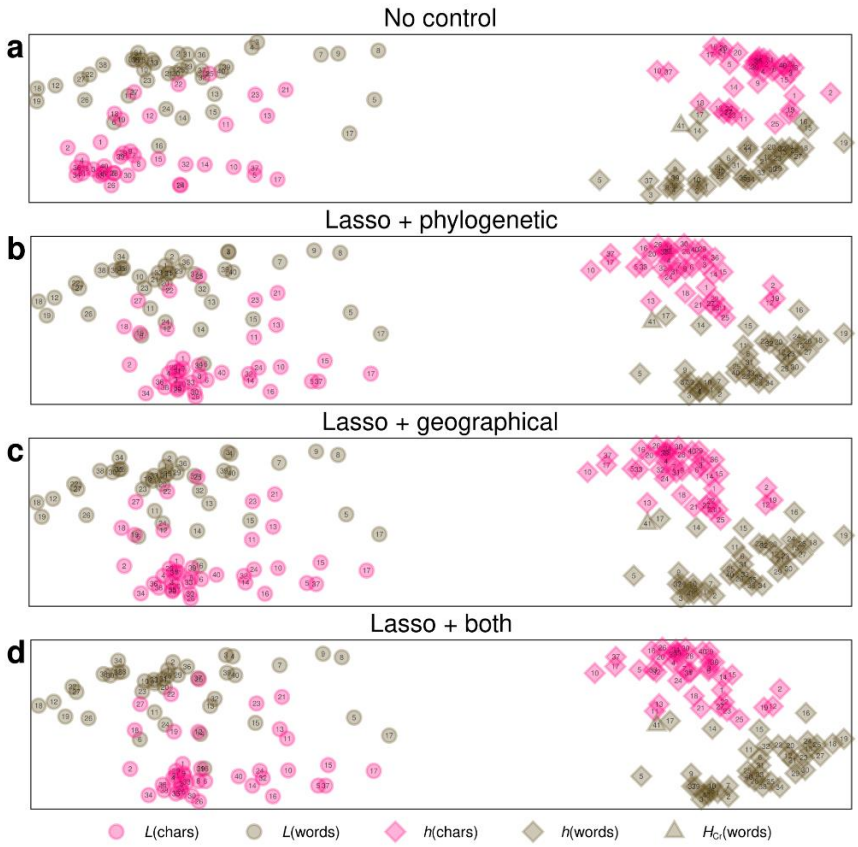
S4. Testing for a potential systematic length bias

In Sect. 3.1.2, we showed that, across corpora and LMs, languages that tend to have a higher entropy rate *h* tend to require fewer symbols to encode messages, i.e., have a lower *L*. Given the almost ubiquitous influence of text length in both corpus and quantitative linguistics, we now test whether the obtained pattern between *h* and *L* holds when we explicitly adjust our estimate for a possible text length bias. To this end, we used additional data that we made available as part of our prior publication [1] in which we prepared a truncated version by first computing λ_{corpus} that denotes the minimum document length in symbols per corpus and per type (characters or words). We then extrapolated entropy rates [1] based on the first λ_{corpus} symbols for each available text that belongs to the corresponding corpus. This procedure ensures that per corpus and per type, *h* is computed for text samples of identical size. We then compute cross-correlations between *h* and *L* across corpora and type as described in Sect. 2.6.1. Supplementary Figure 1 shows that the results for the truncated version are qualitatively identical to the results presented in Sect. 3.1.2.



Supplementary Figure 1: Comparing entropy and length across corpora. To rule out that the above results are mainly driven by a potential text length bias, entropy rates taken from [1] are computed for text samples of identical size per corpus and per type (words, characters). For each correlation type, N_ρ = 12,880 individual correlations are computed. Refer to the caption of Figure 5 for further details.

References

1. Koplenig A, Wolfer S, Meyer P. A large quantitative analysis of written language challenges the idea that all languages are equally complex. Sci Rep. 2023;13: 15351. doi:10.1038/s41598-023-42327-3