_n + \frac{m_n \kappa(y_n)}{m_n + \kappa(y_n)} \left(\overline{\mathbf{X}}_n - \mu(y_n)\right) \left(\overline{\mathbf{X}}_n - \mu(y_n)\right)^T\right),$$

$$(5) \qquad \theta_n \mid \Lambda_n, y_n, \mathbf{X}_n \sim \mathcal{N}\left(\frac{\kappa(y_n)\mu(y_n) + m_n \overline{\mathbf{X}}_n}{m_n + \kappa(y_n)}, \frac{1}{\kappa(y_n) + m_n} \Lambda_n\right),$$

where

$$S_n = \sum_{i=1}^{m_n} \left(X_{n,i} - \overline{\mathbf{X}}_n\right) \left(X_{n,i} - \overline{\mathbf{X}}_n\right)^T.$$

If the parameters $\pi, \mu, \kappa, \nu, \Sigma$ are known, then it is possible to calculate the posterior probabilities of any complexity class for a new observation $\mathbf{X}$ consisting of m parts:

$$\mathbf{P}(y = k|\mathbf{X}) \propto p(\mathbf{X} \mid y = k)\mathbf{P}(y = k) = \pi_k \frac{p(\mathbf{X}, \theta, \Lambda \mid y = k)}{p(\theta, \Lambda \mid \mathbf{X}, y = k)}.$$

Here $p(\mathbf{X} \mid y = k)$ is the density of $\mathbf{X}$ for known y. The right-hand side of the expression above suggests a straightforward way to compute this quantity using (4) and (5). We state the standard result of this calculation:

$$p(\mathbf{X} \mid y = k) = \pi^{-\frac{mM}{2}} |\Sigma(k)|^{\frac{\nu(k)}{2}} \left|\Sigma(k) + S + \frac{m\kappa(k)}{m + \kappa(k)} \left(\overline{\mathbf{X}} - \mu(k)\right) \left(\overline{\mathbf{X}} - \mu(k)\right)^T\right|^{-\frac{\nu(k)+m}{2}} \times$$
$$\Gamma_M\left(\frac{\nu(k) + m}{2}\right) \left[\Gamma_M\left(\frac{\nu(k)}{2}\right)\right]^{-1} \sqrt{\frac{\kappa(k)}{\kappa(k) + m}}$$

with analogous definitions for $\overline{\mathbf{X}}$ and S.

Next, we present an estimation algorithm for the parameters $\pi, \mu, \kappa, \nu, \Sigma$. Since there are unobserved variables θ_n and Λ_n, we employ the EM algorithm McLachlan and Krishnan (2007). To initialize it, we need starting parameter values. A natural initial approximation is given by

$$\mu(y) = \frac{1}{N_y} \sum_{n: y_n = y} \overline{\mathbf{X}}_n,$$

$$\Sigma(y) = \frac{1}{N_y} \sum_{n:y_n=y} \frac{1}{m_n} \sum_{i=1}^{m_n} \left(X_{n,i} - \overline{\mathbf{X}}_n\right) \left(X_{n,i} - \overline{\mathbf{X}}_n\right)^T,$$

$$\pi_k = \frac{N_k}{N},$$

and the parameters $\kappa(y)$ and $\nu(y)$ are initialized to 1.

Then the E-step and M-step are repeated until convergence.

E-step. The conditional distributions of θ_n, Λ_n given $y_n, \mathbf{X}_n$ are calculated for known parameter values. These are given by (4), (5). We define

$$\xi_n = \frac{\kappa(y_n)\mu(y_n) + m_n\overline{\mathbf{X}}_n}{m_n + \kappa(y_n)},$$

$$\kappa_n = m_n + \kappa(y_n),$$

$$\nu_n = \nu(y_n) + m_n,$$

$$\Sigma_n = \Sigma(y_n) + S_n + \frac{m_n\kappa(y_n)}{m_n + \kappa(y_n)} \left(\overline{\mathbf{X}}_n - \mu(y_n)\right) \left(\overline{\mathbf{X}}_n - \mu(y_n)\right)^T.$$

We also define

$$\lambda_n = \mathbf{E}\left(\ln|\Lambda_n| \mid y_n, \mathbf{X}_n\right) = -\sum_{i=1}^{M} \psi\left(\frac{\nu_n}{2} + \frac{1-i}{2}\right) - M\ln 2 + \ln|\Sigma_n|,$$

where ψ is the digamma function; this result is a known property of the Wishart distribution. We also use another property of the Wishart distribution:

$$L_n = \mathbf{E}\left(\Lambda_n^{-1} \mid y_n, \mathbf{X}_n\right) = \nu_n\Sigma_n^{-1}.$$

These values are calculated in the E-step and treated as known in the M-step. Let the corresponding joint density of θ_n, Λ_n (evaluated at point θ, Λ) be denoted by $q_n(\theta, \Lambda)$.

M-step. We need to maximize the function

$$(6) \quad \sum_{n=1}^{N}\left[\int\int q_n(\theta,\Lambda)\left\{\sum_{i=1}^{m_n} \ln\varphi\left(X_{n,i} \mid \theta,\Lambda\right) + \ln\varphi\left(\theta \mid \mu(y_n), \frac{1}{\kappa(y_n)}\Lambda\right) + \ln p_{\mathcal{W}^{-1}}\left(\Lambda \mid \nu(y_n), \Sigma(y_n)\right)\right\} d\theta d\Lambda + \ln\pi_{y_n}\right]$$

with respect to parameters $\mu, \kappa, \nu, \Sigma, \pi$ subject to the constraint $\sum \pi_k = 1$.

The estimates of π_k are straightforward and given by

$$\pi_k = \frac{N_k}{N},$$

which matches the initial values.

In expression (6), only the terms depending on μ, κ, ν, Σ are the prior density terms. Thus, we need to calculate expectations of these terms. Expectation of twice the log prior density of θ equals

$$\int q(\theta \mid \Lambda) 2 \ln \varphi \left(\theta \mid \mu(y_n), \frac{1}{\kappa(y_n)} \Lambda \right) d\theta = \int q(\theta \mid \Lambda) \Big[-M \ln(2\pi) + \ln \kappa(y_n) - \ln|\Lambda| - \kappa(y_n)(\theta - \mu(y_n))^T \Lambda^{-1} (\theta - \mu(y_n)) \Big] d\theta. \tag{7}$$

Now we use the fact that $q_n(\theta|\Lambda)$ is a normal density with mean ξ_n and covariance matrix $(\kappa(y_n)+m_n)^{-1}\Lambda$. Thus the expression (7) equals to

$$-M \ln(2\pi) + \ln \kappa(y_n) - \ln|\Lambda| - \kappa(y_n) \left[(\xi_n - \mu(y_n))^T \Lambda^{-1} (\xi_n - \mu(y_n)) + \frac{1}{\kappa_n} \mathrm{tr}(\Lambda^{-1}\Lambda) \right] \tag{8}$$

In what follows, we use the fact that $\mathrm{tr}(\Lambda^{-1}\Lambda) = M$.

Continuing the calculation of (6) by integrating with respect to Λ, the density $q_n(\Lambda)$ is given by the expression

$$q_n(\Lambda) = p_{\mathcal{W}^{-1}}(\Lambda \mid \nu_n, \Sigma_n).$$

Combining the previous integration result (8) and $\ln p_{\mathcal{W}^{-1}}\big(\Lambda \mid \nu(y_n), \Sigma(y_n)\big)$:

$$-M \ln(2\pi) + \ln \kappa(y_n) - \ln|\Lambda| - \kappa(y_n) \left[(\xi_n - \mu(y_n))^T \Lambda^{-1} (\xi_n - \mu(y_n)) + \frac{1}{\kappa_n} \mathrm{tr}(\Lambda^{-1}\Lambda) \right] + \nu(y_n) \ln|\Sigma(y_n)| - \nu(y_n) M \ln 2 - 2 \ln \Gamma_M \left(\frac{\nu(y_n)}{2} \right) - (\nu(y_n) + M + 1) \ln|\Lambda| - \mathrm{tr}\left(\Sigma(y_n) \Lambda^{-1} \right). \tag{9}$$

Next, we calculate the expectation of (9) with respect to $q_n(\Lambda)$. Here we need $\mathbf{E} \ln|\Lambda|$ and $\mathbf{E}\Lambda^{-1}$, which are given by λ_n and L_n, respectively. Taking this step and summing over n, we obtain

$$\sum_{n=1}^{N} \Bigg[-M \ln(2\pi) + \ln \kappa(y_n) - \lambda_n - \kappa(y_n)(\xi_n - \mu(y_n))^T L_n (\xi_n - \mu(y_n)) - \frac{M\kappa(y_n)}{\kappa_n} + \nu(y_n) \ln|\Sigma(y_n)| - \nu(y_n) M \ln 2 - 2 \ln \Gamma_M \left(\frac{\nu(y_n)}{2} \right) - (\nu(y_n) + M + 1)\lambda_n - \mathrm{tr}\left(\Sigma(y_n) L_n \right) \Bigg] \tag{10}$$

up to a constant.

We now optimize this function.

The derivative of (10) with respect to $\mu(k)$ equals

$$2\kappa(k) \sum_{n:y_n=k} (L_n\xi_n - L_n\mu(k)).$$

Equating this to zero and solving gives

$$\mu(k) = \left(\sum_{n:y_n=k} L_n\right)^{-1} \left(\sum_{n:y_n=k} L_n\xi_n\right).$$

The derivative of (10) with respect to $\kappa(k)$ equals

$$\frac{N_k}{\kappa(k)} - \sum_{n:y_n=k} \left[(\xi_n - \mu(y_n))^T L_n(\xi_n - \mu(y_n)) + \frac{M}{\kappa_n}\right].$$

Therefore,

$$\kappa(k) = N_k \left(\sum_{n:y_n=k} \left[(\xi_n - \mu(y_n))^T L_n(\xi_n - \mu(y_n)) + \frac{M}{\kappa_n}\right]\right)^{-1}.$$

The derivative of (10) with respect to $\Sigma(k)$ equals

$$N_k\nu(k)\Sigma(k)^{-1} - \sum_{n:y_n=k} L_n$$

Equating this to zero and solving gives

$$\Sigma(k) = \left(\frac{1}{N_k\nu(k)} \sum_{n:y_n=k} L_n\right)^{-1}. \tag{11}$$

The derivative of (10) with respect to $\nu(k)$ equals

$$N_k \ln|\Sigma(k)| - N_k M \ln 2 - N_k \sum_{i=1}^{M} \psi\left(\frac{\nu(k)}{2} - \frac{1-i}{2}\right) - \sum_{n:y_n=k} \lambda_n.$$

Substituting (11) into this expression:

$$\ln|\Sigma(k)| = \ln\left|\frac{1}{N_k\nu(k)} \sum_{n:y_n=k} L_n\right|^{-1} = -\ln\nu(k)^{-M}\left|\frac{1}{N_k} \sum_{n:y_n=k} L_n\right| = M\ln\nu(k) - \ln\left|\frac{1}{N_k} \sum_{n:y_n=k} L_n\right|.$$

Thus, to find the M-step estimate of $\nu(k)$, we need to solve the equation

$$(12)\qquad N_k M \ln \nu(k) - N_k \ln \left| \frac{1}{N_k} \sum_{n:y_n=k} L_n \right| - N_k M \ln 2 - N_k \sum_{i=1}^{M} \psi\left(\frac{\nu(k)}{2} - \frac{1-i}{2}\right) - \sum_{n:y_n=k} \lambda_n = 0.$$

This equation is solved numerically, and the resulting root is then used in (11).

In the case of a flat prior distribution for y, the probabilities π_k are fixed: $\pi_k = K^{-1}, k = 1, \ldots, K$. This assumption may be natural since the researcher does not expect any of the complexity classes to be more frequent than others.

Homoscedasticity means that the distribution of Λ_n is independent of y_n. In this case, the algorithm requires only minor modifications. The expressions for ν (12) and Σ (11) are modified so that they depend on all observations, not only those from their corresponding complexity classes. This modification is used when the number of observations is too small to properly estimate the Σ parameters. The results in Section 6 (see Table 5) show that in our case we had sufficient observations to use the heteroscedastic model.

Syntactic Complexity Across Genres in Karel Čapek's Writing

Michaela Nogolová[1*], Xinying Chen[1], Miroslav Kubát[1], Žaneta Stiborská[1]

[1] University of Ostrava
[*] Corresponding author's email: nogolovam@gmail.com

DOI: https://doi.org/10.53482/2026_60_433

ABSTRACT

This paper investigates syntactic complexity across eight genres in the works of Czech writer Karel Čapek using a stylometric approach. The aim of the study is to determine whether syntactic complexity differs among the following genres: novels, short stories, travelogues, poems, newspaper columns, academic essays, children's literature, and personal correspondence. For the analysis, we compute five core complexity metrics: Average Sentence Length (in words and in clauses), Average Clause Length, Mean Dependency Distance (MDD), and Mean Hierarchical Distance (MHD). The results indicate systematic genre-specific differences in syntactic patterning within Čapek's literary work. Academic literature and travelogues consistently occupy the upper end of the complexity scale, whereas novels and short stories show comparatively low complexity, in line with their narrative and accessibility-oriented functions. Poetry is characterised by relatively low average complexity but high variability, and children's literature emerges as unexpectedly complex across several metrics. Personal correspondence and newspaper columns take up intermediate positions. Overall, the findings demonstrate that syntactic complexity offers a useful structural signal for distinguishing genres within a single author and underscore the value of syntactic features in stylometric and genre-oriented research.

Keywords: syntactic complexity, dependency syntax, genre, Karel Čapek

1 Introduction

Quantitative analysis of literary style, or stylometry, has repeatedly shown that linguistic features can reliably distinguish authors, periods and genres (e.g. Eisen et al., 2016; Grieve, 2023). Early work in computational authorship attribution and style analysis predominantly focused on lexical indicators such as word frequencies and function-word profiles (cf. Kestemont, 2014). However, over time stylometric research has increasingly moved beyond purely lexical indicators toward richer feature spaces that also incorporate syntactic, morphosyntactic, and discourse-level variables. (see Herrmann et al., 2021 for an overview). This development reflects the view that syntactic patterns, by capturing aspects of style that are less sensitive to topical variation than vocabulary alone, can provide a robust and theoretically motivated complement to lexical analysis.

The present study focuses on the works of Karel Čapek, one of the most prominent Czech authors of the twentieth century. Čapek's writing is unusually diverse in terms of genre and communicative setting: it includes novels, short stories, travelogues, poems, newspaper articles, academic literature, children's literature and personal correspondence. This makes him an ideal test case for examining the extent to which genre predicts syntactic complexity within a single author's production. Rather than asking whether syntactic complexity can distinguish between different authors, we ask how far syntactic complexity varies within an author as he moves across genres. This intra-author perspective complements more common cross-author stylometric designs and allows us to control for many confounding factors related to individual style. At the same time, we are aware that our empirical findings are restricted to Čapek's texts and do not directly extend to analysed genres in general. The study also builds on previous quantitative work on Čapek's writing, which has so far concentrated mainly on lexical and morphological characteristics (e.g., Kubát, 2016; Čech and Kubát, 2018)

Our analysis focuses on syntactic features. Using automatically parsed dependency trees, we compute five complexity metrics for each text: Average Sentence Length (in words), Average Sentence Length (in clauses), Average Clause Length (in words), Mean Dependency Distance (MDD) and Mean Hierarchical Distance (MHD). Together, these metrics capture both traditional length-based properties and dependency-based structural characteristics.

By integrating detailed syntactic analysis with a genre-focused perspective, the study aims to refine our understanding of Čapek's stylistic versatility and to contribute more generally to research on the relationship between syntax and genre. The findings not only reveal genre-specific patterns in Čapek's writing but also offer a complementary perspective to genre studies based on large, multi-author corpora (cf. Wang and Liu, 2017; Chen and Kubát, 2024).

2 Language Material and Methodology

The corpus used in this study consists of over 700 texts written by Karel Čapek, spanning eight distinct genres: novels, short stories, travelogues, newspaper columns, academic literature, poetry, children's literature, and correspondence. Due to substantial differences in text lengths across genres, the study adopts a standardized approach to define a "text" as a unit. A text is defined as one poem, one short story from a collection, one chapter from a novel, one letter of correspondence, one chapter from a scientific book, one chapter from a travelogue, or one story from a children's book. This approach ensures comparability across genres, provides a more robust dataset for analysing intra-genre variations and enabling statistical evaluation. An overview of the corpus is presented in Table 1.

Table 1: Overview of the language material.

Genres	Number of texts	Total size in tokens	Average size in tokens
novel	252	219078	869.357
short story	71	96318	1356.592
travelogue	132	49508	377.924
newspaper column	92	39042	424.370
academic literature	70	44518	635.971
poem	24	1984	82.667
children literature	26	25944	997.846
correspondence	93	26275	282.527
total	**760**	**502667**	

The syntactic analysis in this study was conducted in following steps. First, all texts were parsed using UDPipe 2.0 (Straka, 2018), employing pre-trained models from the Universal Dependencies (UD) version 2.15 dataset (Zeman et al., 2024). Following the initial parsing, the output was systematically converted into the Surface Syntactic Universal Dependencies (SUD) framework (Gerdes et al., 2018), which emphasizes syntactic functions over content-focus features and is particularly well-suited for assessing structural sentence complexity.

For data consistency, only sentences that (i) have a finite verb or auxiliary as their syntactic root, and (ii) are free of tokens containing abbreviations, numerical digits, or non-standard characters, were retained. This filtering step reduces potential noise in the dependency trees and supports more reliable computation of syntactic indices.

In this analysis, we use Average Sentence Length (ASL), Average Clause Length (ACL), Mean Dependency Distance (MDD), and Mean Hierarchical Distance (MHD). Together, these indicators capture both surface-level and deep structural characteristics of syntax.

ASL was calculated using two complementary methods: one based on the word count per sentence, and another based on the number of clauses per sentence. While the first method captures overall sentence elaboration, the second reflects the density of subordinate structures within sentences.

ACL was derived by dividing the total word count by the total number of clauses, offering insight into the internal complexity of clauses themselves.

MDD, as formulated by Liu (2008), assesses the average linear distance between a dependent and its syntactic head. For each word in the text (excluding root tokens and punctuation), the dependency distance was calculated as the absolute difference between the word's index and that of its syntactic parent. The average was then computed as follows:

$$MDD = \frac{\sum_{i=1}^{n-s} |DD_i|}{n-s}$$

where n denotes the total number of words, s is the number of analysed sentences and *DD* represents the dependency distance of the i^{th} word. Because the root of the sentence is excluded from the calculation, it must be omitted from the total word count; hence, $n - s$.

MHD, proposed by Jing and Liu (2015), offers a hierarchical perspective by measuring the average vertical depth of dependency trees. For each token[1] in the text, the hierarchical distance (HD) is defined as the number of dependency links from the token to the root of the sentence. The average value of these distances provides a measure of how deeply nested syntactic structures are, serving as a proxy for syntactic embedding.

The differences between genres were statistically tested. For each pairwise comparison, we first assessed the normality of the distributions using the Shapiro–Wilk test (Shapiro and Wilk, 1965). If both groups met the normality assumption, we compared their means with an independent-samples t-test. If at least one group deviated from normality, we used the Mann–Whitney U test (Mann and Whitney, 1947) as a non-parametric alternative. Since multiple pairwise comparisons were carried out, we adjusted the resulting p-values for multiple testing using the Benjamini–Hochberg correction (Benjamini and Hochberg, 1995).

3 Results

This section presents the results of the syntactic complexity analysis of Karel Čapek's multi-genre corpus. Focusing first on Average Sentence Length (ASL) in words, based on the results of statistical tests (see Table 2), the genres fall into four distinguishable groups (see Figure 1). The first group, with the lowest[2] ASL, comprises poetry, novels and short stories. Poetry shows both the shortest sentences and the highest variability, which is expected given that poetic syntax must adapt to metrical and prosodic constraints, alternating between very short lines and more extended sentences. Novels and short stories also favour relatively short sentences, in line with their communicative goal of maintaining narrative flow and ensuring processing ease for a broad readership.

A second group is formed by personal correspondence and children's literature, which exhibit intermediate – and in Čapek's case noticeably higher – sentence lengths. In his letters, Čapek writes to specific, often highly educated addressees, including the first Czechoslovak president Tomáš Garrigue Masaryk and the poet, literary critic, journalist and translator S. K. Neumann. In this relatively unconstrained context, he allows himself longer, more expansive sentences, which is reflected in the higher ASL values. The relatively long sentences in children's literature are less intuitive, given the usual expectation of shorter, simpler sentences in texts for young readers. Here, the patterns suggest that Čapek does not

[1] Except for the root and punctuation which are excluded.
[2] In all figures, the boxplots are ordered by their median values, as most distributions deviate from normality and the median therefore provides a more appropriate basis for comparison than the mean.

substantially simplify his syntax for children but relies on other strategies (such as topic choice or lexical transparency) to ensure comprehensibility.

Newspaper articles constitute a separate, third group: their ASL is significantly different from all other genres, and they occupy a position just below the most complex group. This is consistent with the genre's dual function. On the one hand, newspaper columns are information-dense and argumentative, which encourages relatively long, multi-clause sentences; on the other hand, they still address a broad, non-specialist audience, which may limit syntactic complexity compared to strictly academic prose.

The fourth group, with the highest ASL values, consists of academic prose and travelogues. Both genres are typically expository and information-heavy: academic texts condense complex arguments and theoretical discussion into syntactically dense sentences, while travelogues frequently contain extended descriptive passages and background explanations.

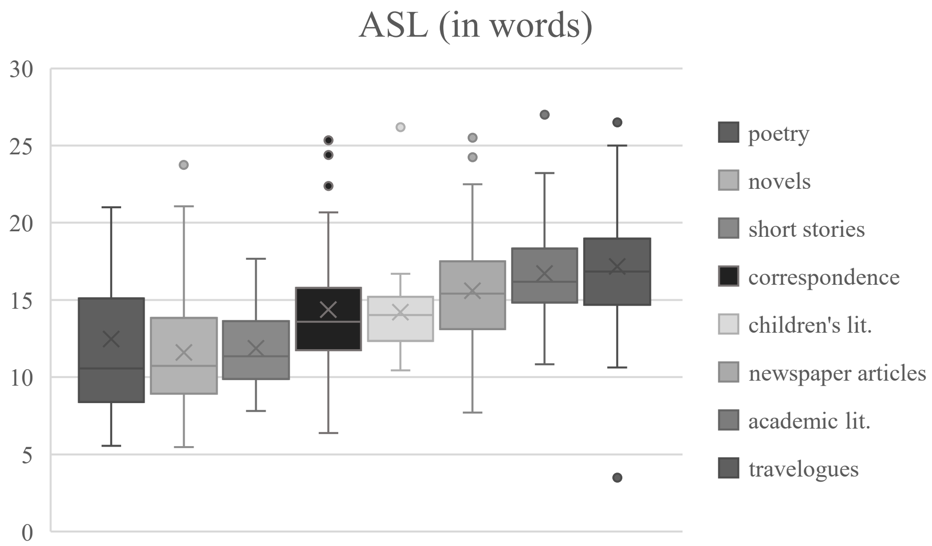


Figure 1: Results of Average Sentence Length in the number of words.

Table 2: Statistical differences (p-value) between genres based on ASL (in words).

Genres	Poetry	Novels	Short stories	Correspondence	Children's lit.	Newspaper articles	Academic lit.
novels	0.745						
short stories	0.264	0.190					
correspondence	**0.003**	**<0.001**	**<0.001**				
children's lit.	**0.007**	**<0.001**	**<0.001**	0.819			
newspaper articles	**<0.001**	**<0.001**	**<0.001**	**0.008**	**0.029**		
academic lit.	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.014**	
travelogues	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.003**	0.599

Turning to Average Sentence Length in clauses (ASL in clauses; see Figure 2 and Table 3) and Average Clause Length in words (ACL; see Figure 3 and Table 4), we find a broadly similar genre hierarchy as for ASL in words, but with some informative shifts in how complexity is realised.

Poetry has the lowest ASL in clauses and one of the lowest ACL values. It also shows considerable variability in ASL in clauses, so that it is not statistically different from most other genres (academic prose, novels, travelogues, short stories and newspapers). This pattern suggests that Čapek flexibly adjusts the number of clauses per sentence in poetry, while keeping the clauses themselves relatively short and fairly stable in length.

Academic prose has the second-lowest median of ASL in clauses (slightly above two clauses per sentence), which might initially seem at odds with its status as a complex genre. However, it exhibits the longest statistically significant ACL of all genres. Rather than piling up large numbers of clauses, Čapek tends to pack a great deal of information into individual clauses, so that much of the complexity resides within clauses rather than in long clause chains. Travelogues show a related pattern: they have somewhat more clauses per sentence than academic literature, but still relatively moderate ASL in clauses and the second-highest ACL. Again, complexity is driven by rich, information-dense clauses rather than by extreme clause stacking.

Novels and short stories have both relatively low ASL in clauses and relatively short clauses. This combination points to syntactically lighter sentences overall, consistent with genres that prioritise readability and narrative flow and avoid very complex structures, both in terms of how many clauses are combined and how long those clauses are. Personal correspondence, by contrast, features more clauses per sentence on average, while clause length remains in the mid-range. Here, Čapek appears to allow himself more expansive sentences made up of several medium-length clauses, which fits the more informal and personalised communicative setting.

Newspaper articles occupy an upper-middle position: they have the third-highest number of clauses per sentence and the third-longest clauses. They thus combine both types of complexity – more clauses and longer ones – but still remain below the most extreme values observed in the corpus. The most striking pattern is found in children's literature, which has the highest ASL in clauses and medium clause lengths. In other words, Čapek builds unusually long clause chains even when writing for children and does not fully reduce clause-based complexity.

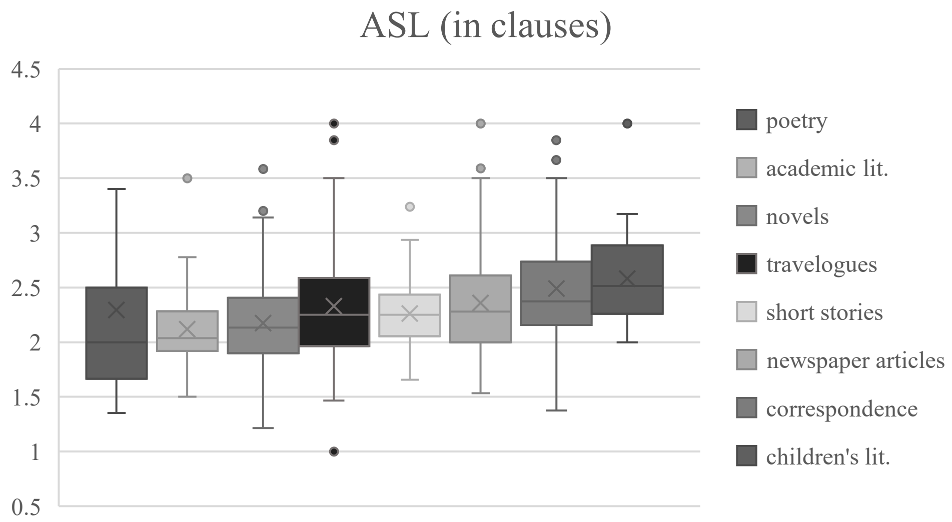


Figure 2: Results of Average Sentence Length in the number of clauses.

Table 3: Statistical differences (p-value) between genres based on ASL (in clauses).

Genres	Poetry	Academic lit.	Novels	Travelogues	Short stories	Newspaper articles	Correspondence
academic lit.	0.597						
novels	0.394	0.252					
travelogues	0.116	**<0.050**	**<0.050**				
short stories	0.130	**<0.050**	**<0.050**	0.900			
newspaper articles	0.085	**0.001**	**<0.050**	0.628	0.541		
correspondence	**<0.050**	**<0.001**	**<0.001**	**<0.050**	**<0.050**	0.054	
children's lit.	**<0.050**	**<0.001**	**<0.001**	**<0.050**	**0.001**	**<0.050**	0.230

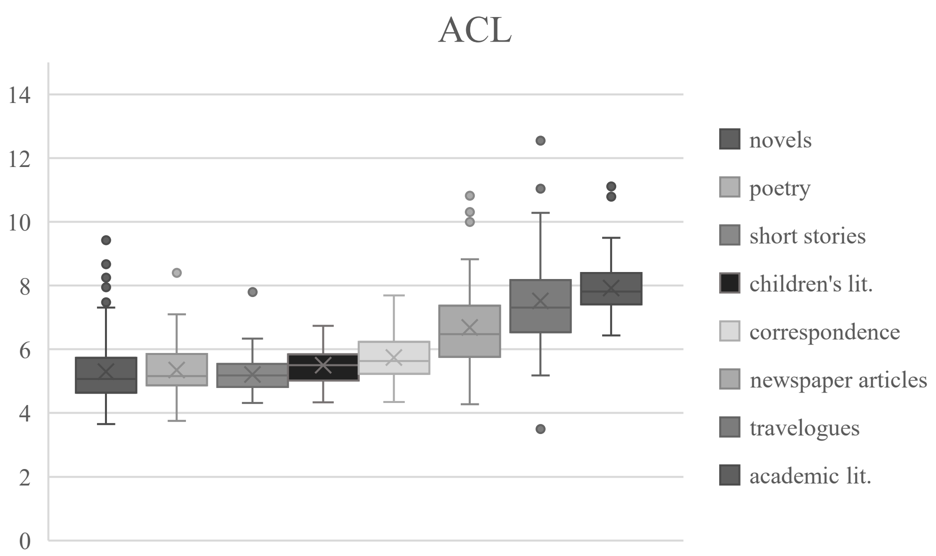


Figure 3: Results of Average Clause Length.

Table 4: Statistical differences (p-value) between genres based on ACL.

Genres	Novels	Poetry	Short stories	Children's lit.	Correspondence	Newspaper articles	Travelogues
poetry	0.584						
short stories	0.691	0.616					
children's lit.	**<0.050**	0.190	**<0.050**				
correspondence	**<0.001**	**<0.050**	**<0.001**	0.182			
newspaper articles	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**		
travelogues	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	
academic lit.	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

For Mean Dependency Distance (MDD), we again find statistically significant differences across genres (see Figure 4 and Table 5), although the overall effect is relatively modest. The clearest contrasts involve novels and short stories on the one hand and children's literature on the other. Novels and short stories have the lowest MDD values, whereas children's literature – somewhat unexpectedly – shows the highest MDD. Children's literature differs significantly from most other genres except correspondence, poetry and travelogues, while novels are significantly lower than all of them. Short stories occupy a slightly higher position than novels and do not differ significantly from poetry or academic prose. Poetry once more exhibits substantial variability but does not systematically occupy either extreme. The generally compressed range of MDD values is in line with Futrell et al. (2015), who argue that language users tend to minimise dependency distance due to working-memory constraints. In Čapek's writing, this suggests that although he clearly differentiates genres through sentence length and clause structure, genre-specific choices are constrained by a more general pressure to keep dependency distances within manageable limits.

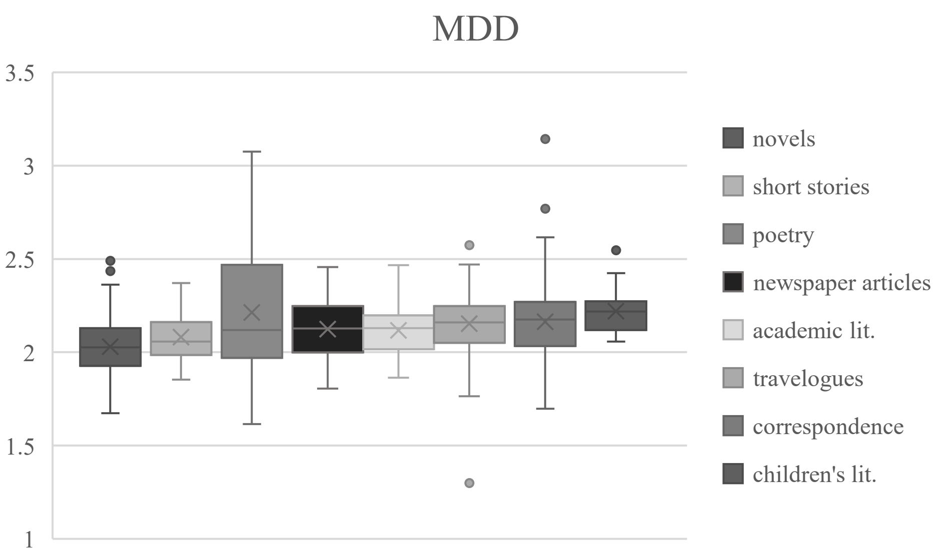


Figure 4: Results of Mean Dependency Distance.

Table 5: Statistical differences (p-value) between genres based on MDD.

Genres	novels	short stories	poetry	newspaper articles	academic lit.	travelogues	correspondence
short stories	**<0.050**						
poetry	**<0.050**	0.267					
newspaper articles	**<0.001**	**<0.050**	0.082				
academic lit.	**<0.001**	0.075	0.074	0.751			
travelogues	**<0.001**	**<0.001**	0.909	0.162	0.051		
correspondence	**<0.001**	**<0.050**	0.906	0.238	0.090	0.900	
children's lit.	**<0.001**	**<0.001**	0.937	**<0.050**	**<0.001**	0.060	0.094

Finally, Mean Hierarchical Distance (MHD) reveals a relatively clear separation between individual genres (see Figure 5). According to the statistical tests, we can again identify four distinct groups (see Table 6). The first group is formed by poetry, which has the lowest MHD. This is in line with the genre's tendency towards structurally flat, often elliptical sentences, in which many relations are left implicit rather than being expressed through deeply nested structures.

The second group comprises novels, short stories and children's literature, which show low to intermediate MHD values. Here, the syntactic structures are usually organised into moderately complex sentences: there is some hierarchical embedding, but it is kept within limits, so that the overall sentence architecture remains relatively transparent. This pattern fits genres that rely on storytelling and scene-building.

The third group, with higher MHD, consists of personal correspondence, newspaper articles and academic literature. The highest, and statistically clearly distinct MHD values are found in travelogues. This suggests that Čapek's travel writing makes particularly extensive use of embedded words in the sentence to layer descriptions, background information and commentary, creating syntactic structures that are not only long but also deeply nested in terms of their hierarchical organisation.

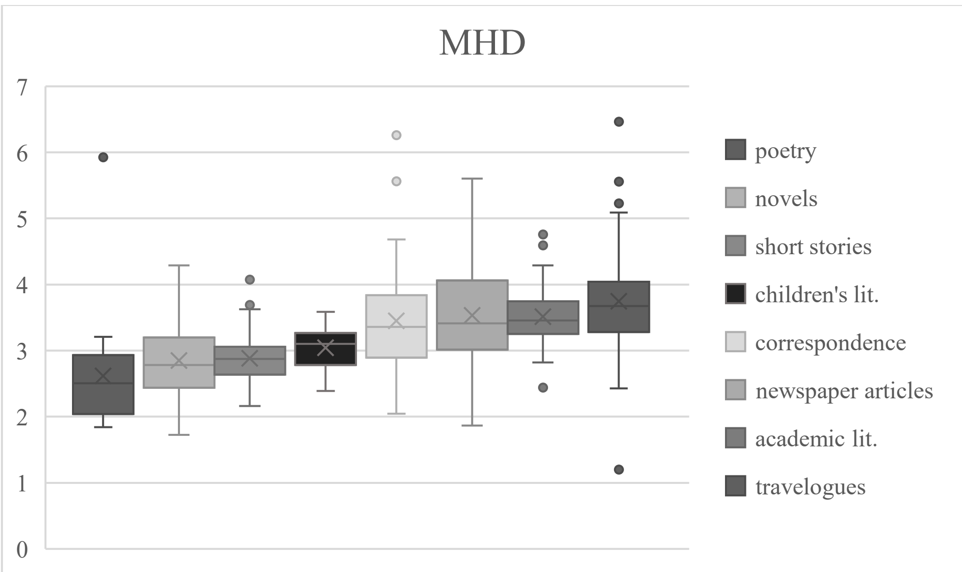


Figure 5: Results of Mean Hierarchical Distance.

Table 6: Statistical differences (p-value) between genres based on MHD.

Genres	poetry	novels	short stories	children's lit.	correspondence	newspaper articles	academic lit.
novels	**<0.050**						
short stories	**<0.001**	0.270					
children's lit.	**<0.001**	**<0.050**	0.060				
correspondence	**<0.001**	**<0.001**	**<0.001**	**<0.050**			
newspaper articles	**<0.001**	**<0.001**	**<0.001**	**0.001**	0.376		
academic lit.	**<0.001**	**<0.001**	**<0.001**	**<0.001**	0.180	0.447	
travelogues	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.001**	**<0.050**	**<0.050**

4 Conclusion

This study examined how syntactic complexity varies across eight genres in the work of a single author, Karel Čapek, using five syntactic metrics. Taken together, the results show that genre is a strong predictor of syntactic profile even when authorial style is held constant. The patterns are far from random: each genre occupies a relatively stable position across the different indices, and these positions align in meaningful ways with the communicative function and audience design of the genres.

Across the metrics, academic prose and travelogues consistently cluster at the upper end of the complexity scale. They feature relatively long clauses and, in the case of travelogues in particular, deeply nested hierarchical structures. At the other end of the spectrum, novels and short stories are systematically less complex: they have shorter sentences, fewer and shorter clauses, and lower dependency and hierarchical distances. This profile fits their role as narrative genres that prioritise readability and narrative flow. Poetry stands somewhat apart: it is typically less complex in terms of average lengths and hierarchical depth, yet it shows the highest variability in most indices. This variability reflects the flexibility of poetic form, where syntactic shaping is strongly influenced by prosodic, rhythmic and stylistic choices.

Perhaps the most striking finding concerns children's literature. Contrary to common expectations about texts for young readers, Čapek's children's books are syntactically far from simplified. They show relatively long sentences, high numbers of clauses per sentence and surprisingly high linear and hierarchical complexity. This suggests that Čapek does not primarily adapt his style for children by reducing syntactic sophistication; instead, he may rely more on other resources – such as familiar topics or transparent lexis – to support comprehension. Personal correspondence and newspaper columns occupy intermediate positions: they permit more expansive and sometimes embedded sentences than narrative prose but remain less structurally extreme than academic writing or travelogues.

A key contribution of this study lies in the demonstration of Čapek's stylistic adaptability. Rather than maintaining a uniform syntactic style across genres, he modulates his writing with a high degree of genre sensitivity. On a broader methodological level, the study illustrates the added value of syntactic analysis in stylometry. While lexical and morphological features have long dominated computational literary studies, the findings here confirm that syntactic metrics offer unique and complementary insights. The combined use of hierarchical and linear syntactic features provides a more holistic view of stylistic variation, particularly when applied across structurally diverse genres.

In conclusion, this research not only sheds light on the genre-specific stylistic practices of Karel Čapek but also contributes to a deeper understanding of how syntax functions as a vehicle of genre differentiation. It opens up avenues for further work on cross-linguistic and multilingual corpora, comparative stylistics and the evolution of genre norms over time. Ultimately, it confirms that syntactic complexity is not merely a property of linguistic form, but a meaningful dimension of literary style, integral to the way texts communicate, persuade and resonate with their readers.

Acknowledgements

This research is supported by Grant SGS04/FF/2025, University of Ostrava.

References

Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), pp. 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Čech, R., Kubát, M. (2018) Morphological Richness of Text. In Fidler, M., Cvrček, V. (eds.). Taming the Corpus. *From Inflection and Lexis to Interpretation*. Springer, pp. 63–77.

Chen, X., Kubát, M. (2024). Quantifying Syntactic Complexity in Czech Texts: An Analysis of Mean Dependency Distance and Average Sentence Length Across Genres. *Journal of Quantitative Linguistics*, 31(3), pp. 260–273, https://doi.org/10.1080/09296174.2024.2370459

Eisen, M., Ribeiro, A., Segarra, S., Egan, G. (2018). Stylometric analysis of Early Modern period English plays, *Digital Scholarship in the Humanities*, 33(3), pp. 500–528, https://doi.org/10.1093/llc/fqx059

Futrell, R., Mahowald, K., Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages, *PNAS.* 112(33), pp. 10336–10341, https://doi.org/10.1073/pnas.1502134112

Gerdes, K., Guillaume, B., Kahane, S., Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In de Marneffe, M.-C., Lynn, T., Schuster, S. *Proceedings of the Second Workshop on Universal Dependencies* (UDW 2018), Brussels, Belgium. Association for Computational Linguistics, pp. 66–74.

Grieve, J. (2023). Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1), pp 47–77. https://doi.org/10.1515/cllt-2022-0040

Herrmann, J., B., Jacobs, A., M., Piper, A. (2021). Computational Stylistics. In Kuiken, D., Jacobs, A., M. (Eds.) *Handbook of Empirical Literary Studies*. De Gruyter, pp. 451–486. https://doi.org/10.1515/9783110645958-018

Jing, Y., Liu, H. (2015). Mean Hierarchical Distance Augmenting Mean Dependency Distance. In Nivre, J., Hajíčová, E. (Eds.) *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, Sweden, pp. 161–170.

Kestemont, M. (2014). Function Words in Authorship Attribution. From Black Magic to Theory?. In Feldman, A., Kazantseva, A., Szpakowicz, S. *Proceedings of the 3rd Workshop on Computational Linguistics for Literature* (CLFL), 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Kubát, M. (2016). Kvantitativní analýza žánrů. University of Ostrava.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), pp. 159–191. https://doi.org/10.17791/jcs.2008.9.2.159

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4), pp. 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college‐level ESL writers' language development. *TESOL quarterly*, 45(1), pp. 36–62. https://doi.org/10.5054/tq.2011.240859

Mann, H. B., Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 18, pp. 50–60.

Shapiro, S. S., Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52, pp. 591–611.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 197–207.

Wang, Y., Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59, pp. 135–147. https://doi.org/10.1016/j.langsci.2016.09.006

Zeman, D. et al. (2024). Universal Dependencies 2.15, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Prague: Charles University, Accessible at http://hdl.handle.net/11234/1-5787.

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	ɛ̃fɔrmasjõ	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]:

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

[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)

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

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, 舁く, 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).

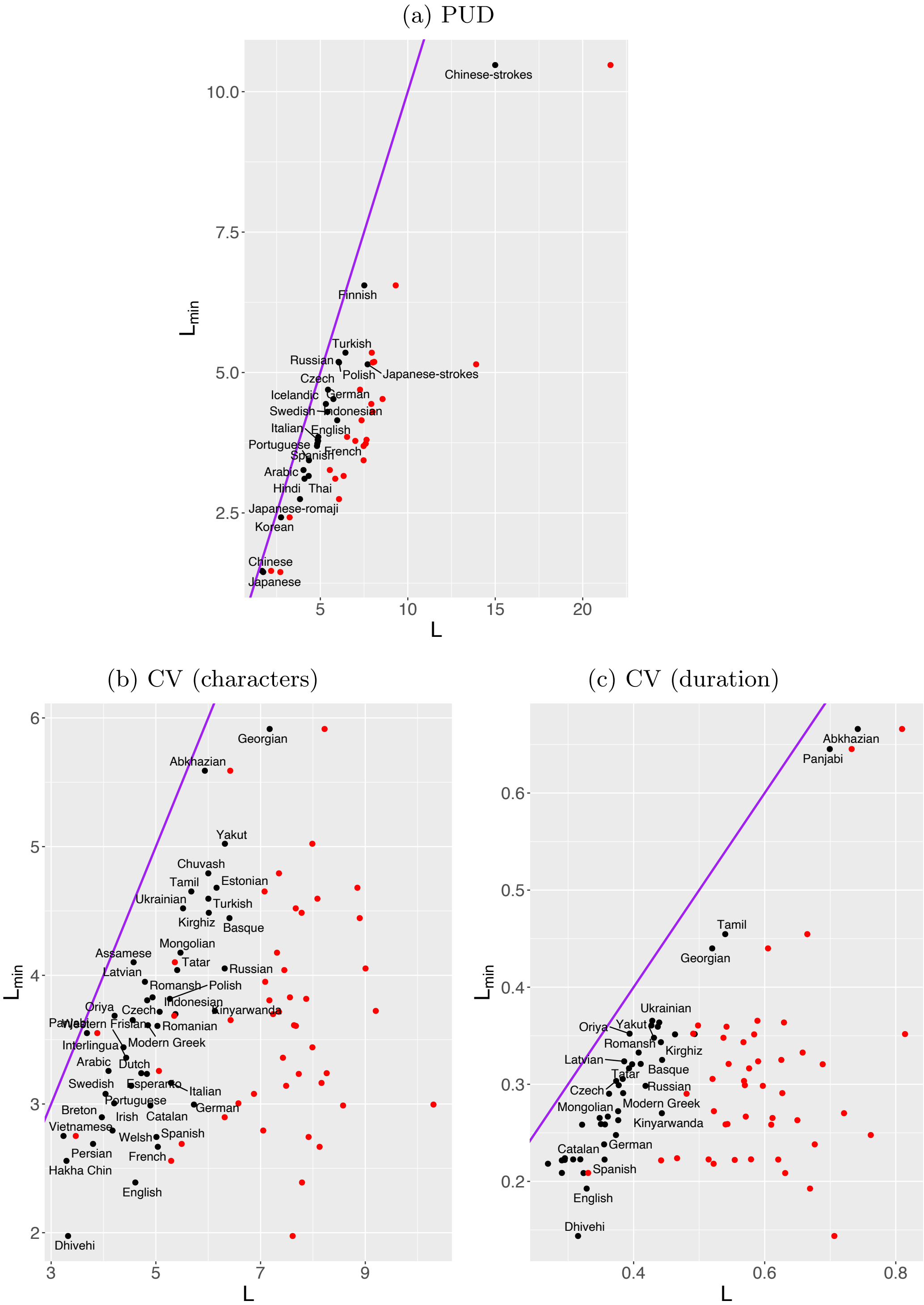


Figure 1: The ingredients of the ratio η: L_{min} (the minimum baseline) versus L (the actual word length). The purple line is the identity function (slope of 1 and zero intercept). Red points indicate the points that are expected under the null hypothesis (where $L = L_r$ is expected). (a) PUD collection.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

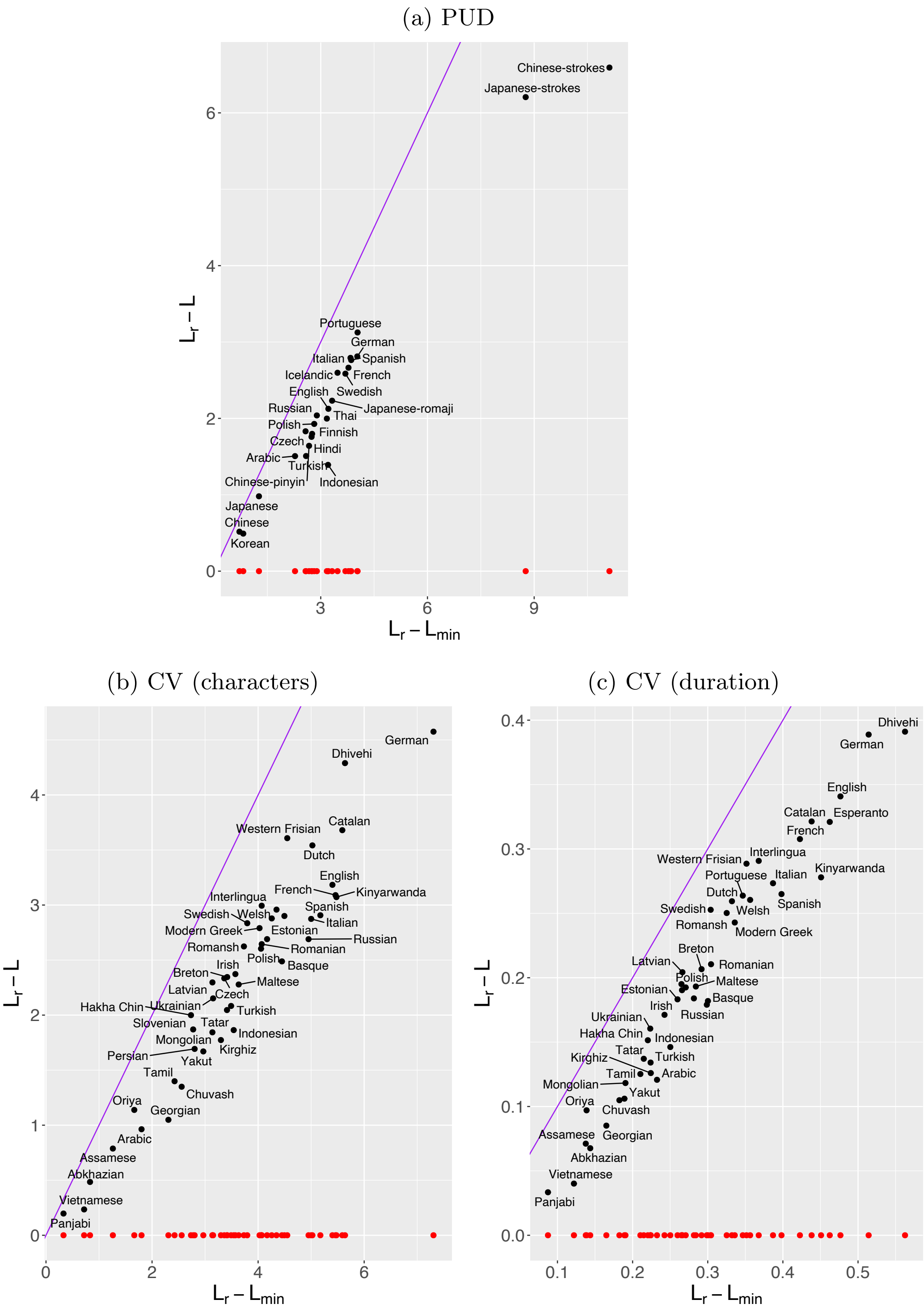


Figure 2: The ingredients of the ratio Ψ: the gap $L_r - L$ versus the gap $L_r - L_{min}$. The purple line is the identity function (slope of 1 and zero intercept). Red points indicate the points that are expected under the null hypothesis (where $L = L_r$ is expected). (a) PUD collection. (b) CV collection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

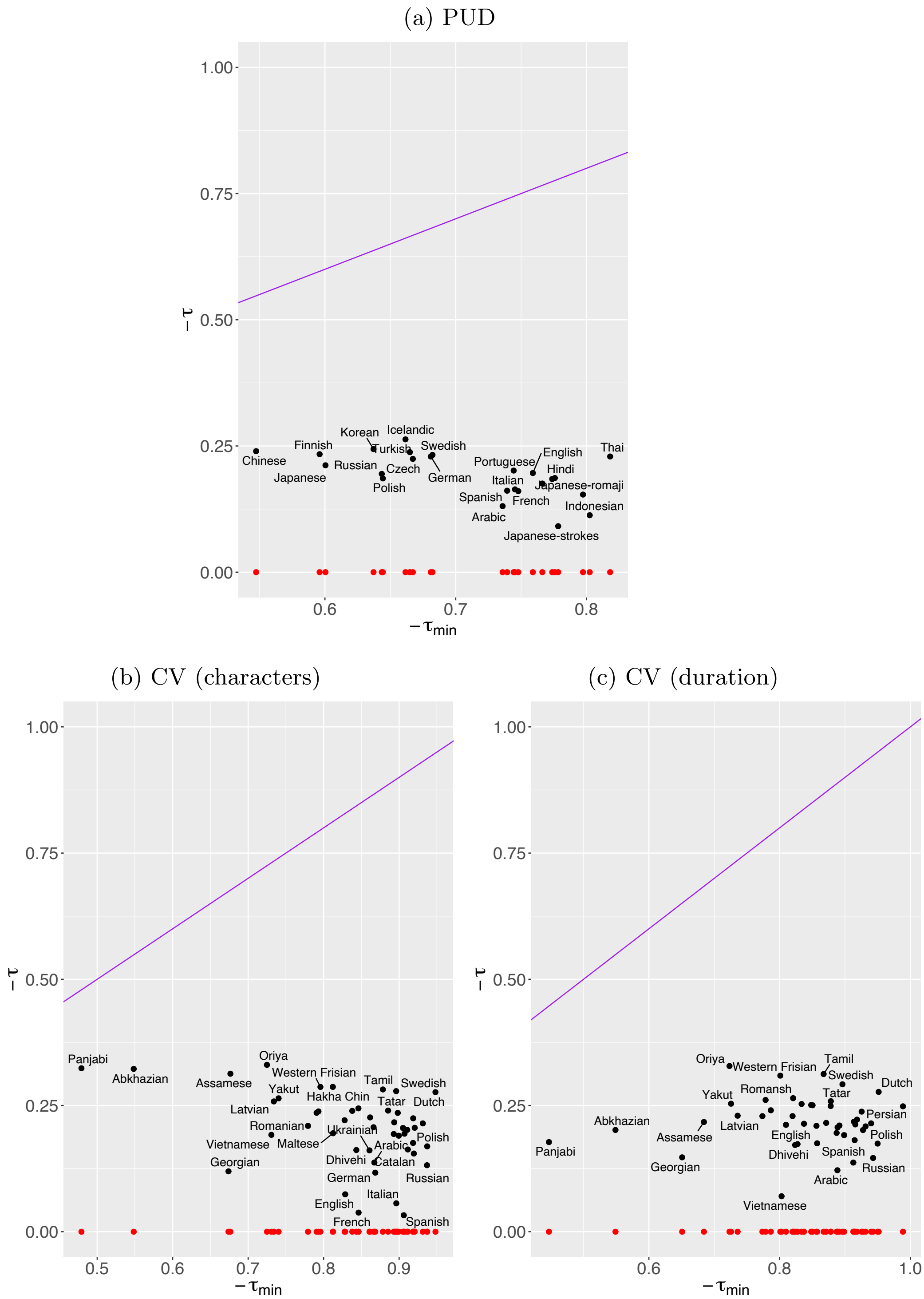


Figure 3: The ingredients of the ratio Ω: negative Kendall τ correlation between frequency and length, versus the negative τ_{min}, the minimum baseline for τ. Here we use the definition of Ω Equation 6 with $\tau_r = 0$, hence the minus sign in the correlations displayed by each axis. Red points indicate 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

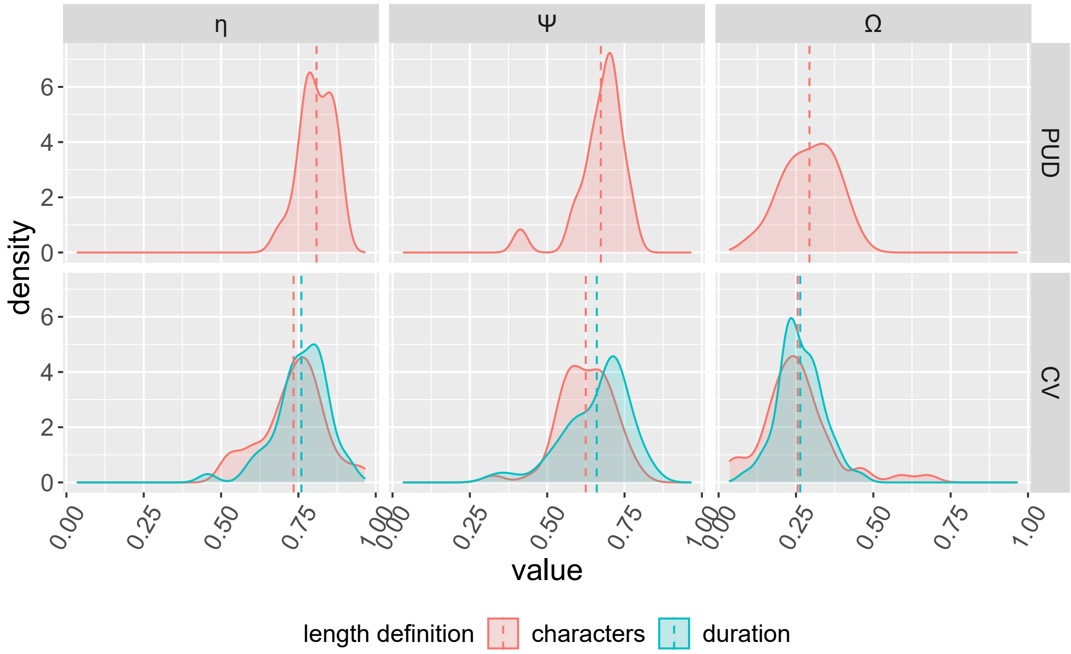


Figure 4: Density distribution of the scores η, Ψ and Ω for length in characters or duration over languages in the PUD and CV collections. For length in characters, we only use immediate word constituents for the sake of homogeneity. Accordingly, scores in strokes or in romanizations for Chinese and 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.

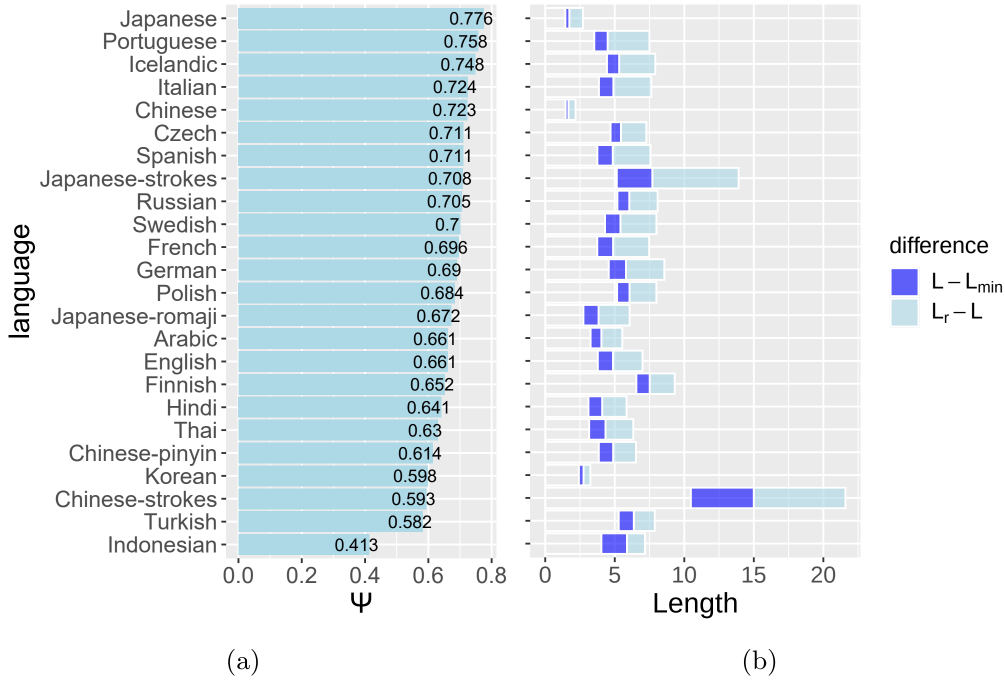


Figure 5: Optimality of word lengths in PUD according to Ψ with length measured in characters. (a) Absolute values of Ψ. (b) Visual depiction of the composition of $\Psi = (L_r - L)/(L_r - L_{min})$ using bars that consist of two segments, a dark blue one followed by a light blue one. The bar starts 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).

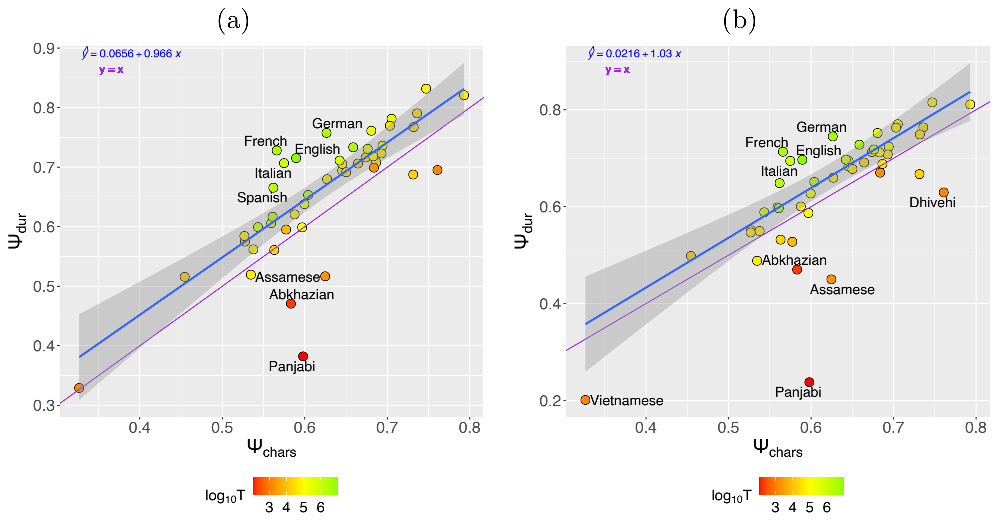


Figure 6: Ψ measured in duration (Ψ_{dur}) versus Ψ measured in characters (Ψ_{chars}) in the CV collection. Languages are color-coded by number of tokens in logarithmic scale to indicate orders of magnitude in sample size. We fit a robust linear model by Theil-Sen regression (blue line), which is then 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.

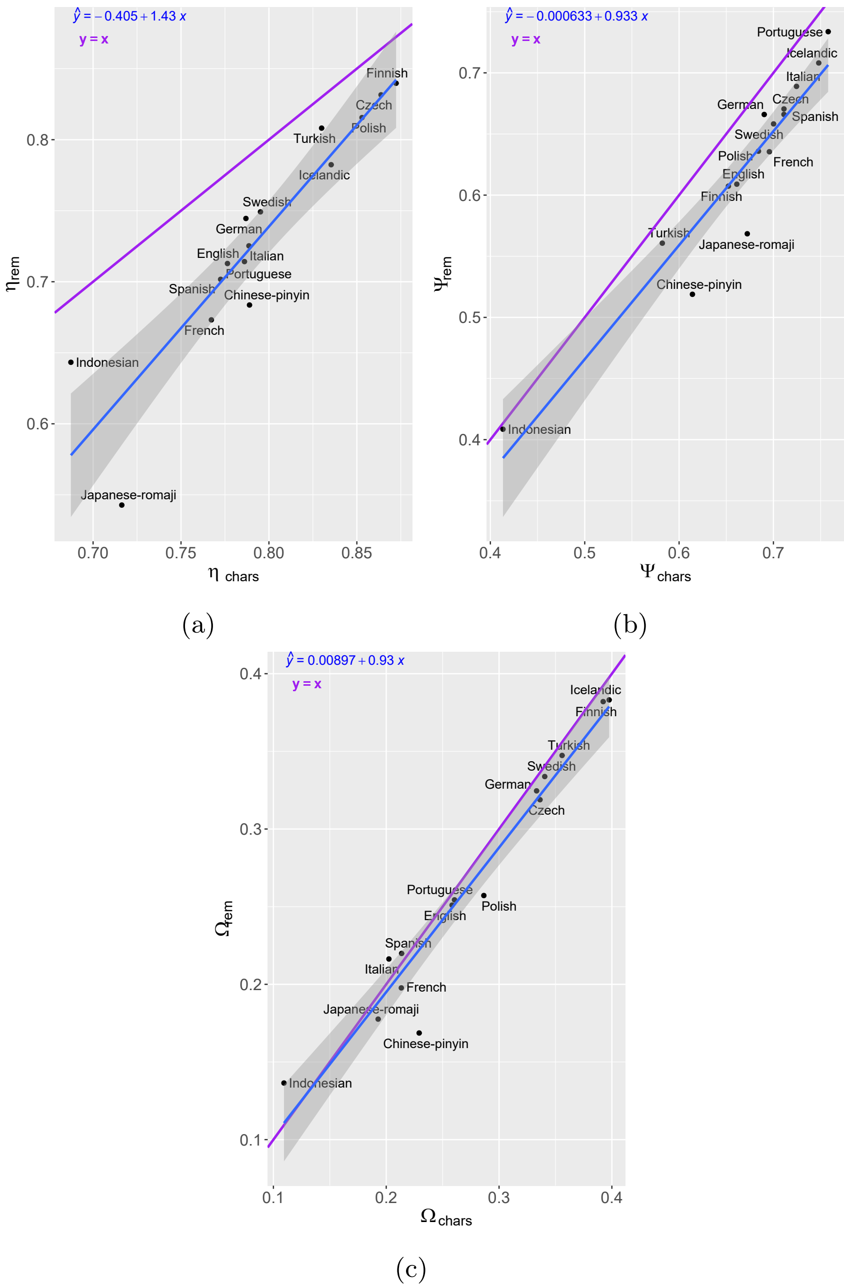


Figure 7: The value of a score after removing vowels, indicated with the subindex *rem* (e.g. Ψ_{rem}), as a function of the original value with all characters (e.g. Ψ_{chars}) in languages using the Latin script in the PUD collection. For a given language, the new value of the score is obtained after 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: パプアニューギニア ('Papua New Guinea'); フォルクスワーゲン ('Volkswagen'); ディジボーデンベルク ('Disibodenberg'); オーバーマルスベルク ('Obermarsberg'); エンターテインメント ('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
...	...	...	...	...	...	...	...	...
パプアニューギニア	3	9	conservationists	2	16	mendokumentasikannya	1	20
フォルクスワーゲン	2	9	catastrophically	1	16	mendokumentasikan	1	17
ディジボーデンベルク	1	10	hydroelectricity	1	16	mendemonstrasikan	1	17
オーバーマルスベルク	1	10	technologically	1	15	mempertemukannya	1	16
エンターテインメント	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^{RO}_{min} 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^{RO}_{min} coincides with the ordering of the l_i's in the matrix such that L is minimized, namely, $L = L^{RO}_{min}$ (Ferrer-i-Cancho, Bentz, and Seguin, 2022). $\tau^{RO}_{min} = -1$ if there are no ties both in the p_i's and the l_i's; otherwise (the common situation), $\tau^{RO}_{min} > -1$ (Section C, Property C.1).

τ^{RO}_{min} 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^{RO}_{min} be the minimum value of L under rank ordering, defined as*

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

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.

$\square$

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)

$$COV(aX + b, cY + d) = acCOV(X, Y)$$
$$\sigma(aX + b) = |a|\sigma(X)$$
$$\sigma(cY + d) = |c|\sigma(Y),$$

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*

$$
\begin{aligned}
\Psi = \Psi' \quad & \Omega = \Omega' \\
r = r' \quad & \tau = \tau' \\
\eta = \eta'' &
\end{aligned}
$$

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

$$
\begin{aligned}
L' &= aL + b \\
L'_r &= aL_r + b.
\end{aligned}
$$

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).

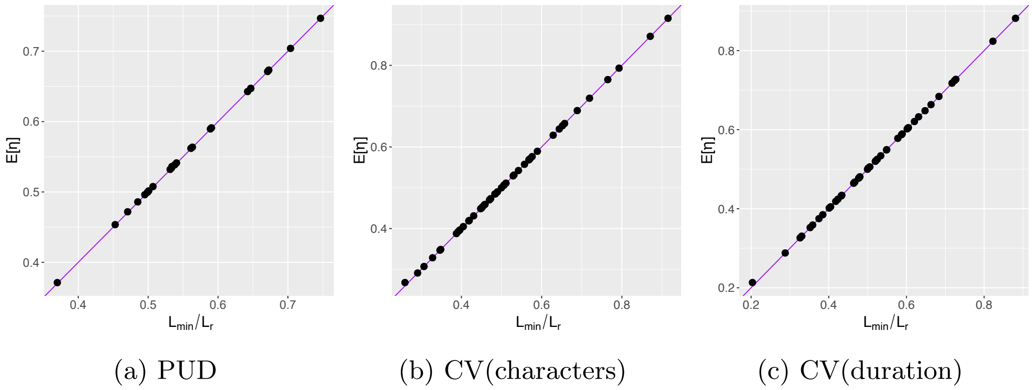


Figure D1: The Monte Carlo estimate of $\mathbb{E}[\eta]$ against its theoretical lower bound, namely $\frac{L_{min}}{L_r}$. The purple line indicates the identity function, $\mathbb{E}[\eta] = \frac{L_{min}}{L_r}$. (a) PUD collection with length measured in characters. (b) CV collection with length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

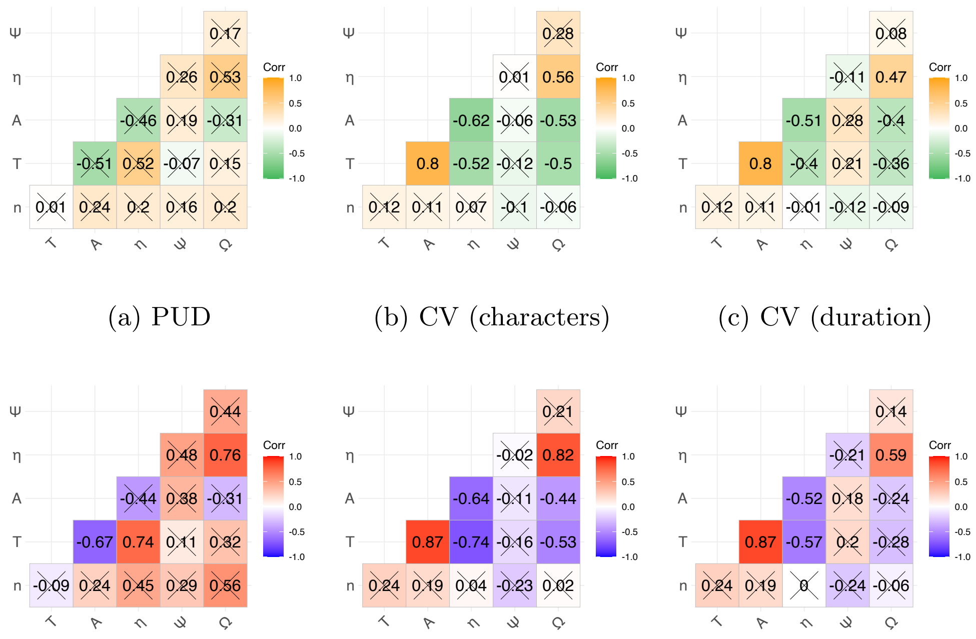
Appendix E Desirable properties of the scores

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.

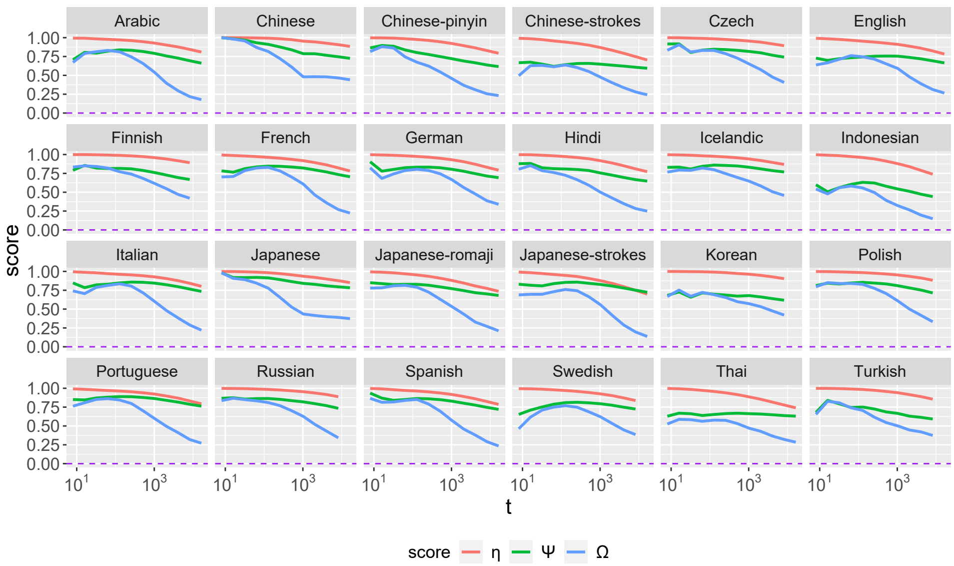


Figure E2: Convergence of the scores in the PUD collection. Values of the optimality scores (η, Ψ, and Ω) for increasing number of tokens t. For each value of t, we show the average value of the score over 10^2 random experiments (only those for which a value could be computed). The dashed line marks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

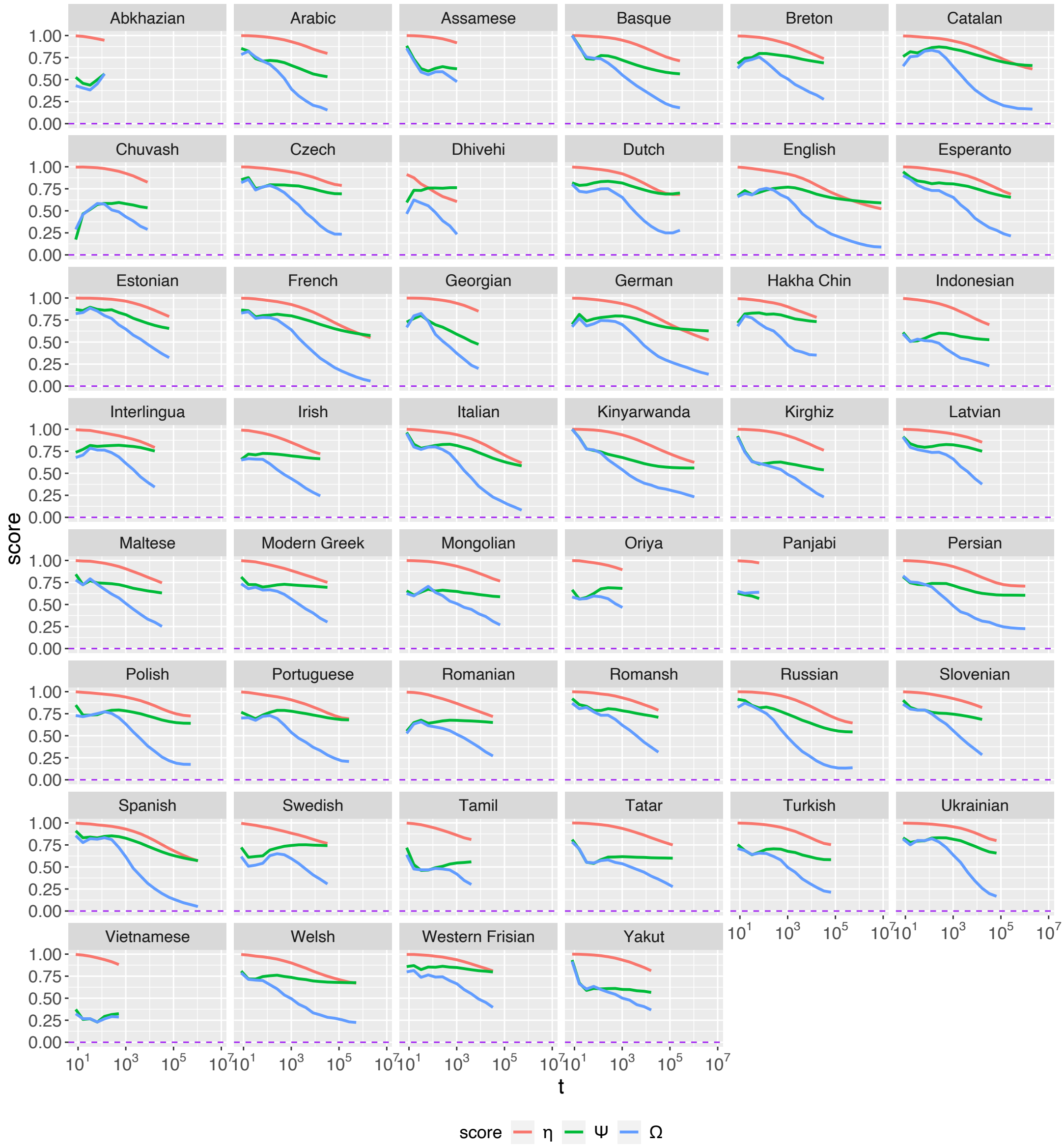


Figure E3: Convergence of the scores in the CV collection with word length measured in characters. The format is the same as in Figure E2.

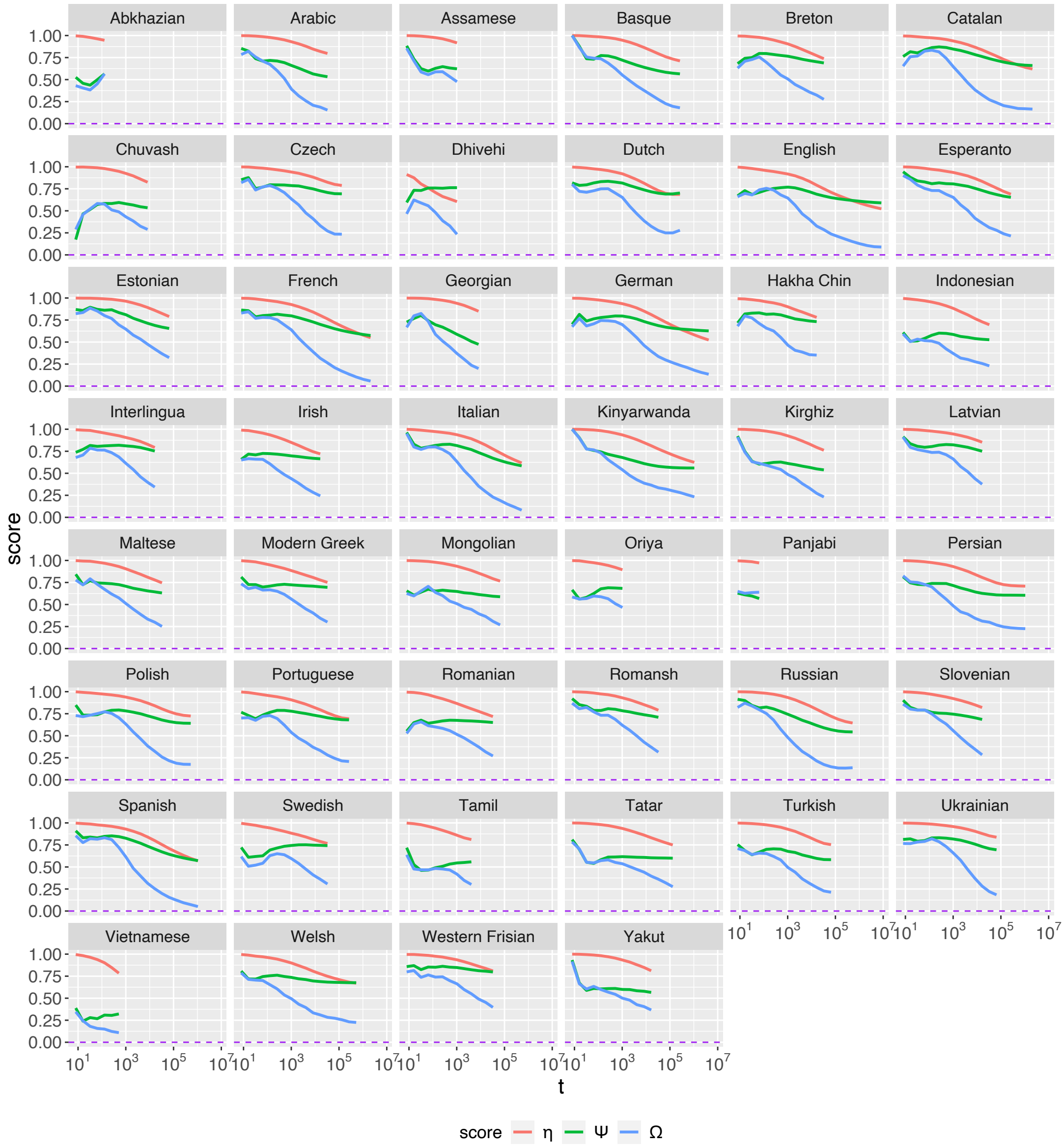
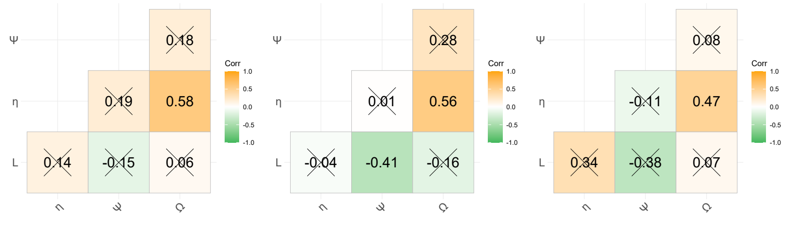


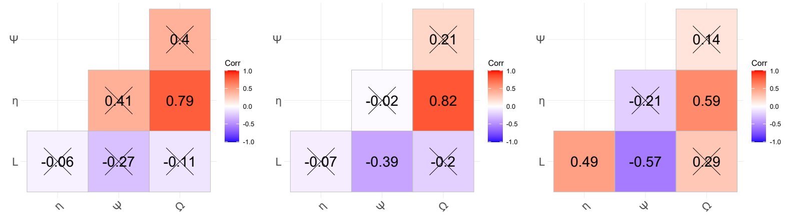
Figure E4: Convergence of the scores in the CV collection with word length measured in duration. The format is the same as in Figure E2.

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.