Human languages trade off complexity against efficiency

Alexander Koplenig (koplenig@ids-mannheim.de)
Leibniz Institute for the German Language (IDS)  https://orcid.org/0000-0002-9630-9680

Sascha Wolfer
Institute for the German Language

Peter Meyer
Leibniz Institute for the German Language (IDS)

Article

Keywords:

Posted Date: March 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1462001/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Human languages trade off complexity against efficiency

Authors: Alexander Koplenig\textsuperscript{1,*}, Sascha Wolfer\textsuperscript{1}, Peter Meyer\textsuperscript{1}

\textsuperscript{1}Leibniz-Institute for the German Language (IDS), Mannheim, Germany.
*Corresponding author. Email: koplenig@ids-mannheim.de

Abstract: One of the fundamental questions about human language is whether all languages are equally complex. To answer this long-standing question, we conduct a large scale quantitative cross-linguistic analysis of written language by training a language model on more than 6,500 different documents as represented in 41 multilingual text collections consisting of \( \sim3.5 \) billion words or \( \sim9.0 \) billion characters and covering 2,069 different languages that are spoken as a native language by more than 90\% of the world population or \( \sim46\% \) of all languages that have a standardized written representation. Statistically inferring the entropy of each language-model as an index of (un)predictability/complexity allows us to refute the equi-complexity hypothesis, but also unveils a previously undocumented complexity-efficiency trade-off: high entropy languages are information-theoretically more efficient because they tend to need fewer symbols to encode messages. Our findings additionally contribute to debates about language evolution/diversity by showing that this trade-off is partly shaped by the social environment in which languages are being used.
Language is one of our most complex traits\textsuperscript{1}. But how complex is it? And are all of the \textasciitilde{7,000} distinct languages on earth equally complex – or not\textsuperscript{2}? Quantifying the statistical structure and complexity of human language is essential to understanding a large variety of phenomena in linguistics, the study of human culture and natural language processing from language learning to language evolution and from the role of culture in shaping cognitive skills to the creation of artificial intelligence\textsuperscript{1,3–11}. Information theory, an area of mathematics that links probability and communication\textsuperscript{12}, provides a mathematical framework for dealing with the above-mentioned phenomena. Providing a blueprint for the quantitative study of the statistical character of language, Shannon\textsuperscript{13,14} showed that prediction, compression and understanding are intimately related\textsuperscript{15} by demonstrating that humans are very good (in fact until very recently much better than any machine\textsuperscript{7,16,17}) in predicting subsequent linguistic material based on previous input. Since Shannon’s work, numerous studies have revealed that adults, children and even infants show an extraordinary ability to (unconsciously) exploit statistical information on different levels in the input they receive in order to efficiently process language\textsuperscript{18–20}. Text compression is equivalent to machine learning of natural languages\textsuperscript{8} and can thus be used to quantitatively study human language\textsuperscript{21–23}. One of the key quantities in information theory is the average per-symbol information content or entropy rate $h$\textsuperscript{12}: since, due to grammatical, phonological, lexical and other regularities governing language use, not every sequence of symbols is allowed\textsuperscript{21}, $h$ both (i) measures how much choice a writer has when selecting successive symbols and (ii) approximates the reader’s uncertainty about upcoming symbols\textsuperscript{13}. Kolmogorov showed that $h$ can also be interpreted as a complexity metric: the harder a text is to predict the greater is its complexity\textsuperscript{24}.

Against this background, parallel corpora offer an intriguing source of data because they can be considered translational equivalents\textsuperscript{25}: parallel texts are basically texts in different languages containing the same message, but are different regarding the used language. Therefore,
information theory can be used to study cross-linguistic complexity variation, since obtained
differences cannot be attributed to differences in content, style or register\textsuperscript{26,27} (for a discussion of
potentials confounds see refs.\textsuperscript{25,27–31}). Leveraging available corpora and multilingual text
collections\textsuperscript{32–35}, we compiled a database of parallel texts comprising a large variety of different
text types, e.g. religious texts, legalese texts, subtitles for various movies and talks, or machine
translations. Where necessary (and possible), we developed computational routines that made
sure that the resulting corpora are as parallel as possible (see Supplementary Information (SI):
Corpora for details). In addition, we added comparable corpora, i.e. texts that are not parallel but
come from comparable sources and are therefore similar in content, again comprising very
different text types/genres, e.g. newspaper texts, web crawls, Wikipedia articles, Ubuntu
localization files, or translated example sentences from a free collaborative online database.
Furthermore, we calculated Gibbs-Shannon unigram entropies $H$ (see equation (1)) based on
word frequency information from the Crúbadán project\textsuperscript{36} that aims at creating text corpora for a
large number of (especially under-resourced) languages. In total, we analysed 41 different
multilingual corpora (Figure 1a) by compressing $\sim$30.2 million strings of 6,513 documents
consisting of $\sim$3.5 billion words or $\sim$9.0 billion characters covering 2,069 different languages that
are spoken as a native language by more than 90\% of the world population or $\sim$46\% of all
languages that have a standardized written representation (see SI: Coverage). In Figures 1b-e, we
further describe our database (also see Supplementary Table 3). Figure 1b shows that most
corpora only consist of a few tens of texts [$N_{\text{median}} = 40$]. For some corpora, the reason for this is
rather simple, e.g. the European constitution was only translated into the languages of the
European Union. However, for other corpora, e.g. the 13 subtitle corpora, translations into further
languages are not available. On the other side of the spectrum, we have 11 corpora that consist of
more than 100 different documents. Figure 1c complements this observation by showing that our
database is also unbalanced at the language level: while we have more than 100 languages with at least 10 available data points, i.e. documents, we only have less than four available data points for most languages (~84%). This reflects the fact that especially for languages that are spoken only by a small number of people, there exists only a very limited number of documents that are electronically available\(^\text{36}\). Correspondingly, Figure 1d shows that our database is biased towards languages with more speakers. For example, while the estimate for median number of speakers for all documented languages is 8,000, the median number for which we have available data is 30,000. Finally, Figure 1e shows that many documents are rather short, e.g. 25% of the documents are below 14,575 characters or 3,181 words. However, 200 documents are longer than 1 million characters, 49 documents are longer than 10 million characters and the longest documents are several hundred million words or more than a billion characters long. In what follows, we adapt our analysis strategy accordingly by both using state-of-the-art statistical methods that allow for unbalanced datasets and by statistically comparing the diversity structure found in smaller corpora (i.e. corpora consisting of shorter documents and/or corpora with only a limited number of available documents) with the underlying structure found in bigger corpora (i.e. longer documents and/or available data points for many languages). Here, the idea is that if the results in both smaller and bigger corpora point in the same direction, then this strengthens the claim that those results are more than just an artefact resulting from the unbalancedness of the database. In addition, we draw balanced subsets from our database and use both parametric and non-parametric methods to evaluate the results.

Figure 1f visualizes how we adapt Shannon’s information-theoretic view of communication to analyse parallel corpora: \([1]\) a source message is \([2]\) encoded into a signal, i.e. translated into different languages.\(^\text{37}\) \([3]\) To estimate the average per-symbol information content, we use Prediction by partial matching (PPM), a prominent computational language model originally
developed for data compression\textsuperscript{38,39} to calculate the compression rate $r$ (compressed size/message length $L$ in symbols) as an index of (un)predictability/complexity for both words and characters as information encoding units.\textsuperscript{37} [4] Since $r$ only provides upper bounds for $h$, we calculate $r$ for subsets of increasing length (red circles) and fit a nonlinear ansatz\textsuperscript{40,41} (yellow line) with interval constraints by log-least squares where initial values are approximated in linear space to estimate the asymptotic value of $h$ (blue line; see Methods: Estimating entropy and SI: Ansatz functions for details and Supplementary Figures 1a-c for illustrations of our statistical estimation approach).

[FIGURE 1]

Figure 1 | Dataset and study design. (a) Collected corpora and their geographical distribution. Asterisks indicate fully parallel corpora (see Supplementary Information: Corpora for details). (b) Number of documents per corpus. Corpora are ranked by the number of available documents. (c) Number of documents per language. Languages are ranked by the number of available data points. (d) Kernel density estimations of the number of speakers per language (logged). Emerald line – all existing languages. Cranberry line – languages for which we have available documents. Dotted vertical lines correspond to median values. (e) Distribution of document length in characters and words. (f) Adaption of Shannon’s information-theoretic view of communication for the analysis of parallel corpora.

Results

Idealized language learning. $PPM$ can be seen as very simple model of an idealized language learner\textsuperscript{21,42,43}: in order to compress, $PPM$, a variable-order Markov model, uses a set of previous symbols as context to predict the most probable next symbol based on the conditional probability distribution generated from the text it has already observed. It is a dynamic and adaptive method: without prior knowledge of the source, it acquires a representation of the probabilistic structure of the input in one single pass. In Table 1, we illustrate that every time $PPM$ encounters new text, it updates its language-model (i.e. its probability distribution) and with growing input, it gets better in compressing/predicting subsequent linguistic data\textsuperscript{23}, or put differently, $PPM$ learns to exploit the statistical structure of the input, paralleling human language learning\textsuperscript{20,21} with interesting
applications in natural language processing\textsuperscript{45,46}, language production\textsuperscript{47,48} and – more generally – machine learning of patterns to predict (into) the future\textsuperscript{8,48,49}.

Table 1 | Illustration of PPM in action. 1st column: on the level of words, we trained a PPM model of order 5 to predict the next 5 words; on the level of characters, we trained a PPM model of order 10 to predict the next 30 characters [prediction also ends when the end of a sentence is reached; both models were smoothed by linearly interpolating all probabilities, i.e. predictions of $n$-grams of all orders are blended together. Higher weights are assigned to $n$-grams of higher order, where individual weights are determined by calculating the number of different words observed after a specific $n$-gram]. 2nd column: models were trained on the entire Project Gutenberg works of Mark Twain\textsuperscript{44}. Size refers to the number of sentences used for training. 3rd column: the context is presented in bold face, predictions are in regular font. Characters were mapped to lower-case and punctuation marks were treated as words. Output was manually corrected for capitalization to improve readability. Further details and – given that the examples are cherry-picked – training data, trained (unsmoothed and smoothed) $n$-gram models and an open source Java program to interactively test PPM are available at https://osf.io/f5mke/.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Size</th>
<th>Context</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>$10^1$</td>
<td>I was ashamed of all the swarms that come ...</td>
<td>I was ashamed of all the swarms that come ...</td>
</tr>
<tr>
<td></td>
<td>$10^2$</td>
<td>I was ashamed of himself he did bet half ...</td>
<td>I was ashamed of himself he did bet half ...</td>
</tr>
<tr>
<td></td>
<td>$10^3$</td>
<td>I was ashamed of the country to part with ...</td>
<td>I was ashamed of the country to part with ...</td>
</tr>
<tr>
<td></td>
<td>$10^4$</td>
<td>I was ashamed of myself and felt unspeakably ridiculous ...</td>
<td>I was ashamed of myself and felt unspeakably ridiculous ...</td>
</tr>
<tr>
<td></td>
<td>$10^5$</td>
<td>I was ashamed of my performance at the time ...</td>
<td>I was ashamed of my performance at the time ...</td>
</tr>
<tr>
<td>Character</td>
<td>$10^1$</td>
<td>It reminds to the will of god.</td>
<td>It reminds to the will of god.</td>
</tr>
<tr>
<td></td>
<td>$10^2$</td>
<td>It reminds, and then wilted to the ...</td>
<td>It reminds, and then wilted to the ...</td>
</tr>
<tr>
<td></td>
<td>$10^3$</td>
<td>It reminds, and the rest of the ...</td>
<td>It reminds, and the rest of the ...</td>
</tr>
<tr>
<td></td>
<td>$10^4$</td>
<td>It reminds one that it was not ...</td>
<td>It reminds one that it was not ...</td>
</tr>
<tr>
<td></td>
<td>$10^5$</td>
<td>It reminds me of a tale that ...</td>
<td>It reminds me of a tale that ...</td>
</tr>
<tr>
<td>Character</td>
<td>$10^1$</td>
<td>He died and made him feel good.</td>
<td>He died and made him feel good.</td>
</tr>
<tr>
<td></td>
<td>$10^2$</td>
<td>He died and registered their names on t ...</td>
<td>He died and registered their names on t ...</td>
</tr>
<tr>
<td></td>
<td>$10^3$</td>
<td>He died and took the crown of thorns.</td>
<td>He died and took the crown of thorns.</td>
</tr>
<tr>
<td></td>
<td>$10^4$</td>
<td>He died and was buried in a couple of f ...</td>
<td>He died and was buried in a couple of f ...</td>
</tr>
<tr>
<td></td>
<td>$10^5$</td>
<td>He died at the same time the ship was t ...</td>
<td>He died at the same time the ship was t ...</td>
</tr>
<tr>
<td>Character</td>
<td>$10^1$</td>
<td>So we speriences of my life, this secre ...</td>
<td>So we speriences of my life, this secre ...</td>
</tr>
<tr>
<td></td>
<td>$10^2$</td>
<td>So we speed of two or three hours the s ...</td>
<td>So we speed of two or three hours the s ...</td>
</tr>
<tr>
<td></td>
<td>$10^3$</td>
<td>So we spent part of a book of mine on t ...</td>
<td>So we spent part of a book of mine on t ...</td>
</tr>
<tr>
<td></td>
<td>$10^4$</td>
<td>So we spent a week there, at the very s ...</td>
<td>So we spent a week there, at the very s ...</td>
</tr>
<tr>
<td></td>
<td>$10^5$</td>
<td>So we spent a moment or two and a half ...</td>
<td>So we spent a moment or two and a half ...</td>
</tr>
</tbody>
</table>

Figures 2a,b visualize this online learning behavior for the United Nations Parallel Corpus (UNPC)\textsuperscript{50} consisting of various documents in the six official languages of the United Nations. From a linguistic point of view, the entropy rate determines how hard it is to make accurate predictions once the statistical structure of the input language has been learned\textsuperscript{40}. Therefore, when estimating the entropy rate via compression, it is essential to take into account that the compression algorithm needs (a certain amount of) input to learn how to exploit the statistical structure of the input in order to make accurate predictions. As written above, we therefore use an
ansatz to estimate asymptotic entropy rates. Figures 2a,b visually indicate that our ansatz fits the observed curves very well (solid lines; see Supplementary Figures 2 – 19 for all corpora). Figure 2c confirms this impression: 99% of the ~5,000 compression series show an accuracy ratio, i.e. an approximate average percentage difference, between (held-out) observed and predicted values that is within 1% (see equation (8)). The ansatz has three parameters (see equations (5) – (7)): the limiting entropy rate \( h \), a proportionality constant and an exponent \( b \). While \( h \) quantifies how difficult it is to predict, \( b \) quantifies how difficult it is to learn to predict, as aptly put by ref.40: lower \( b \)-values are indicative of slower convergence, i.e. learning is more difficult (see Supplementary Figure 1b for an illustration). By correlating \( b \), i.e. learning difficulty, with \( h \), i.e. predictability per corpus (partialling out the influence of \( L \)), we arrive at a rather unexpected conclusion (Figure 2d): languages that are harder to predict tend to be easier/faster to learn for PPM. Given statistical theories of language learnability6,21,23, we would have expected the opposite. Understanding this paradox could be the subject of papers to come.

[FIGURE 2]

Figure 2 | PPM as an idealized language learner. (a and b) Compression rates as function of length for the UNPC illustrate that without prior knowledge of the source, PPM acquires a representation of its probabilistic structure in one single pass, i.e. PPM learns to predict. (c) The distribution of relative accuracy ratios indicates that the curve can be accurately modelled by the three parameter ansatz. (d) For each corpus, on both symbolic levels (characters/words), we calculated the partial Spearman correlation coefficient between the learning speed \( b \) and the entropy rate \( h \), after removing the effect of the message length \( L \). Since lower \( b \)-values are indicative of slower convergence (see Supplementary Figure 1b), a positive correlation between \( h \) and \( b \) supports the idea that documents that are harder to predict (i.e. higher \( h \)) are easier to learn to predict (i.e. higher \( b \); compare Supplementary Figure 20).

Figure 2a shows that it is considerably more difficult to predict Chinese characters \( (h = 3.03) \), obviously due to the fact that written Mandarin Chinese employs a logographic system where individual characters typically represent words/morphemes compared to the other five languages
that employ alphabetic systems where symbols typically represent phonemes (here $h \in [0.89, 1.51]$) which affects the capacity of the communication channel$^{13,31}$. However, on the word level (Figure 2b), Chinese ($h = 5.51$) occupies a middle ground ($h \in [4.27, 6.25]$). Correspondingly, variability in $h/r$ tends to be smaller for words than for characters (see Supplementary Figure 21).

**Comparing complexity rankings across corpora.** To evaluate the similarity of language complexity rankings between different corpora, we calculated Spearman correlation coefficients $\rho$ between the values of $h$ for all corpus pairs with a least five common languages ($N = 3,056$) on both symbolic levels (words/characters; see SI: Correlation matrix for details). Figure 3 visualizes our first main empirical finding: complexity rankings are very similar across different corpora. The median correlation across corpora is $\rho_{\text{med}} = 0.742$ for words (first quartile $Q_1 = 0.571$) and $\rho_{\text{med}} = 0.590$ for characters ($Q_1 = 0.346$). Even if we compare complexity across corpora and across symbols, i.e. we correlate the distribution calculated for words in one corpus with the distribution calculated for characters in another corpus, there tends to be a positive statistical association ($\rho_{\text{med}} = 0.367, Q_1 = 0.201$). Figure 3 also shows that the same relationship holds (i) if we only calculate correlations for fully parallel corpora and (ii) if we replicate the analysis for all documents that use the most widely adopted writing system, Latin script (~80% of all our documents), to rule out the possibility that these results are mainly driven by the fact that – as mentioned above – different languages use different writing systems. Similar results are obtained if we use $r$ instead of $h$ to evaluate the similarity of complexity rankings between different corpora ($\rho_{\text{med}} = 0.815$ for words; $\rho_{\text{med}} = 0.820$ for characters and $\rho_{\text{med}} = 0.514$ across symbols; compare Supplementary Figure 22).
A complexity-efficiency trade-off. Comparing Figure 2a and Figure 2b points to another interesting aspect: higher complexity seems to co-occur with shorter message length in the UNPC. Correspondingly, there is a strong negative relationship between $h$ and $L$ for both words and characters. The same is true for the correlation between $r$ and $L$ (all $\rho_s = -0.94$). This prompted us to compare those variables across corpora by calculating $\rho$ between $h$, $r$, and $L$ for all corpus pairs and both symbolic levels. Figure 4a presents a correlogram of the resulting correlation matrix. A visual inspection reveals a striking overall pattern that is quantitatively confirmed by Figure 4d where we use principal-component factoring to analyze the matrix: $\sim 40\%$ of the variance in the matrix is explained by a factor that points towards a previously undocumented complexity-length trade-off: languages that tend to be more complex tend to need fewer symbols to encode messages.

To test whether this pattern merely arose from methodological factors, such as how entropy is estimated or the fact that most, if not all, quantities in the context of word frequency distributions
vary systematically with the text length\textsuperscript{51–53}, we used a different entropy estimation method and adjusted our estimates for the systematic influence of the text length to calculate \( \rho \) between entropy and length for 29 books of the Biblical canon in 146 different languages (see Methods: Correlations between and within languages). Per type (word/character) we calculated (i) Spearman correlations between entropy rates within languages and length and (ii) Spearman correlations between entropy rates between languages and length. Figure 4e shows that the trade-off only holds between (29 books, each with 146 observations), but not within languages (146 languages, each with 29 observations): between languages, i.e. within books, the median Spearman correlation between entropy and length is \( \rho_{med} = -.79 \) for characters and \( \rho_{med} = -.88 \) for words. Within languages, \( \rho_{med} = .03 \) for characters and \( \rho_{med} = -.01 \) for words. Thus, there is only evidence for an entropy-length trade-off between languages, but not within languages.

Together, we arrive at our second main empirical result: from an information theoretic point of view, the message length quantifies efficiency – the shorter the encoded message the higher the efficiency\textsuperscript{6}. Therefore, our results suggest that human languages trade off efficiency against complexity as visualized in Figure 4f: to encode a message with fewer symbols, more information is transmitted per symbol.

Notably, a trade-off would imply that the product of entropy and length is roughly constant. Interestingly, that product is the size of the compressed version of a text and thus an estimate of the Kolmogorov complexity \( K \) that measures the absolute information contained therein\textsuperscript{24,54}. To further investigate the entropy-length trade-off with respect to \( K \), we tested how well the distribution of one variable in one corpus can be predicted using its distribution in another corpus by computing both ordinary least squares (\textit{OLS}) and linear mixed effects (\textit{LMM}) models where – if possible – several random intercepts are included to account for the genealogical and
geographic relatedness of languages\textsuperscript{55}. In addition to $h$, $r$ and $L$, we also included estimates for $H$ for all corpora and both symbolic levels (equation (1)). For $K$, we produced two estimates, (i) the observed compressed length of a document $K_{\text{obs}}$, i.e. an actual (upper bound) approximation of the “algorithmic (descriptive) complexity”\textsuperscript{54} of the text; (ii) the extrapolated compressed length of a document $K_{\text{ext}} = h \cdot L$, an approximation for the algorithmic complexity where, compared to $K_{\text{obs}}$, sender and receiver share \textit{a priori} knowledge regarding the structure and the rules of the language the transmitted text is written in\textsuperscript{54}. In total, we ran 148,544 models (see Methods: Comparing predictability across languages for details regarding statistical modelling and random baseline generation). Figure 5 visualizes the results: compared to all three entropic measures and lengths, both $K_{\text{obs}}$ and $K_{\text{ext}}$ only explain little variance and thus support the idea of an entropy-length trade-off. For example, focussing on the $OLS$ models, the random baseline is 7.57%. Both the median explained variance for $K_{\text{obs}}$ ($R^{2}_{med} = 5.81\%$) and for $K_{\text{ext}}$ ($R^{2}_{med} = 4.40\%$) are below this baseline. For length, the corresponding quantity is $R^{2}_{med} = 20.85\%$. Even higher values are obtained for the three entropic measures ($R^{2}_{med} = 49.17\%$ for $r$, $R^{2}_{med} = 29.99\%$ for $h$ and $R^{2}_{med} = 35.50\%$ for $H$). A similar structure is found for the $LMM$s, with both $R^{2}_{med} = 3.38\%$ for $K_{\text{obs}}$ and $R^{2}_{med} = 2.68\%$ for $K_{\text{ext}}$ being below the random baseline of 3.72\%. Again, (i) length has a much better median model fit with $R^{2}_{med} = 9.60\%$ and (ii) even better fits are obtained for the three entropic measures ($R^{2}_{med} = 34.87\%$ for $r$, $R^{2}_{med} = 17.95\%$ for $h$ and $R^{2}_{med} = 22.36\%$ for $H$).

[FIGURE 5]

\textbf{Figure 5} | Exploring the entropy-length trade-off. For each corpus pair $(c_1, c_2)$ we computed separate $OLS$s and $LMM$s both by and across symbols of $V^{c_1}$ on $V^{c_2}$ where $V$ denotes one of the variables listed on the ordinate. Solid vertical lines represent median value model-fits. Solid horizontal lines encompass 50\% of model-fits. Dotted vertical lines correspond to random baselines (see Methods: Comparing predictability across languages for details).
In SI: Comparing predictability across languages we further show that the small amount of variance explained by $K_{obs}$ and $K_{ext}$ can at least partly be attributed to the influence of different writing systems. We thus can say that knowledge of the distribution of both length and one of the three entropic variables in one corpus allows us to make informed predictions about the corresponding distributions in another corpus while this is hardly the case for both $K_{ext}$ and $K_{obs}$. On the one hand, this strengthens the conclusion that efficiency/length is traded off against (an increase in) complexity/entropy. On the other hand, the observation that distributions for $K$ are far less similar across different corpora could point toward the idea that the absolute amount of information in individual texts is invariant across different languages and thus potentially complementing related observations for written\textsuperscript{56} and spoken language\textsuperscript{5,11}. More research is clearly needed to explore this possibility.

**Limitations/extension.** We proceed by discussing and evaluating several potential limitations/extensions of our approach.

Firstly, we tested if the obtained similarity of complexity rankings between different corpora can mainly be attributed to the degree that different languages make use of inflectional morphology\textsuperscript{26}. We used the Treetagger\textsuperscript{57} with a corresponding language-specific parameter file to lemmatize the European constitution corpus (EUconst, see SI: Corpora) prior to estimation of $h$ on both symbolic levels to remove the effect of inflectional morphology\textsuperscript{26}. For each available corpus, we then calculated pairwise Spearman correlations of the lemmatized entropy rates of EUconst with the (unlemmatized) $h$, $r$ and $L$ of all other corpora. Figures 6a,b indicate that our results are also valid for lemmatized texts on both symbolic levels.
Secondly, due to the large amount of textual data, we used an off-the-shelf compressor that is optimized for speed and memory usage \(^8,40,7\)-zip PPMd\(^58\). While PPM consistently performs well on text compression benchmarks\(^8,48\), its language model is rather simple (see Methods: Estimating entropy). To rule out the possibility that more complex language models would lead to different results, we compressed the Bible (OT) corpus (see SI: Corpora) again, but used a much more sophisticated algorithm called CMIX\(^59\) which achieves state-of-the-art compression rates, but is slower than PPMd by several orders of magnitude (cf. Supplementary Table).

Compared to PPMd, CMIX uses an ensemble of several hundreds of independent prediction models that are combined using a context mixing algorithm that is based on a neural network architecture\(^48,60\). Some of the contexts are allowed to be non-contiguous in order to capture longer-term dependencies\(^48\) and CMIX uses long short-term memory\(^61 (LSTM)\) trained by backpropagation as a byte-level mixer\(^59\). In addition, instead of estimating on the level of either characters or words, we tokenize our text into sub-word units by byte pair encoding (BPE)\(^27,62\) which plays an important role in many state-of-the-art natural language model applications such as GPT-3\(^63\) or SentencePiece\(^64\) and provides strong baseline results on a multilingual corpus\(^65\).

Applying BPE results in a sequence of sub-word units, e.g. “|he |may |give |me |a |kind|ly |re|cep|tion |”. We then compressed each such sequence with CMIX and computed compression rates \(r_{CMIX}\) (see Methods: BPE-CMIX for details). We then calculated pairwise Spearman
correlations between $r_{CMIX}$ of Bible (OT) with $h$, $r$ and $L$ of all other corpora. Figure 6c indicates that a CMIX BPE model also supports our results.

Thirdly, our study is confined to written language. To test a potential relationship with spoken language, we use data from ref.\textsuperscript{5} who show that the information rate measured in bits per second (\textit{bits/sec}) is very similar across different languages and thus pointing towards a trade-off between utterance speed and predictability per syllable\textsuperscript{5,11}. To this end, ref.\textsuperscript{5} compiled a cross-linguistic corpus of 15 short parallel texts that are each being read by 10 different native speakers of 17 different languages belonging to 9 different language families. We fit an \textit{LMM} where the information rate is the outcome and a fixed effect for sex and (crossed) random intercepts for speaker, text, language and language family are included and where different residual variances by sex are estimated to allow for heteroscedasticity in the residual errors as observed by ref.\textsuperscript{5}. We then use this model to obtain the best linear unbiased predictions (\textit{BLUPs}) that are due to language (see Methods: Spoken information rates for details). We then calculated pairwise Spearman correlations between language specific \textit{BLUPs} and $h$, $r$ and $L$ for each of the available corpora on both symbolic levels. Figure 6d reveals an interesting effect that is more pronounced on the level of words: a language that tends to transfer more information per second on the spoken level tends to be less complex/more predictable and also tends to produce longer texts on the written level. More research is clearly needed to understand this relationship. In general, the smaller correlation strengths on the level of characters obtained in most scenarios can very likely, at least to a certain extent, be explained by noise that is due to cross-linguistic differences in the mapping between phonemes and graphemes\textsuperscript{26,66}. Conversely, the stronger correlations on the level of words, where idiosyncrasies/vagaries of the writing system arguably play a less important role, strengthen our overall conclusions.
**Differences across populations.** Finally, we tested an obvious follow-up question: since the structures of speaker populations vary in many ways\(^3,\,6\), it is important to evaluate whether differences in efficiency/complexity may “reflect adaptions to different environments”\(^3\) in which the languages are used. To account for the non-independence of data-points\(^55\), we ran separate LMMs that include random intercepts for language, language family, macro-area, country, writing system and corpus on both symbolic levels where \(h, r, H\) and length are predicted by a fixed effect for the estimated speaker population size (logged) as a proxy for population structure\(^67\) (see Methods: Differences across populations for details). Table 2 presents our third main empirical result: languages with more speakers tend to have higher entropic values, i.e. are more complex. However, we also find that languages with more speakers tend to produce shorter messages, i.e. are more efficient (all parametric \(p\)-values < .00001). Since the sample of languages for which we have available documents cannot be considered a random sample of the population of all languages and given the complex structure of our database (Figure 1b-e), the appropriateness of standard parametric tests can be questioned\(^31\). Therefore, we incorporated two additional tests: (i) an information-theoretic approach where we calculate differences in Akaike’s information criterion (\(AIC\))\(^68\) between full and reduced LMMs where only random effects are included; (ii) a generic variant of the non-parametric *Freedman-Lane* permutation test\(^31,\,69\) that evaluates the hypothesis of whether a predictor provides information about an outcome by permuting residuals from reduced LMMs (equations (17) – (19)). Like the parametric \(p\)-values, the \(\Delta AIC\)-values show that there is very strong support for the full model – i.e. a model that includes a fixed effect for speaker population size – in 8 out of 9 cases and strong support in the remaining case (see Methods: Differences across populations for details).
Table 2 | Differences across populations. Results of LMMs that use speaker population size (logged) as fixed effect and (crossed) random intercepts for language, language family, macro-area, country, writing system and corpus. 1st column: outcome variable (standardized per corpus). 2nd column: symbol type. 3rd column: estimated \( \beta_1 \). 4th column: \( p \)-value based on parametric significance tests. 5th column: marginal \( R^2 \) of the model, i.e. explained variance by the fixed factor (logged) population size (%). 6th column: conditional \( R^2 \) of the model, i.e. explained variance by both the fixed factor and the random intercepts (%). 7th column: number of cases; 8th column: \( AIC \) of the reduced model, i.e. a LMM with no fixed factor but only the random intercepts. 9th column: \( AIC \) of the full model. 10th column: difference in \( AIC \), i.e. \( AIC_r - AIC_f \). 11th column: Result of the permutation test, i.e. count of \( |\hat{\beta}_1/\hat{\sigma}_1| \geq |\beta_1/\sigma_1| \) in relation to the number of repetitions. The \( p \)-value in parentheses is calculated as that count divided by the number of repetitions, see Methods: Differences across populations. NB: Only if the \( p \)-value after 100 repetitions was below .1, the test was re-run with 1,000 repetitions (see Methods: Differences across populations for details).

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Symbol</th>
<th>( \hat{\beta}_1 )</th>
<th>( p )-parametric</th>
<th>( R^2_{LMM(m)} )</th>
<th>( R^2_{LMM(c)} )</th>
<th>( N )</th>
<th>( AIC_r )</th>
<th>( AIC_f )</th>
<th>( \Delta AIC )</th>
<th>Permutation test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy rate</td>
<td>Characters</td>
<td>0.045</td>
<td>&lt;.0000001</td>
<td>0.71</td>
<td>88.50</td>
<td>3,849</td>
<td>7816.7</td>
<td>7793.4</td>
<td>23.3</td>
<td>0 of 1,000 (0.000)</td>
</tr>
<tr>
<td>Entropy rate</td>
<td>Words</td>
<td>0.056</td>
<td>&lt;.0000001</td>
<td>1.39</td>
<td>88.16</td>
<td>3,849</td>
<td>7757.4</td>
<td>7735.6</td>
<td>21.8</td>
<td>10 of 1,000 (0.010)</td>
</tr>
<tr>
<td>Compression rate</td>
<td>Characters</td>
<td>0.066</td>
<td>&lt;.0000001</td>
<td>1.50</td>
<td>93.99</td>
<td>3,849</td>
<td>5699.0</td>
<td>5586.5</td>
<td>112.5</td>
<td>0 of 1,000 (0.000)</td>
</tr>
<tr>
<td>Compression rate</td>
<td>Words</td>
<td>0.067</td>
<td>&lt;.0000001</td>
<td>1.89</td>
<td>93.15</td>
<td>3,849</td>
<td>6438.9</td>
<td>6400.6</td>
<td>38.3</td>
<td>9 of 1,000 (0.009)</td>
</tr>
<tr>
<td>Unigram entropy</td>
<td>Characters</td>
<td>0.030</td>
<td>&lt;.000001</td>
<td>0.42</td>
<td>94.22</td>
<td>3,849</td>
<td>4888.8</td>
<td>4878.7</td>
<td>10.1</td>
<td>12 of 100 (0.120)</td>
</tr>
<tr>
<td>Unigram entropy</td>
<td>Words</td>
<td>0.110</td>
<td>&lt;.0000001</td>
<td>38.78</td>
<td>90.06</td>
<td>1,914</td>
<td>5342.3</td>
<td>5351.1</td>
<td>81.3</td>
<td>0 of 1,000 (0.000)</td>
</tr>
<tr>
<td>Unigram entropy</td>
<td>(Crúbadán)*</td>
<td>0.066</td>
<td>&lt;.0000001</td>
<td>2.45</td>
<td>85.12</td>
<td>3,849</td>
<td>7676.8</td>
<td>7640.3</td>
<td>36.6</td>
<td>1 of 1,000 (0.001)</td>
</tr>
<tr>
<td>Message length</td>
<td>Characters</td>
<td>-0.050</td>
<td>&lt;.00000005</td>
<td>1.01</td>
<td>83.86</td>
<td>3,849</td>
<td>8784.7</td>
<td>8767.5</td>
<td>17.2</td>
<td>0 of 1,000 (0.000)</td>
</tr>
<tr>
<td>Message length</td>
<td>Words</td>
<td>-0.061</td>
<td>&lt;.0000001</td>
<td>1.38</td>
<td>86.48</td>
<td>3,849</td>
<td>8732.8</td>
<td>8711.1</td>
<td>21.7</td>
<td>4 of 1,000 (0.004)</td>
</tr>
</tbody>
</table>

*Outcomes for the Crúbadán data were not standardized and the model does not contain a random effect for corpus but additional fixed effects for text length, available number of documents (both logged) and a binary variable indicating if the word frequency list is truncated (no/yes; see Methods: Differences across populations and SI: S.3.7 for details).

The permutation test also dismisses the idea that the influence of the speaker population size can easily be explained as accidental\(^70\): in all but one case, the permutation \( p \)-value is below .011 with a median value of .001. In the remaining model, 12 of 100 permutations displayed at least the observed absolute \( z \)-statistic. On both symbolic levels, the results support the idea that languages with more speakers tend to have higher entropies: for both \( h \), \( r \) and \( H \) the estimated coefficients \([\beta_1]\) are positive. On the other hand, the direction of the relationship is negative for \( L \) and thus indicates that languages that are spoken by more people tend to produce shorter messages.
To avoid possible confounding influences, we ran additional models where we restricted the analyses to fully parallel corpora. The results are qualitatively very similar compared to the unrestricted analyses (see Supplementary Table 6). As explained in Methods: Differences across populations, outcomes were standardized per corpus. Supplementary Table 7 demonstrates that if we log outcomes instead, the analyses still support all overall conclusions.

We then tested whether several other environmental variables (within-country linguistic diversity, geographical range, altitude, climate, distance to water resources, latitude; see SI: Code and data for details) also predict entropy/efficiency: Supplementary Table 8 demonstrates that the results do not support the idea of a relationship for those environmental variables, with the results being strongest for geographical range sizes. Here the direction of the estimated $\beta_1$-coefficients points towards the idea that languages with a greater geographical range are less predictable but produce shorter messages. However, only the parametric significance tests support this conclusion, while it is unsupported by both the permutation test and especially the difference-in-$AIC$ approach. For the remaining variables, there is essentially no support for any relationship.

Next, we tested whether speaker population size is associated with $K$. Supplementary Table 9 demonstrates that this is not the case for neither $K_{ext}$ nor $K_{obs}$. Furthermore, the variance explained by both the fixed and the random factors $[R^2_{LMM(c)}]$ is markedly lower for models compared to the models of Table 2 and Supplementary Table 6. Conversely, this result supports the idea of a trade-off between entropy and length. Finally, Supplementary Table 10 shows that the overall conclusions for speaker population size still hold if we include covariates in our crossed-effects models, i.e. if we allow the effect of population size to vary across different groups, which represent deviations from the overall mean linear effect of speaker population size (see SI:
equation (S2)). For a discussion/test of model assumptions and an evaluation of the reliability of obtained results, see Supplementary Information: Differences across populations.

**Discussion**

A central goal of linguistics is to understand the diverse ways in which human language can be organized. In this paper, we present results of a large cross-linguistic analysis of written language. Our computational results support the idea that the long-held belief of a principle of “invariance of language complexity” is likely incorrect⁴: a language with high/low entropy in one corpus also tends to be more/less complex in another corpus (Figure 3 and Supplementary Figure 22). This observation, which constitutes our first main empirical finding, prompted us to visually explore the compression rates for different languages in different corpora (e.g. Figure 2a,b). On this basis, we arrived at our second main empirical finding (Figure 4) that can be seen as a specific instantiation of the compensation hypothesis¹¹,⁷¹: since high entropy languages tend to need fewer symbols to encode messages, this implies that higher complexity is compensated by higher efficiency, because, under an idealized information-theoretic view where human languages can be understood as *codes* that allow the transmission of messages⁶,³⁷,⁷², message length quantifies efficiency⁶.

Finally, we showed that languages with more speakers tend to produce shorter messages, i.e. are more efficient. We also found that languages with more speakers have higher entropic values, i.e. are more complex (Table 2). This result stands in contrast to previous studies that argued that languages spoken in larger communities tend to be less complex³,⁶⁷,⁷³. In SI: Individual regressions, we discuss this apparent paradox by showing that smaller corpora tend to be biased towards languages with more speakers. This has important ramifications for any quantitative typological investigation since it demonstrates that analyses of smaller corpora can lead to
potentially incorrect conclusions as already discussed by ref.\textsuperscript{74} and thus highlights the importance of endeavours such as the Crúbadán Project\textsuperscript{16} given that research shows that – comparable to global biodiversity – the world’s language diversity is currently under great threat\textsuperscript{75,76}.

Taken together, the statistical relationship between entropy/length and speaker population size constitutes our third main empirical result: the complexity-efficiency trade-off is (partly) shaped by the social environment in which languages are used. This points towards the idea that language complexity is an evolving variable\textsuperscript{2}, as our results imply that when a speaker community gets larger across time, we should observe a correlated trend of increasing entropy rates and decreasing message lengths. This could be an interesting avenue for future research.

**Methods**

**Corpora.** All corpora used in this paper and details regarding data preparation and compression can be found in SI: Corpora.

**Estimating entropy.** A text $t$ can be represented as a sequence of $L$ symbols that are drawn (with replacement) from an alphabet consisting of $C$ different symbol types. Depending on the chosen level of analysis, symbols are taken to be either (Unicode) characters or word types. We can then count how often each symbol $i$ appears in $t$ and call the resulting token frequency $f_i$, and can then represent $t$ as a distribution of symbol frequencies. In order to quantify the amount of information contained in $t$, we can calculate the Gibbs-Shannon entropy $H$ of this distribution as\textsuperscript{12}:

$$H(t) = -\sum_{i=1}^{C} p_i \cdot \log_2 (p_i)$$  \hspace{1cm} (1)

where $p_i = \frac{f_i}{L}$ is the maximum likelihood estimator of the probability of $i$ in $t$ consisting of $L = \sum_{i=1}^{C} f_i$ tokens, where $C$ is the number of symbol types in $t$. $H(t)$ can be interpreted as the average number of guesses that are needed to correctly predict the type of a symbol token that is randomly sampled from $t$. The entropy rate or per-symbol entropy\textsuperscript{12,77} can be formally defined as:

$$h_t = \lim_{l \to \infty} \frac{1}{l} H_t(t) = \lim_{l \to \infty} \frac{1}{l} H(s_1, s_2, \ldots, s_l)$$
where $s_1, s_2, \ldots, s_l$ is a consecutive sequence of symbols of length $l$ and $H_l(t)$ denotes the so-called block entropy of block size $l$.\textsuperscript{41} Given stationarity\textsuperscript{12,77}, this is equivalent to

$$h_t = \lim_{l \to \infty} H(s_l | s_{1}, s_2, \ldots, s_{l-1})$$

(2)

Thus, $h_t$ is the conditional entropy of $s_t$ given all preceding tokens\textsuperscript{12,77}. In analogy to $H(t)$, $h$ can be informally defined as the average number of (yes/no) guesses that are needed to guess the next symbol of a sequence and thus incorporating the notion that prediction and understanding are intimately related\textsuperscript{8}.

To estimate $h_t$ computationally, we use a data compression algorithm, since the true probability distribution for natural language is unknown\textsuperscript{8,41}. The algorithm generates a language model, i.e. an estimate of the probability distribution of $t$ that can then be used for encoding, e.g. via arithmetic coding.\textsuperscript{8,40,78} We use PPM as implemented in the 7-zip software package, which is based on Dmitry Shkarin’s PPMd.\textsuperscript{58} The algorithm makes an assumption of the Markov property: To encode/predict the next symbol, the algorithm uses the last $o$ symbols that immediately precede the symbol of interest. If the order $o$ context has not been seen before, the algorithm attempts to make a prediction based on the last $o-1$ symbols. This is repeated until a match is found, or, if no match is found until order 0, then a fixed prediction is made. In general, let $R_t$ denote the size (in bits) of the compressed text. Then the compression rate $r_t = R_t / L_t$ is an upper bound on the underlying entropy rate $h_t$, i.e.:

$$r_t \geq h_t$$

(3)

Importantly, $h_t$ is defined in the limit, i.e. for a text $t$ whose length $L_t$ tends to infinity.\textsuperscript{12,77} Given stationarity and ergodicity\textsuperscript{12}, the following equality holds for universal compressors\textsuperscript{40}:

$$\lim_{L_t \to \infty} r_t = h_t$$

(4)

Or put differently, the entropy rate measures how difficult it is to predict subsequent text based on the preceding input when the optimal compression scheme is known\textsuperscript{40}. Equation (4) implies that convergence to the source entropy is only guaranteed in the limit\textsuperscript{79}, i.e. when the text size approaches infinity. One way to take into account the dependence on $L$ is to use extrapolation when estimating $h_t$ via compression\textsuperscript{40}. However, the (probabilistic) relationship between (the convergence of) $h_t$ and $L$ is unknown. For brevity, we drop the subscript $t$ in what follows. To estimate $h$, we use a variant of the following ansatz suggested by ref.\textsuperscript{41}:

$$r_t = h + A \cdot \log \frac{l}{lb}$$

(5)
where \( A > 0, b > 0 \) and – assuming that the entropy rate is positive \(- h > 0\); \( r_i = R(X_i^l) / l \) denotes the number of bps that are needed to compress the first \( l \) symbols of \( t \). In general, the idea of the ansatz is to calculate the compression rate for different subsequences of \( t \) of increasing length. This gives us a measure of how well language learning succeeds\(^{23,80}\). For example, we can feed the compressor with the first \( l = 1 \cdot m \) symbols and calculate the compression rate for this subsequence where \( m \) is some pre-defined chunk size, e.g. 1,000 symbols. After that, the compression rate is calculated for the first \( l = 2 \cdot m \) symbols and the compression rate is calculated again. This procedure is repeated until the end of \( t \) is reached. The resulting series of compression rates for texts that consist of \( 1, 2, \ldots, [L/m] \) chunks can then be used to fit the three parameters to the data. We fit the following nonlinear ansatz function by log-least squares:

\[
\begin{align*}
    r_l &= \exp \left( h^* + \exp \left( A' \right) \cdot \frac{\log l}{\log (l)} \right) + \delta_l \tag{6}
\end{align*}
\]

where \( \delta_l \) is an independent and identically distributed (i.i.d) error term and \( \exp() \) denotes the exponential function (see SI: Ansatz functions where we discuss other ansatz functions and different error specifications that have been suggested in the literature\(^{40,41,81}\) and justify our choice). Since we want \( A \) and \( b \) to be positive, we set interval constraints that make sure that the optimization algorithm will not search in the negative subspace by fitting both parameters as exponentials, i.e. we estimate \( A' = \log (A) \) and \( b' = \log (b) \). The limiting entropy rate of equation (5) can be recovered from equation (6) as \( h = \exp (h^*) \).

Since achieving convergence of the parameter estimates turned out to be difficult, we approximate initial values in linear space, i.e., for each value of \( \phi = .01, .02, \ldots, 10 \), we calculate \( \Phi = \frac{\log l}{\log (\phi)} \) and fit the following linear regression by OLS:

\[
\begin{align*}
    \log(r_l) &= \beta_h + \beta_A \Phi + \delta_l \tag{7}
\end{align*}
\]

where \( \delta_l \) is an i.i.d error term. To provide initial values to fit equation (6), we pick the solution of equation (7) where the root mean squared error is smallest and where \( \beta_A > 0 \), then \( h^* \) is initialized as \( \beta_h \), \( A' \) is initialized as \( \exp (\beta_A) \) and \( b' \) is initialized as \( \exp (\phi_m) \) where \( \phi_m \) denotes the value of \( \phi \) corresponding to the selected \( \Phi \).

As written above, the model is fit by log-least squares, i.e. \((\log(\hat{r}_l) - \log(r_l))^2 = \log(\hat{r}_l/r_l)^2\) where \( r_l \) and \( \hat{r}_l \) denote the observed and the predicted compression rate, respectively. To assess the model fit, we fit both equation (6) and equation (7) to only the first 90% of the data points and use the last 10% as test data. Let \( \tau = 1, 2, \ldots, T \) denote the holdout data points. On this basis, we calculate the model fit as a measure of prediction accuracy\(^{82}\):
\[ M = \frac{1}{T} \sqrt{\sum_{T=1}^{T} \log \left( \frac{\hat{r}_T}{r_T} \right)^2} \]  

\( M \)-values are reported as percentages by multiplying the above equation by 100. Note that as long as the difference between \( r_T \) and \( \hat{r}_T \) is relatively small, \( \log \left( \frac{\hat{r}_T}{r_T} \right) \approx \frac{\hat{r}_T - r_T}{r_T} \). Thus we can interpret \( M \) as measuring the approximate (absolute) average percentage difference between \( r_T \) and \( \hat{r}_T \).

In order to avoid relying too much on the *ansatz* whose appropriateness can only be verified numerically, we additionally use the compression rate (denoted as \( r \) in what follows) at \( T = \lfloor L/m \rfloor \) as an observed unbiased upper-bound-estimate for the underlying entropy rate.

**Correlations between and within languages.** For this analysis, each Bible translation\(^{33}\) was split into 66 separate 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 146 different languages.

To rule out the possibility that our results are mere artefacts resulting from the fact that most, if not all, quantities in the context of word frequency distributions vary systematically with the text length\(^{51–53}\), we first computed the minimum text length in symbols (words/characters) per language across the 29 books and call that minimum \( \lambda_i \). Likewise, we computed the minimum text length in symbols per book across the 146 languages and call that minimum \( \lambda_c \). We then truncated each book at the respective minima and used the truncated books to calculate the corresponding entropy rates \( h_c \) and \( h_i \). Additionally, here we use the non-parametric entropy rate estimator by ref.\(^{83}\) that was already used in several studies\(^{31,56,74,77,84}\). This serves to check whether our results systematically depend on the PPM algorithm. Based on \( \lambda_c \), the entropy rate \( h_c \) is calculated as (likewise for \( h_i \)):

\[ h = \left[ \frac{1}{\lambda_c} \sum_{l=2}^{A_l} \frac{A_l}{\log_2(l)} \right]^{-1} \]  

Here, the key quantity of interest is the match-length \( A_l \). 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 before this position and can be used to estimate \( h \), since it was shown that \( A_l \) grows like \( \log_2(l) / h \)\(^{79,83,85}\). More details and an open source Java program to efficiently obtain match-lengths in texts can be found in ref.\(^{56}\).

Per symbol type (word/character) we calculated Spearman correlations between \( h_i / h_c \) and the corresponding initial text length. Note that even without truncating, i.e. without adjusting for the systematic influence of the text length, the results are much clearer *between* than *within* languages:
between languages $\rho_{med} = -.83$ for characters and $\rho_{med} = -.93$ for words. Within languages, $\rho_{med} = -.34$ for characters and $\rho_{med} = -.17$ for words.

**Comparing predictability across languages.** For the OLS case, we made use of the fact that for simple OLS regression the square of the Pearson correlation coefficient $\rho_p$ is equivalent to $R^2$, i.e. the coefficient of determination. For each corpus pair $(c_1, c_2)$ we computed pairwise $\rho_p$-coefficients both by and across symbol types between $V^{c_1}$ and $V^{c_2}$ where $V$ denotes one of the following variables $r, h, H, L, K_{obs}$ or $K_{ext}$. Both the outcome and the predictor were logged. We restricted the computation to corpus pairs with at least 5 available shared languages (for languages with more than one available translation in a corpus, all quantities were averaged). To generate random baselines, we randomly permuted the values of $V^{c_2}$ and recalculated $\rho_p$. As a random baseline, we computed the upper quartile for the set of all squared random $\rho_p$-coefficients. This means that 75% of all models show a coefficient of determination that is lower than or equal to this random baseline.

For the LMM case, we regressed $V^{c_1}$ on a fixed effect for $V^{c_2}$. Again, both outcome and predictor were logged and computations were restricted to corpus pairs with at least 5 available shared languages. We fitted the following crossed-effects models:

$$V_{imag}^{c_1} = \beta_0 + \beta_1 V_{imag}^{c_2} + \mu_m + \alpha_a + \zeta_g + \epsilon_{imag}$$  \hspace{1cm} (10)

for $i = 1, \ldots, I$ different languages (identified by their ISO codes), $m = 1, \ldots, M$ macro-areas (Africa, Australia, Eurasia, North America, Papuasia or South America), $a = 1, \ldots, A$ countries and $g = 1, \ldots, G$ language families with $\mu_m \sim \text{Gaussian}(0, \sigma_m^2); \alpha_a \sim \text{Gaussian}(0, \sigma_a^2); \zeta_g \sim \text{Gaussian}(0, \sigma_g^2); \epsilon_{imag} \sim \text{Gaussian}(0, \sigma_{\epsilon}^2)$ all independently and where $\sigma_m^2, \sigma_a^2, \sigma_g^2$ and $\sigma_{\epsilon}^2$ are the variances of $\mu_m, \alpha_a, \zeta_g$ and $\epsilon_{imag}$. The fixed portion of the model, $\beta_0 + \beta_1 V_{imag}^{c_2}$ is analogous to the linear predictor from a standard OLS regression and the random portion of the model, i.e. $\mu_m + \alpha_a + \zeta_g + \epsilon_{imag}$, incorporates group-specific shifts for language family, country and macro-area to account for genealogical and geographic relatedness of languages, i.e. $\sigma_m^2, \sigma_a^2$ and $\sigma_g^2$ (languages were excluded from the analyses if information for one or more of the grouping factors was missing). All LMMs were fitted by restricted maximum likelihood (REML). Note that for some corpus pairs not all groups did not vary, for example because all languages are located in one macro-area (e.g. in case of the European Constitution data, all languages are located in the Eurasian macro-area). In a similar vein, fitting an LMM does not make much sense if each group of each random factor consists of exactly one member. To solve this problem, our model automatically checks the composition of each grouping factor for each corpus pair and only included it if it consisted of at least two different groups and if at least one of those groups consisted of
more than one member. Models were fitted with gradient-based maximization first. If gradient-based maximization did not converge, models were re-fitted with expectation-maximization (EM) only and we accepted any solution after a maximal number of EM iterations of 1,000.

As shown in ref.\(^\text{87}\), equation (27), the variance of the fixed component of the model can be estimated as:

\[
\sigma_f^2 = \text{var}(\beta^f V^c_{\text{imag}})
\]

This can be computed by predicting values based on the estimated fixed effects of the model followed by a calculation of the variance of these fitted values. The variance of the full model can then be decomposed as:

\[
\sigma_f^2 + \sigma_m^2 + \sigma_a^2 + \sigma_g^2 + \sigma_e^2
\]

(12)

On this basis, ref.\(^\text{87}\), equation (26), define an \(R^2\) as measure of explained variance of the fixed portion (\(m\) indicates marginal \(R^2\)) of the LMM as follows:

\[
R^2_{\text{LMM}(m)} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_m^2 + \sigma_a^2 + \sigma_g^2 + \sigma_e^2}
\]

(13)

We calculated \(R^2_{\text{LMM}(m)}\) for each model and proceeded as for the OLS-version described above.

In a similar vein, it is possible to define a conditional \(R^2\) version, i.e. the variance explained by both the fixed and the random factor, as in ref.\(^\text{87}\), equation (30):

\[
R^2_{\text{LMM}(c)} = \frac{\sigma_f^2 + \sigma_m^2 + \sigma_a^2 + \sigma_g^2}{\sigma_f^2 + \sigma_m^2 + \sigma_a^2 + \sigma_g^2 + \sigma_e^2}
\]

(14)

**BPE-CMIX.** We downloaded the most current version (v19) of CMIX from ref.\(^\text{59}\). For this analysis, each Bible translation\(^\text{33}\) was split into 66 separate books of the Biblical canon. We only kept translations with available information for all 39 books of the Old Testament (OT) of the Christian biblical canon. For languages with more than one available OT translation, we randomly sampled one translation. In total, we have available translations for 147 different languages. Before compression, byte pair encoding\(^\text{62}\) was applied to each translation. Following ref.\(^\text{27}\), the number of BPE merges was set to 0.4 \(\cdot C\) where \(C\) is the number of different word types observed in a given translation. After tokenization into sub-word units, we replaced each distinct sub-word unit by a unique symbol and CMIX is then used (without further preprocessing and without an additional dictionary) to compress both the resulting symbol sequence and the mapping of sub-word units to 1-4 byte symbols in order to the calculate compression ratio of text \(t\) as

\[
r^t_{\text{CMIX}} = \frac{(R^t_{\text{seq}} + R^t_{\text{dict}})}{L^t_{\text{BPE}}}
\]

(15)
where $R_{seq}^t$ refers to the compressed length of the $BP$-encoded symbol sequence, $R_{dict}^t$ refers to the compressed length of the mapping of sub-word units to byte symbols and $L_{BPE}^t$ denotes the length of $t$ in sub-word unit tokens.

**Differences across populations.** We fit the following crossed-effects models on both symbolic levels (characters/words):

$$y_{tkmagwi} = \beta_0 + \beta_1 x_{tkmagwi} + \kappa_k + \mu_m + \alpha_a + \zeta_g + \omega_w + \iota_i + \epsilon_{tkmagwi}$$  

(16)

for $t = 1, \ldots, T$ different texts, $k = 1, \ldots, K$ corpora, $m = 1, \ldots, M$ macro-areas, $a = 1, \ldots, A$ countries, $g = 1, \ldots, G$ language families, $w = 1, \ldots, W$ writing systems and $i = 1, \ldots, I$ languages, all i.i.d. and independently. Thus, each model contains group-specific shifts for language, language family, country, macro-area, writing system and corpus to account for the relatedness of documents and languages$^{55}$; $x$ denotes the (logged) speaker population size; $y$ denotes one of the following outcome variables: $r$, $h$, $H$ or $L$. Additionally, we fit models for $H_{Crubadan}$ as an outcome that is calculated based on Crúbadán word frequency information$^{36}$. Here, models do not contain a random effect for corpus, but models additionally contain fixed effects for text length, available number of documents (both logged) and a binary variable indicating whether the word frequency list is truncated (no/yes; see SI: 3.7 for details). All models are fitted by REML. Since LMM convergence behaviour for this complex model specification turned out to be problematic, outcomes were standardized per corpus, i.e. the corpus-specific mean was subtracted from each observed value and the result was divided by the corpus-specific standard deviation. Correspondingly, outcomes for the Crúbadán models were not standardized. We then extract the estimate for $\beta_1$ and its parametric $p$-value that is based on the absolute value of the $z$-statistic, i.e. $|\hat{\beta}_1|/\hat{\sigma}_{\hat{\beta}_1}$ where $\hat{\sigma}_{\hat{\beta}_1}$ is the standard error of $\hat{\beta}_1$. As measures of effect size, we calculate both $R^2_{LMM(m)}$, i.e. explained variance by the fixed factor and $R^2_{LMM(c)}$, i.e. explained variance by both the fixed factor and the random intercepts (see equation (13,14) and ref.$^{37}$ for details).

To accommodate for the fact that the appropriateness of standard parametric significance testing can be called into question because our data cannot be considered a random sample of the population of all languages, we present two further analyses:

(i) We calculate $AIC$, i.e. Akaike’s information criterion$^{68}$ for each full model and call the resulting quantity $AIC_i$. To test whether the inclusion of the fixed effect is warranted, we fit a reduced model, i.e. we refit the model without the fixed effect and call the resulting quantity $AIC_i$. Since smaller $AIC$ values are indicative of a better-fitting model, we then calculate $\Delta AIC$, i.e. the difference between $AIC_i$ and $AIC_i$. Only if this value is positive, the inclusion of the fixed effect seems appropriate and the greater the value of $\Delta AIC$, the greater the support for the full model. As threshold values, we use the guideline presented in
ref.\textsuperscript{88} that is based on ref.\textsuperscript{89} and translated into $\Delta$AIC values by ref.\textsuperscript{90}: there is limited support for the full model if $\Delta$AIC ranges between 0 and 4.6; moderate support if $\Delta$AIC ranges between 4.6+ and 9.2; strong support if $\Delta$AIC ranges between 9.2+ and 13.8 and very strong support for values above 13.8.

(ii) We implemented a generic variation of the Freedman-Lane permutation procedure that does not make any assumptions about the mechanism that generated the data\textsuperscript{69,70,91}. Here, we wish to test the null hypothesis that $x$ provides no information about the outcome $y$, i.e. that $\beta_1 = 0$. In that case, equation (16) would reduce to:

$$y_{tkmagwi} = \beta_0 + \kappa_k + \mu_m + \alpha_a + \zeta_g + \omega_w + \iota_i + \epsilon_{tkmagwi}$$

(17)

Since the residuals $\epsilon_{tkmagwi}$ are supposed to be i.i.d. with zero expectation, they are exchangeable under the null hypothesis\textsuperscript{92}. Or put differently, if the null hypothesis is true, we do not lose “anything essential in the data”\textsuperscript{69} by permuting $\epsilon_{tkmagwi}$, because the residuals from the reduced model (equation (17)) should not be different from the full model (equation (16)) and can thus be used to generate the reference distribution of the test statistic. Thus, we repeatedly fit

$$y^*_{tkmagwi} = \beta_0 + \hat{\beta}_1 x_{tkmagwi} + \kappa_k + \hat{\mu}_m + \hat{\alpha}_a + \hat{\zeta}_g + \hat{\omega}_w + \hat{\iota}_i + \epsilon^*_{tkmagwi}$$

(18)

where $y^*_{tkmagwi}$ is computed based on equation (17) as the sum of the fitted values, i.e. $\bar{y}_{tkmagwi}$, and the randomly permuted residuals, i.e. $\epsilon^*_{tkmagwi}$:

$$y^*_{tkmagwi} = \bar{y}_{tkmagwi} + \epsilon^*_{tkmagwi} = \hat{\beta}_0 + \hat{\kappa}_k + \hat{\mu}_m + \hat{\alpha}_a + \hat{\zeta}_g + \hat{\omega}_w + \hat{\iota}_i + \epsilon^*_{tkmagwi}$$

(19)

We then count the number of times the absolute value of the $z$-statistic of the estimated parameter for $x_{tkmagwi}$, i.e. $\beta_1^*/\hat{\beta}_1$, where $\hat{\beta}_1$ is the standard error of $\beta_1^*$, was found to be at least as high as $\beta_1^*/\hat{\beta}_1$, i.e. the observed $z$-statistic based on equation (16), call the resulting quantity $\varphi$ and divide $\varphi$ by the number of repetitions. The result is the permutation $p$-value. Since, due to the complex model structure, gradient-based maximization turned out to be far too time consuming, models were fitted with expectation-maximization (EM) only and we accepted any solution after a maximal number of EM iterations of 20. In addition, we first ran the permutation procedure with 100 repetitions. If $\varphi$ was below 10, i.e. the corresponding $p$-value was below .1, we re-ran the permutation procedure with 1,000 repetitions.

**Spoken information rates.** Using data from ref.\textsuperscript{5}, we fit the following crossed-effects model by REML:

$$z_{stgi} = \beta_0 + \beta_1 \text{male}_{stgi} + \tau_s + \tau_t + \zeta_g + \iota_i + \epsilon^f_{stgi} (1 - \text{male}_{stgi}) + \epsilon^m_{stgi} \text{male}_{stgi}$$

(20)

for $s = 1, \ldots, 170$ speakers, $t = 1, \ldots, 15$ different texts, $g = 1, \ldots, 9$ language families and $i = 1, \ldots, 17$ languages, all i.i.d. and independently where $z$ denotes the information rate measured in bits/sec and
male is a binary indicator variable that is equal to 0/1 if speaker s is female/male. Since ref. 5 observed heteroskedastic residual errors, we allow for heteroscedasticity with respect to the fixed effect of sex by estimating one distinct variance for each sex while maintaining independence of residual errors, i.e. $\varepsilon^f_{stgl}$ denotes the residual error for females and $\varepsilon^m_{stgl}$ denotes the residual error for males, resulting in a better fit to the data [$AIC = 10238.21$, compared to a model without distinct variances where $AIC = 10245.95$]. This model is used to calculate BLUPs of $\mu_i$ which are then used as estimates of the effect of language on information rates.

References


70. Freedman, D. A. & Lane, D. Significance testing in a nonstochastic setting. in *A Festschrift for Erich L. Lehmann* 185–208 (Wadsworth, 1983).


74. Koplenig, A. Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *R. Soc. Open Sci.* **6**, 181274 (2019).


Acknowledgments: We acknowledge the many people who compiled publicly available parallel text data and provided them to the community, especially Jörg Tiedemann for the OPUS corpus collection, Kevin Scannell for the Crúbadán project, Thomas Mayer and Michael Cysouw for the Parallel Bible Corpus, Natalia Levshina for the ParTy Corpus, and Dirk Goldhahn and the Leipzig Corpora Collection team. We thank Sarah Signer for proofreading. We also thank Chris Bentz, Łukasz Dębowski and Kumiko Tanaka-Ishii for input and feedback.

Funding: We received no external funding.

Author contributions:

Conceptualization: AK, PM
Data curation: AK
Formal Analysis: AK
Investigation: AK
Methodology: AK
Software: AK
Validation: AK, SW
Visualization: AK
Writing – original draft: AK
Writing – review & editing: AK, PM, SW

Competing interests: We declare no competing interests.

Data and materials availability: All parallel text data were taken from the sources mentioned in the supplementary information. Code and data are described in SI: Code and Data and are available at https://osf.io/f5mke/.
Figures

Figure 1

Dataset and study design. (a) Collected corpora and their geographical distribution. Asterisks indicate fully parallel corpora (see Supplementary Information: Corpora for details). (b) Number of documents per corpus. Corpora are ranked by the number of available documents. (c) Number of documents per language. Languages are ranked by the number of available data points. (d) Kernel density estimations of the number of speakers per language (logged). Emerald line – all existing languages. Cranberry line – languages for which we have available documents. Dotted vertical lines correspond to median values. (e) Distribution of document length in characters and words. (f) Adaption of Shannon’s information-theoretic view of communication for the analysis of parallel corpora.

Figure 2

PPM as an idealized language learner. (a and b) Compression rates as function of length for the UNPC illustrate that without prior knowledge of the source, PPM acquires a representation of its probabilistic structure in one single pass, i.e. **PPM learns to predict**. (c) The distribution of relative accuracy ratios indicates that the curve can be accurately modelled by the three parameter ansatz. (d) For each corpus, on both symbolic levels (characters/words), we calculated the partial Spearman correlation coefficient between the learning speed $b$ and the entropy rate $h$, after removing the effect of the message length $L$. Since lower $b$-values are indicative of slower convergence (see Supplementary Figure 1b), a positive correlation between $h$ and $b$ supports the idea that documents that are harder to predict (i.e. higher $h$) are easier to learn to predict (i.e. higher $b$; compare Supplementary Figure 20).

Figure 3

Testing the similarity of complexity rankings. Distribution of pairwise Spearman correlations for $h$ between all corpus pairs and on/across both symbolic levels (words/characters) for (1) all corpora, (2) fully parallel corpora and (3) documents that use Latin script.

Figure 4

Comparing efficiency and complexity across corpora. (a) Correlogram where each square represents a pairwise correlation for each combination of $r, h, H$ and $L$ across corpora and symbols as visualized in
insets (b,c). Legend to the colors is shown on the right and on top. (d) A principal-component factoring reveals that ~50% of the variance in the matrix can be reduced to two factors, one main factor that represents the strong negative correlation between length and entropy and one factor that separates symbol types. (e) Distribution of $\rho$s for 29 books of the Biblical canon in 146 different languages between and within languages. (f) Languages trade off entropy against message length – to encode a source message with fewer symbols, more information has to be transmitted per symbol.

**Figure 5**

**Exploring the entropy-length trade-off.** For each corpus pair ($c_1$, $c_2$) we computed separate $OLS$s and $LMM$s both by $V^{c_1}$ and $V^{c_2}$ across symbols of on where denotes one of the variables listed on the ordinate. Solid vertical lines represent median value model-fits. Solid horizontal lines encompass 50% of model-fits. Dotted vertical lines correspond to random baselines (see Methods: Comparing predictability across languages for details).

**Figure 6**

**Evaluating limitations and extensions.** (a and b) Distribution of pairwise $\rho$s between $h_{\text{lemmatized}}$ estimated on the character level (a) and on the word level (b) with both $h$, $r$ and $L$ on both symbolic levels. (c) Distribution of pairwise $\rho$s between $r_{\text{CMIX}}$ estimated on the level of sub-word units with both $h$, $r$ and $L$ on both symbolic levels (see Methods: BPE-CMIX for details) (d) Distribution of pairwise $\rho$s between estimated $BLUPs$ of language on information rate with both $h$, $r$ and $L$ on both symbolic levels (see Methods: Spoken information rates for details). *NB:* In each case, computations of $\rho$s were restricted to pairs with at least 5 available shared languages.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformationv1.00.docx](#)