S5. Associations between and within languages

To evaluate whether the entropy-length trade-off occurs *between* languages but not *within* languages, we systematically adjust our estimates for the influence of text length, as this bias is especially relevant within languages [1–3]. Since there is a direct correspondence between entropy rates and text length at the word and character levels, but not for BPE, we focus on words and characters as information encoding units. Additionally, we want to ensure that the observed patterns did not arise merely from the way we statistically inferred entropy rates using LMs. Therefore, here we thus use the non-parametric entropy rate estimator by [4] that was already used in several studies [5–9] and that is computed as

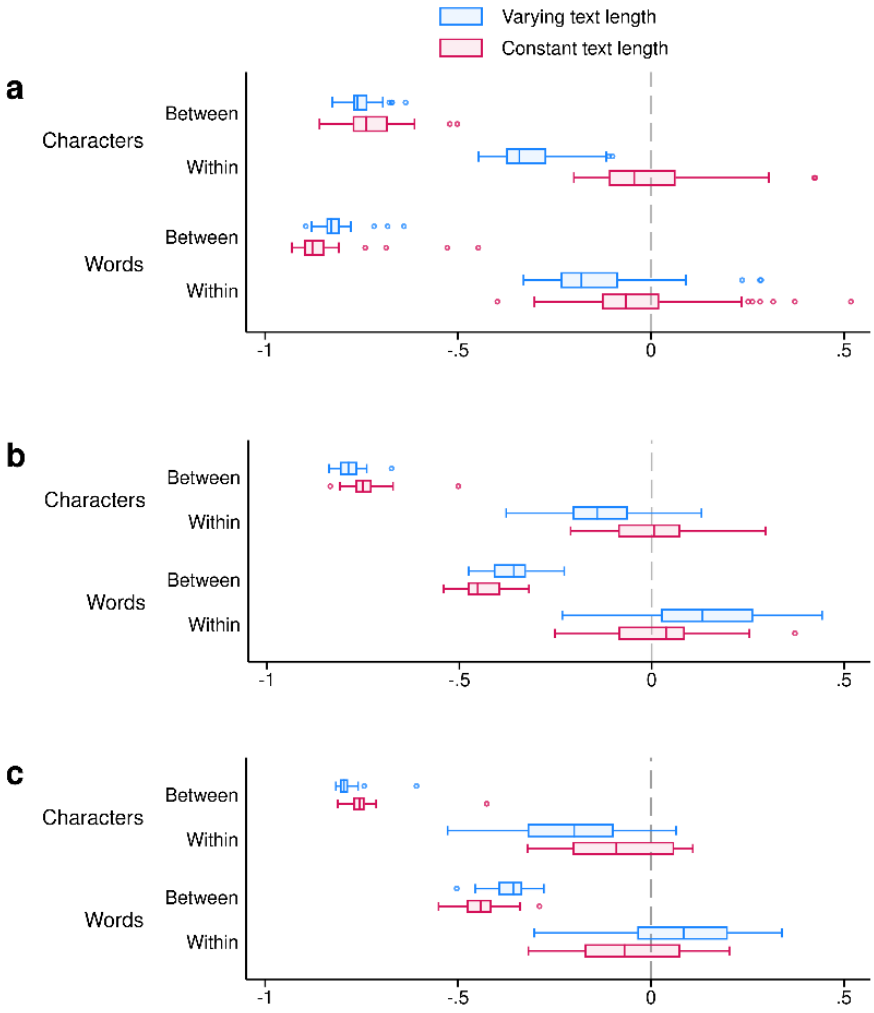
$$\hat{h}^K(\kappa) = \left[\frac{1}{L}\sum_{l=2}^{L}\frac{\Lambda_l}{\log_2(l)}\right]^{-1} \quad (1)$$

The key quantity is the match-length Λ_l. It measures the length (in symbols) of the shortest substring starting at position l that is not also a substring of the part of the corresponding document κ before this position and can be used to estimate h, since it was shown that Λ_l grows like $\log_2 l/h$ [4,10,11]. More details and an open source Java program to efficiently obtain match-lengths in texts can be found in [8].

As a first analysis, we use the PBC (cf. Sect. 2.1.1) and split each translation into the 66 different books of the Biblical canon. We only kept translations with available information for all the 66 books. We then kept all 29 books with a median length of at least 10,000 words. For languages with more than one available Bible translation, we randomly sampled one translation. In total, we have available translations for 144 different languages. We then compute two sets of correlations: (i) Pearson correlations, ρ_{within}, between entropy rates, $\hat{h}^K$, and L *within* languages. For each of 144 languages, we have 29 Bible books. Per language, we compute the Pearson correlation between $\hat{h}^K$ and L. (ii) Pearson correlations, adjusted for geographical proximity, ρ_{geo}, between $\hat{h}^K$ and L *between* languages for 29 Bible books, each with parallel translations into 144 languages.

To account for a potential text length bias, we first compute the minimum text length in symbols (words/characters) per language across the 29 Bible books and call that minimum $ג_b$. Similarly, we compute the minimum text length in symbols per Bible book across the 144 languages and call that minimum $ג_w$. We then truncated each book at the respective minima and used the truncated books to calculate the corresponding entropy rates $\hat{h}^K_{ג_b}$ and $\hat{h}^K_{ג_w}$. We repeated (i) and (ii) with these truncated entropy rates as input. Supplementary Figure 2 shows that the trade-

off only holds *between*, but not *within* languages: *Within* languages (144 languages, each with 29 Bible books from the same Bible translation), there is only weak evidence for a trade-off if we do not control for text length. As soon as we add that control, any evidence for a trade-off disappears. *Between* languages (29 Bible books, each with parallel translations into 144 languages), there is clear evidence for a trade-off, regardless of whether we control for text length.



Supplementary Figure 2: Evaluating the entropy-length trade-off *between* and *within* languages. Per symbolic level (characters, words), we compute (i) ρ_{within}, between $\hat{h}^K$ and L *within* languages and (ii) ρ_{geo} between $\hat{h}^K$ and L *between* languages. To account for potential text length bias, we compute entropy rates for text samples of identical size ($\hat{h}^K_{\lambda_w}$, $\hat{h}^K_{\lambda_b}$) and use those as input to calculate ρ_{within} and ρ_{geo}. Blue colour: Distribution of ρs with $\hat{h}^K$ as input. Red colour: Distribution of ρs with $\hat{h}^K_{\lambda_w}$ or $\hat{h}^K_{\lambda_b}$ as input. (**a**) Distribution of ρs for 29 books of the Biblical canon in 144 different languages between and within languages. (**b** and **c**) Distribution of ρs for 43 different text samples in 43 languages between and within languages. (**b**) Parallel text samples. (**c**) Non-parallel text samples.

We continue by demonstrating that an entropy-length trade-off *between* languages also occurs when the content of the message is not fully controlled/parallel. For this, we used the Quran corpus consisting of parallel translations of 6,233 sentences/verses into 43 different languages (see Sect. 2.1.1). We randomly distributed the sentences into $i = 1, 2, \ldots, 43$ different samples $s_1, s_2, \ldots, s_{43}$ of approximately equal size where, across languages, each sample consists of the same sentences that are arranged in the same order. Thus, each s_i in each language *ceteris paribus* contains a different message with the statistical characteristics of the source text, i.e., the corresponding Quran translation. We then continue as above to compute ρ_{within}. Supplementary Figure 2 shows that w*ithin* languages (43 languages, each with 43 different text samples), there is no noteworthy negative correlation between entropy and length on either symbolic level, irrespective of whether correlations are unadjusted or adjusted for text length. Hence, *within* languages, i.e., across comparable samples in a given language, there is no evidence that higher entropy is compensated by shorter length. We then randomly re-arranged the data into a *between*-languages format by assigning one randomly chosen s_i from each language to one of 43 different text collections, so that each text collection contains 43 different samples, one from each language. Supplementary Figure 2 demonstrates that here results point again towards a trade-off between *h* and *L between* languages (parallel documents in different languages) on both levels and for both types of text length adjustments.

We round off this section by eliminating the parallelism across languages. Instead of preparing 43 different language-specific samples that consist of the same set of sentences in the same order across languages, we randomly distribute the sentences across languages without maintaining any parallelism across languages. This approach tests whether the entropy-length trade-off persists in a completely randomized context. Supplementary Figure 2 shows that, after this random distribution, the results are fully comparable once again: there is no evidence for a trade-off between entropy and length *within* languages. However, a trade-off is observed *between* languages on both symbolic levels and for both types of text length adjustments.

These findings suggest that the entropy-length trade-off is a robust phenomenon primarily observed *between* languages rather than *within* languages.

References

1. Baayen RH. Word Frequency Distributions. Dordrecht: Kluwer Academic Publishers; 2001.

2. Tweedie FJ, Baayen RH. How Variable May a Constant be? Measures of Lexical Richness in Perspective. Computers and the Humanities. 1998;32: 323–352.

3. Koplenig A, Wolfer S, Müller-Spitzer C. Studying Lexical Dynamics and Language Change via Generalized Entropies: The Problem of Sample Size. Entropy. 2019;21. doi:10.3390/e21050464

4. Kontoyiannis I, Algoet PH, Suhov YuM, Wyner AJ. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. IEEE Transactions on Information Theory. 1998;44: 1319–1327. doi:10.1109/18.669425

5. Bentz C, Alikaniotis D, Cysouw M, Ferrer-i-Cancho R. The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. Entropy. 2017;19: 275. doi:10.3390/e19060275

6. Montemurro MA, Zanette DH. Universal Entropy of Word Ordering Across Linguistic Families. Breakspear M, editor. PLoS ONE. 2011;6: e19875. doi:10.1371/journal.pone.0019875

7. Koplenig A. Quantifying the efficiency of written language. Linguistics Vanguard. 2021;7: 20190057. doi:10.1515/lingvan-2019-0057

8. Koplenig A, Meyer P, Wolfer S, Müller-Spitzer C. The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. Smith K, editor. PLOS ONE. 2017;12: e0173614. doi:10.1371/journal.pone.0173614

9. Koplenig A. Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. Royal Society Open Science. 2019;6: 181274. doi:10.1098/rsos.181274

10. Wyner AD, Ziv J. Some Asymptotic Properties of the Entropy of a Stationary Ergodic Data Source with Applications to Data Compression. IEEE Trans Inf Theor. 1989;35: 1250–1258. doi:10.1109/18.45281

11. Ornstein DS, Weiss B. Entropy and Data Compression Schemes. IEEE Trans Inf Theor. 1993;39: 78–83. doi:10.1109/18.179344