Human languages trade off complexity against efficiency

Authors: Alexander Koplenig[1]*, Sascha Wolfer[1], Jan Oliver Rüdiger[1], Peter Meyer[1]

[1]Department of Lexical Studies, Leibniz Institute for the German Language (IDS), Mannheim, Germany.

*Corresponding author. E-mail: koplenig@ids-mannheim.de

1

9 **Abstract**

10 From a cross-linguistic perspective, language models are interesting because they can be used

11 as idealised language learners that learn to produce and process language by being trained on

12 a corpus of linguistic input. In this paper, we train different language models, from simple

13 statistical models to advanced neural networks, on a database of 41 multilingual text

14 collections comprising a wide variety of text types, which together include nearly 3 billion

15 words across more than 6,500 documents in over 2,000 languages. We use the trained models

16 to estimate entropy rates, a complexity measure derived from information theory. To compare

17 entropy rates across both models and languages, we develop a quantitative approach that

18 combines machine learning with semiparametric spatial filtering methods to account for both

19 language- and document-specific characteristics, as well as phylogenetic and geographical

20 language relationships. We first establish that entropy rate distributions are highly consistent

21 across different language models, suggesting that the choice of model may have minimal

22 impact on cross-linguistic investigations. On the basis of a much broader range of language

23 models than in previous studies, we confirm results showing systematic differences in entropy

24 rates, i.e. text complexity, across languages. These results challenge the long-held notion that

25 all languages are equally complex. We then show that higher entropy rate tends to co-occur

26 with shorter text length, and argue that this inverse relationship between complexity and

27 length implies a compensatory mechanism whereby increased complexity is offset by

28 increased efficiency. Finally, we introduce a multi-model multilevel inference approach to

29 show that this complexity-efficiency trade-off is partly influenced by the social environment

30 in which languages are used: languages spoken by larger communities tend to have higher

31 entropy rates while using fewer symbols to encode messages.

32

33

2

Author Summary

To better understand human language as a complex system, our study leverages the power of computational language models and quantitative statistical analysis. We custom-trained seven different language models – from simple statistical models to state-of-the-art deep neural networks – on a database of 41 multilingual text collections, covering over 2,000 languages, more than 6,500 documents, and almost 3 billion words. We develop a novel statistical framework that makes it possible to account for language- and document-specific features, as well as phylogenetic and geographic relationships among languages. Our results show that, despite their architectural differences, these models produced remarkably consistent rankings of language complexity, challenging the long-held notion that all languages are equally complex. Moreover, we discovered a trade-off between complexity and efficiency: languages with higher complexity tend to use fewer symbols. We show that this trade-off is influenced by the social environments in which languages are used, with larger communities tending to use more complex but more efficient languages. These findings suggest that language complexity is not uniform but rather an adaptive response to the demands of communication within varying social contexts.

3

1 Introduction

1.1 Research objective

A model that assigns probabilities to sequences of linguistic symbols is called a language model (LM) [1]. While originally only trained on (vast amounts of) textual data to predict upcoming linguistic material [2], modern LMs demonstrate impressive and at times surprising capabilities in a wide range of scientific applications beyond linguistic tasks, such as predicting protein structures [3], forecasting time series [4], accelerating drug and material discovery [5,6], analysing genomic and epi-genomic data [7], and enhancing climate modelling [8].

In the fields of linguistics and natural language processing (NLP), LMs excel in traditional tasks, as evidenced by their ability to perform zero-shot learning, where they effectively generalise to new tasks without specific training, as shown by [9]. This highlights their potential to acquire human-like grammatical language through statistical learning, without relying on a built-in grammar [10]. On this basis, a vibrant research field has emerged, where LMs are being used as computational working models [11] or models of languages [12] to study different aspects of language processing and comprehension [2].

While we do not claim that LMs truly *understand* language [11] or attribute meaning in a similar way as humans do [13], we believe that LMs are interesting from a cross-linguistic perspective, because they can be used as idealised language learners that learn to produce and process language by being trained on a corpus of linguistic input [10,11,14]. A central goal of linguistics is to understand the diverse ways in which human language can be organised. By training an LM on linguistic material in different languages, researchers can investigate how these models learn and generalise linguistic structures and rules across different languages and language families.

4

In this paper, we use this framework to enhance our understanding of human language as a complex system. To this end, we custom-trained seven different LMs on a database of more than 6,500 different documents, covering over 2,000 languages.

The main contributions of this paper are fourfold:

- We develop a novel quantitative framework that integrates machine learning with semiparametric spatial filtering methods. This framework allows for the simultaneous consideration of language- and document-specific features, as well as phylogenetic and geographic relationships among languages. Additionally, we introduce a multi-model multilevel inference approach designed to test whether cross-linguistic variation is statistically associated with sociodemographic factors, while accounting for phylogenetic and spatial autocorrelation.
- We measure language complexity using information-theoretic entropy rates and extend previous work by applying this method to a broader range of language models (LMs) – from simple statistical models to machine learning models, among them advanced deep learning models including transformers. Additionally, we expand the analysis to include sub-word but supra-character encoding levels, offering a more detailed and layered perspective on language complexity. We demonstrate that, despite their pronounced architectural differences, the investigated LMs produce remarkably consistent rankings of language complexity across all considered symbolic levels. This challenges the idea that all languages are equally complex.
- We present, discuss, and evaluate evidence for a previously undocumented trade-off between complexity and efficiency: higher entropy rates tend to co-occur with shorter text lengths. From an information-theoretic perspective, message length quantifies efficiency – the shorter the message, the higher the efficiency [15]. We argue that this

5

103   inverse relationship between complexity and length implies a compensatory

104   mechanism, whereby increased complexity is offset by greater efficiency.

105   -   We show that this trade-off is influenced by the social environments in which

106   languages are used, with larger communities tending to use more complex but more

107   efficient languages.

108   1.2 Background and related work

109 Parallel corpora are valuable in a cross-linguistic context because they enable systematic

110 comparisons of language processing across different languages [16]. By providing translations

111 of the same text in multiple languages, parallel corpora allow researchers to study how LMs

112 handle similar content across varying grammatical structures and lexicons. These datasets also

113 facilitate the testing and understanding of linguistic laws, i.e., statistical patterns shared across

114 human languages [17], or the examination of whether languages adapt to the geographical or

115 sociodemographic environments in which they are learned and used [18–22]. Yet another idea

116 – dating back to the work of Greenberg [23] – that was revived with the development of large

117 parallel corpora [24] is to use parallel texts to classify and compare languages [16]. Examples

118 of such cross-linguistic studies are [25–29]. However, the majority of the aforementioned

119 studies are based on very peculiar text types, especially translations of the Bible and there are

120 several important challenges that the use of the Bible as a parallel text source pose [20,30,31].

121 To address this limitation, we leveraged available corpora and multilingual text collections

122 [32–35] and compiled a database of parallel texts comprising a large variety of different text

123 types, e.g. religious texts, legalese texts, subtitles for various movies and talks, and machine

124 translations. In addition, we added comparable corpora, i.e., texts that are not parallel but

125 come from comparable sources and are therefore similar in content, again comprising very

126 different text types/genres, e.g. newspaper texts, web crawls, Wikipedia articles, translation

127 tables for system messages in the Ubuntu operating system, or translated example sentences

6

128 from a free collaborative online database. Furthermore, we added information from the

129 Crúbadán project [36] that aims at creating text corpora for a large number of (especially

130 under-resourced) languages. In total, the compiled database contains 41 different multilingual

131 corpora, comprising nearly 3 billion words or nearly 9 billion Unicode characters across more

132 than 6,500 documents and covering over 2,000 languages. These languages are spoken as a

133 native language by more than 90% of the world's population and represent almost half of all

134 languages with a standardised written representation.

135 In a recent paper [37], we presented the first results based on this database. In this study, our

136 primary focus was a cross-linguistic examination of language complexity, a topic that has

137 garnered significant attention in linguistics and related fields over the past two decades – for

138 an overview, see [38]. In our study, we quantitatively evaluated the so called equi-complexity

139 hypothesis that suggests that all human languages, despite their diverse and varied nature,

140 have the same level of overall complexity [39]. To overcome the difficulty of measuring

141 overall language complexity [40], we leveraged information theory, an area of mathematics

142 that links probability and communication [41] and provides notions of complexity that are

143 both objective and theory-neutral [42]. To this end, we trained a simple statistical LM on each

144 of our documents and statistically analysed the training process to infer the average per-

145 symbol information content or entropy rate of each document, which can be interpreted as a

146 measure of complexity [43,44]: the harder it is, on average, to predict upcoming text, the

147 higher the entropy rate, the greater is the complexity of the text as a whole [10,45–47]. We

148 argued the entropy rate can thus also be used to compare the complexity of different

149 languages. We then statistically compared complexity rankings across different corpora by

150 calculating correlation coefficients between the entropy rates across all possible pairs of

151 multilingual corpora. For example, we correlated the entropy rate rankings derived from a

152 corpus of movie subtitles in various languages with those from a similarly diverse corpus of

7

153 Wikipedia sentences. This approach, applied comprehensively to all pairs among our 41

154 different multilingual corpora, makes it possible to assess the consistency of complexity

155 rankings across various types of linguistic data. From an information-theoretic point of view,

156 we showed that our results constitute evidence against the equi-complexity hypothesis: a

157 language with high/low entropy rate in one corpus also tends to be more/less complex in

158 another corpus.

1.3   Research Question and Scope

159

160 As higher complexity in language results in more demanding processing efforts,

161 encompassing both language production and comprehension, our study [37] naturally leads to

162 the question: *Why is there a trend towards increased complexity in certain languages?* In the

163 present study we pursue this issue by providing evidence suggesting that languages with high

164 entropy rates, which indicate greater complexity, require fewer symbols to encode messages.

165 This points to a compensatory mechanism: higher complexity is balanced by increased

166 efficiency, in terms of shorter message lengths.

167 The main purpose of this study is to present, discuss and evaluate evidence for such a

168 complexity-efficiency trade-off, while also striving to enhance the following empirical and

169 methodological aspects of our prior work.

170 First, in our previous paper we trained a rather simple statistical LM. However, as Baroni [48]

171 pointed out, different LMs have uniquely structured internal architectures, and thus cannot be

172 viewed as "blank slates". Instead, they should be regarded as algorithmic linguistic theories

173 "encoding non-trivial structural priors facilitating language acquisition and processing" [48].

174 These architectural differences affect how LMs acquire and process language. For instance, n-

175 gram models primarily rely on local context and assume fixed-length dependencies. To

176 estimate the probability of the next symbol, n-gram models build a conditional probability

177 distribution based on a limited context window that only takes into account the preceding few

8

178   words (e.g. n = 2, 3, 4, 5) [2]. This restriction on context can lead to a loss in the model's

179   ability to capture long-range dependencies that are essential for grammatical complexity in

180   human languages. More advanced models, such as transformers and other neural network

181   architectures, are equipped with mechanisms like self-attention that allow them to capture

182   complex, long-range dependencies and richer contextual information from the input data [2].

183   These architectural differences enable different models to capture linguistic complexity at

184   various levels, from basic statistical patterns to more nuanced syntactic and semantic

185   relationships.

186   To extend our previous work, we thus train various types of LMs on our data, ranging from

187   simple statistical n-gram models to state-of-the-art transformer models. This approach allows

188   us to compare how models with varying levels of sophistication interpret and represent

189   language complexity, providing a more comprehensive understanding of the phenomenon

190   across models. Surprisingly, we show that, from a cross-linguistic perspective, the type of LM

191   has relatively little impact on the obtained results.

192   Secondly, we improve and extend our prior work methodologically by developing a machine

193   learning method that fully accounts for the relatedness of languages: when comparing

194   languages, statistical challenges arise due to the fact that closely related languages often

195   exhibit more similarities among themselves in various aspects than they do with more

196   distantly related languages. Additionally, languages originating from the same regions often

197   tend to be influenced by common factors, further complicating the analysis [49–51]. While we

198   have included language family, macro-area and country as factors to account for the

199   genealogical and geographic relatedness of languages in our prior paper, this approach ignores

200   variation within language families and geographical units as pointed out in several recent

201   studies [49–53]. To address this issue, we develop two quantitative approaches: (i) a

202   semiparametric machine learning estimation method capable of simultaneously controlling for

203 document- and language-specific characteristics while directly modelling potential effects due

204 to phylogenetic relatedness and geographic proximity; (ii) a multi-model multilevel inference

205 approach designed to test whether cross-linguistic outcomes are statistically associated with

206 sociodemographic factors, while accounting for phylogenetic and spatial autocorrelation via

207 the inclusion of random effects and slopes.

1.4   Structure of the paper

209 The structure of this paper is as follows: the next section introduces the multilingual database

210 and details the procedures for compiling the text data (Sect. 2.1). This is followed by a

211 description of the sociodemographic and linguistic variables considered in this study (Sect.

212 2.2). We then introduce the investigated LMs (Sect. 2.3) and describe how the textual data

213 was pre-processed (Sect. 2.4). The methodology for estimating entropy is presented in Sect.

214 2.5. Sect. 2.6 is devoted to statistical methods. We first present a novel method for evaluating

215 the similarity of entropy rates and length distributions across different corpora (Sect. 2.6.1),

216 followed by a description of the multi-model inference approach used to examine the impact

217 of the number language users on the entropy-length trade-off (Sect. 2.6.2).

218 In Sect. 3, we present our findings. The paper concludes with the "Discussion" section, where

219 we evaluate potential limitations of our approach, examine the relevance of the complexity-

220 efficiency trade-off, and outline directions for future research (Sect. 4).

221 All data and code (Stata v18.0 and Python v3.6.8) needed to replicate our analyses are

222 available at https://osf.io/xdwjc/. In addition, interactive results and visualisations are

223 available online at https://www.owid.de/plus/tradeoffvis/.

2 Materials and methods

225 Some material in this section is recycled from our prior publications [37,54], in accordance

226 with the guidelines provided by the Text Recycling Research Project [55].

2.1 Database

In what follows, we give an overview regarding the database. Additional in-depth details regarding all corpora can be found in [37]. In total, we analysed 41 different multilingual text collections. 40 text collections consist of actual full text data, while the remaining collection consists of word frequency information from the Crúbadán project [36]. Of the 40 full-text collections, 33 are fully parallel and 7 corpora contain comparable documents. The full-text corpora can be loosely categorised into the following text types: 5 *religious* text collections, 4 news/Wikipedia/*Web crawls* text collections, 5 text collections containing *legalese* texts, 22 multilingual *subtitle* corpora and 4 collections of *other* text types.

2.1.1 Text types

Religious texts

The two parallel text collections *BibleNT* and *BibleOT* are both part of the Parallel Bible Corpus (PBC) made available by Mayer and Cysouw [33] and containing a total of 1,568 unique translations of the Bible. The *BibleNT* text collection consists of all 27 books that belong to the New Testament of the biblical canon. In total, $N_D = 1,459$ different documents, i.e., translations of the New Testament into another language, are available for $N_L = 1,093$ individual languages. The median length of individual documents is $\tilde{L}_W = 227,391$ words and $\tilde{L}_C = 1,190,294$ characters. Correspondingly, the *BibleOT* text collection consists of all 39 books that belong to the Old Testament ($N_D = 254$; $N_L = 147$; $\tilde{L}_W = 642,772$; $\tilde{L}_C = 3,259,354$). Each translation was pre-tokenised, Unicode normalised, and spaces were inserted between words, punctuation marks, and non-alphabetic symbols by the corpus compilers, who also manually checked and corrected the texts where necessary [33]. Additionally, great care was taken to account for language-specific peculiarities. For instance, in the Austronesian language Arifama-Miniafia, the right single quotation mark represents the

251 glottal stop [33]. Another example is tone marking, which in languages such as Chinantec,

252 Mazatec, or Nambikuára is represented by raised numbers indicating tones. For instance, in

253 the Sochiapan Chinantec translation of the sentence "I am the God of Abraham," the text

254 reads: "Jná¹³ bíh¹ la³² Dió³² Juo¹³ Há²bran²¹" Automatically tokenising this sentence into words

255 (see Sect. 2.4) would mistakenly split words into separate parts, e.g., "Há bran." For further

256 information, see Bentz et al. [27], Appendix A.

257 Another distinctive feature of the PBC, compared to the other multilingual text collections in

258 corpus, is the usage of full vocalisation using tashkīl diacritics, as highlighted by Gutierrez-

259 Vasques et al. [56]. In Arabic, short vowels are usually not written – most registers omit them,

260 while some use diacritical marks to indicate their presence. For example, in Egyptian Arabic,

261 the PBC encodes the word for "peace" as "salām" (سَلَام), marking the short vowels with

262 diacritics. In other text collections, only the 'skeleton' letters "sl'm" (سلام) are provided,

263 leaving the short vowels implied. As discussed further in Sect. 2.4, this lack of vowel

264 representation complicates computational and quantitative analyses. The quality and

265 consistency of the PBC in encoding such features enhances the reliability of our cross-

266 linguistic results. For further information regarding text pre-processing, see Sect. 2.4.

267 The two parallel text collections *WatchtowerV1* and *WatchtowerV2* are also part of the

268 Parallel Bible Corpus, both containing translations of different introductory texts of the

269 Jehovah's Witnesses' official web site [16] (*WatchtowerV1*: $N_D = 142$; $N_L = 140$; $\tilde{L}_W =$

270 129,008; $\tilde{L}_C = 659,563$; *WatchtowerV2*: $N_D = 265$; $N_L = 260$; $\tilde{L}_W = 7,194$; $\tilde{L}_C = 35,608$).

271 The *Quran* collection consists of parallel translations of the central text of the Islam

272 downloaded from http://tanzil.net/trans/ (accessed 4/30/20, $N_D = 43$; $N_L = 43$; $\tilde{L}_W =$

273 182,950; $\tilde{L}_C = 860,590$).

Web crawls

The *GlobalVoices* comparable collection consist of contributions to the citizen media platform Global Voices. Raw text files of all articles were downloaded from http://casmacat.eu/corpus/global-voices.html (version: 2018Q4; accessed 4/30/20, $N_D = 40$; $N_L = 39$; $\tilde{L}_W = 20{,}021$; $\tilde{L}_C = 112{,}541$). The other three collections were compiled based on plain text files from the Leipzig Corpora Collection (LCC) [35] that presents corpora in a uniform format. Here we focus on three collections that we name as follows (i) *LCCnews*, i.e., text material of crawled newspapers available online ($N_D = 112$; $N_L = 85$; $\tilde{L}_W = 196{,}899$; $\tilde{L}_C = 1{,}119{,}551$), (ii) *LCCweb*, i.e., text material crawled from randomly chosen web pages ($N_D = 87$; $N_L = 85$; $\tilde{L}_W = 195{,}953$; $\tilde{L}_C = 1{,}127{,}076$), and (iii) *LCCwiki*, i.e., text material from Wikipedia dumps ($N_D = 171$; $N_L = 171$; $\tilde{L}_W = 185{,}770$; $\tilde{L}_C = 1{,}038{,}774$). Each document in each corpus consists of 10,000 randomly shuffled sentences in the corresponding language.

Legalese texts

To compile the *UDHR* parallel collection, we downloaded parallel translations of the Universal Declaration of Human Rights from https://unicode.org/udhr/ (accessed 4/30/20, $N_D = 452$; $N_L = 399$; $\tilde{L}_W = 1{,}978$; $\tilde{L}_C = 10{,}822$). The other four legalese parallel text collections are all obtained from the *OPUS* project [32]. The *EUconst* collection consists of different translations of the European Constitution ($N_D = 21$; $N_L = 21$; $\tilde{L}_W = 92{,}607$; $\tilde{L}_C = 620{,}502$). *Europarl* is a corpus of documents extracted from the European Parliament web site ($N_D = 21$; $N_L = 21$; $\tilde{L}_W = 5{,}362{,}935$; $\tilde{L}_C = 31{,}959{,}314$). The collection *EUmed* is compiled from PDF documents from the European Medicines Agency ($N_D = 22$; $N_L = 22$; $\tilde{L}_W = 3{,}241{,}844$; $\tilde{L}_C = 18{,}714{,}208$). *UNPC* consists of manually translated documents in the six official languages of the United Nations ($N_D = 6$; $N_L = 6$; $\tilde{L}_W = 341{,}723{,}872$; $\tilde{L}_C = 879{,}903{,}168$).

13

Subtitles

The parallel subtitle collections consist of two types: subtitles of movies and subtitles of TED talks. The 13 subtitle collections are based on the ParTy corpus [34]. The Technology, Entertainment, Design (TED) talk subtitles were downloaded from https://amara.org/en/teams/ted/videos/ (accessed 4/30/20). Information regarding movie/talk titles, number of translations/languages per corpus and median lengths are provided in Table 1.

Table 1: Overview of the subtitle text collections. 1st column: collection. 2nd column: collection id. 3rd column: movie/talk title. 4th column: number of documents. 5th column: number of different languages (*ISO-639-3* codes). 6th/7th column: median text length in words/characters.

Collection	ID	Title	N_D	N_L	$\tilde{L}_W$	$\tilde{L}_C$
Movies	MSub01	Amelie	29	29	7,349	34,207
	MSub02	Avatar	26	26	9,695	43,988
	MSub03	Black Swan	37	37	4,661	20,444
	MSub04	Bridge of Spies	16	16	12,690	60,718
	MSub05	Das Leben der Anderen	15	15	8,908	42,103
	MSub06	Frozen	27	27	7,832	34,262
	MSub07	Gone Girl	12	12	17,317	79,075
	MSub08	Grand Budapest Hotel	9	9	9,598	46,148
	MSub09	Imitation game	15	15	10,303	49,983
	MSub10	Inception	28	28	10,937	53,059
	MSub11	Ironlady	14	14	9,679	45,133
	MSub12	Noah	30	30	5,516	25,804
	MSub13	Spectre	14	14	7,533	34,984
TED Talks	TEDt01	Bring on the learning revolution	50	50	3,075	14,455
	TEDt02	Do schools kill creativity	60	59	3,555	17,360
	TEDt03	Doing the impossible cutting through fear	61	60	3,763	18,468
	TEDt04	Modern Warrior	31	31	2,099	10,409
	TEDt05	My philosophy for a happy life	31	31	1,853	8,658
	TEDt06	Secondary sugar kills	28	28	1,369	6,318
	TEDt07	Speak to the heart	74	71	1,376	7,017
	TEDt08	Success is a continuous journey	49	48	770	3,669
	TEDt09	Why is x the unknown	52	51	562	2,746

Other

The comparable collection *Ubuntu* consists of Ubuntu localization files. Texts are available from OPUS [32] ($N_D = 86$; $N_L = 86$; $\tilde{L}_W = 5,780$; $\tilde{L}_C = 35,888$). To compile the parallel *Google Translate* collection, we used Google Translate (https://translate.google.com; accessed 09/04/2019) to machine translate a short passage excerpt from the book "Aldono al

14

315   la Dua Libro de l' Lingvo Internacia" in the constructed language Esperanto, written by its

316   inventor L.L. Zamenhof (downloaded from

317   http://esperanto.davidgsimpson.com/librejo/index.html on 09/04/2019) into 102 languages

318   (see [37] for the translated passage; $N_D = 102$; $N_L = 102$; $\tilde{L}_W = 716$; $\tilde{L}_C = 3{,}875$). The two

319   comparable corpora *Tatoeba V1* and *Tatoeba V2* are compiled based on *Tatoeba*, a

320   collaborative online platform that makes available sentences translated into different

321   languages. Raw data were downloaded from https://tatoeba.org/deu/downloads (accessed

322   4/30/20; *V1*: $N_D = 183$; $N_L = 183$; $\tilde{L}_W = 656$; $\tilde{L}_C = 3{,}034$; *V2*: $N_D = 123$; $N_L = 123$; $\tilde{L}_W =$

323   $3{,}438$; $\tilde{L}_C = 15{,}710$).

324   **Word list**

325   To generate the word list data, we downloaded all available lists from the Crúbadán project

326   [36] from http://crubadan.org/files/ (accessed 4/30/20). In total, we arrived at 2,216 word

327   frequency lists for a total of $N_L = 1{,}943$ different languages ($\tilde{L}_W = 101{,}079$).

328   **2.1.2 Overview of the database**

329   Figure 1 displays a map highlighting the geographical distribution of languages for the

330   compiled multilingual database. The figure reveals an imbalance at the language level within

331   the database: over 100 languages have at least 10 documents, but approximately 75% of

332   languages have fewer than four documents. This scarcity reflects the limited electronic

333   availability of documents in languages spoken by smaller populations [36]. This is

334   exemplified by the contrast in median speaker numbers: while the median for all non-extinct

335   languages documented by Ethnologue stands at 8,000 [57], the median for languages

336   represented with at least one document in our database is significantly higher, at 30,000. In

337   addition, the majority of our text collections contain a comparatively small number of

338   individual documents, with a median of 40 documents per corpus. This limited size can be

15

339 attributed to specific reasons in certain cases; for example, the *EUconst* collection is naturally

340 restricted to translations of the European Constitution into the official languages of the

341 European Union. In contrast, for other collections such as the subtitle corpora, translations

342 into further languages were not available when we compiled the database. On the other side of

343 the spectrum, we have 11 multilingual text collections that consist of more than 100 different

344 documents. As described above, documents are rather short, e.g. 25% of the documents are

345 below 14,575 characters or 3,181 words. However, 200 documents are longer than 1 million

346 characters, 49 documents are longer than 10 million characters and the longest documents are

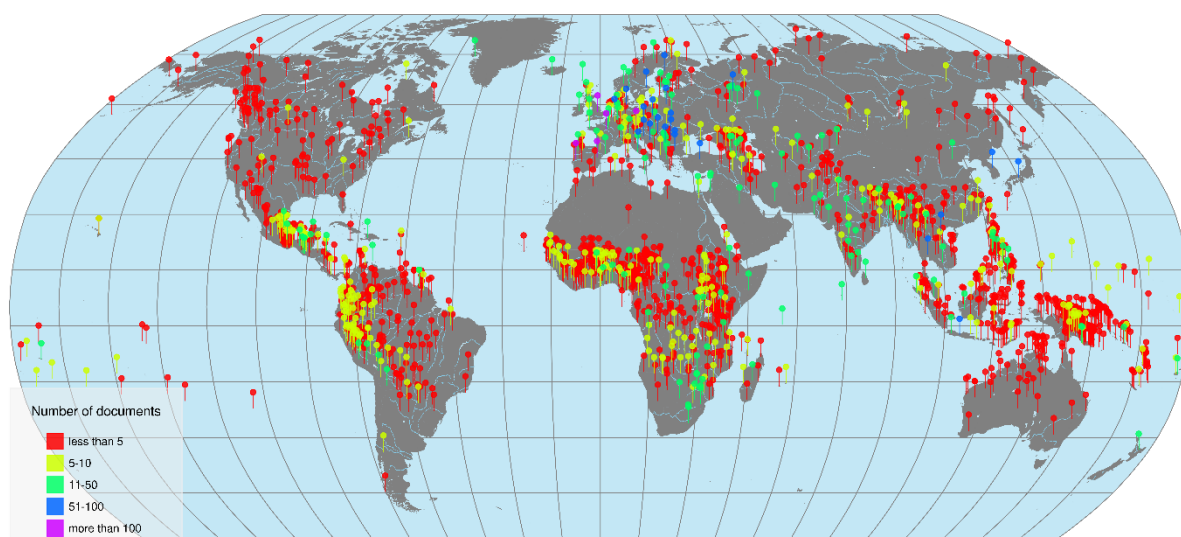347 several hundred million words and more than a billion characters long.

348

349 **Figure 1: Global distribution of collected documents per language.** Approximately 76% of languages have
350 fewer than five documents. On the other side of this spectrum, over 160 languages have more than 10
351 documents. This imbalance reflects the limited electronic availability of documents in languages spoken by
352 smaller populations [36].
353

354 In what follows we statistically compare the structure found in smaller corpora (i.e., those

355 consisting of shorter documents and/or a limited number of available documents) with the

356 structure found in larger corpora (i.e., those consisting of longer documents and/or data points

357 for many languages). The idea is that if the results from both smaller and larger corpora align,

358 this strengthens the claim that these results are not merely artefacts resulting from database

16

359   bias. Additionally, we include control covariates, such as the number of corpora per language,

360   to account for the unbalanced nature of our database, as described in the next section.

361   **2.2   Sociodemographic and linguistic variables**

362   Information on speaker population size, corpus, language family, language (identified by its

363   *ISO*-639-3 code), macro-area, writing script, speaker population size, longitude and latitude

364   are taken from [37]. Writing script refers to the writing system in which the corresponding

365   document is written in. In total, we have information for $N_D = 6{,}513$ different documents

366   using 51 different writing scripts. The vast majority ($N_D = 5{,}183$ or 79.58%) of our

367   documents are written in Latin script, followed by Cyrillic ($N_D = 463$ or 7.11%), Arabic

368   ($N_D = 156$ or 2.40%) and Devanagari ($N_D = 98$ or 1.50%). Expanded Graded

369   Intergenerational Disruption Scale (EGIDS) level information was initially sourced from [58],

370   which is reported in Glottolog [59] (v4.2.1). Country is defined by Ethnologue [60,61] as the

371   primary country/country of origin of the language in question [57]. To ensure completeness,

372   we manually supplemented missing data from [58] by cross-referencing with Glottolog and

373   Ethnologue. The EGIDS level serves as a measure of a language's endangerment status [62].

374   We use the EGIDS level as a covariate to control for potential translation effects [63,64], as

375   languages with lower EGIDS levels are presumably more likely to be used as source

376   languages, while languages with higher EGIDS levels are presumably more likely to be used

377   as target languages. For example, an EGIDS level of 0 (labelled "International") pertains to

378   the six official United Nations languages: Arabic, Chinese, English, French, Russian, and

379   Spanish. On the other hand, languages with values of five and above pertain to languages that

380   are not used in formal education, mass media or by the government, and they may

381   consequently be more susceptible to (more) pronounced "translationese" influences [64]. With

382   a similar logic in mind and to account for the unbalancedness of our database (see Sect.

17

383 2.1.2), we also consider the number of corpora with a least one available document per

384 language as an additional control variable in what follows.

385 Further information regarding the classification of languages into macro-family and sub-

386 family are taken from [65]. We manually added information for languages that was missing

387 by using publicly available genealogical classifications (see the script

388 'prepare_language_info.do' available at https://osf.io/tkgph/ for details). Classifications in [65]

389 are given as comma-separated values. We define the first value as the macro-family and the

390 second one as the sub-family, e.g. for the language "Ghotuo" the classification is "Niger-

391 Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Edoid, North-Central, Ghotuo-Uneme-

392 Yekhee", so the macro-family is "Niger-Congo" and the sub-family is "Atlantic-Congo".

393 Additionally, we use a phylogenetic similarity matrix also provided by [65] that is based on

394 word lists from the Automated Similarity Judgment Program (ASJP) [66]. Information on the

395 number of countries in which each language is spoken was sourced from Glottolog (v4.2.1).

396 We manually supplemented missing data by cross-referencing with Ethnologue [60,61]. The

397 rationale behind considering this variable as a potential covariate is to account for the varying

398 degrees of pluricentrism [67]. For instance, languages such as Chinese or Spanish are spoken

399 in several countries and may therefore have different codified standard forms.

400 For further information and a discussion of potential caveats and problems regarding the

401 assignment of environmental variables to individual languages in order to reflect local

402 grouping structure, see [68,69].

403 **2.3 Language models**

404 We use general-purpose data compression algorithms, taking advantage of the fact that

405 language modelling and lossless compression are essentially equivalent [70–72]. All data

406 compression algorithms consist of a model and a coder [14]. Our focus is on the class of

407 (lossless) compressors where the algorithm uses training data to estimate a model, i.e., a
18

408 conditional probability distribution, that can be used to generate predictions about upcoming

409 symbols. To perform compression, the predicted probabilities are then used to encode

410 symbols using a technique called arithmetic encoding [73].

411

412 **Table 2: Language models.** The table lists each investigated LM along with its implementation techniques,
413 source, and time required to train it on a document of median length. Training times are provided in seconds and
414 categorised based on the hardware used: Central Processing Unit (CPU) and High-Performance Computing
415 (HPC) cluster with Graphics Processing Unit (GPU) support. The first five LMs were run exclusively on a CPU,
416 while the remaining two LMs were run on a GPU. For comparison, we also include the computation time
417 required if these two models are run on a CPU. Further implementation details are given in Supporting
418 Information S1.

LM	Technique/Algorithm	Model specification	Source	Time
PPM2	N-gram modelling [74], prediction by partial matching [75], memory: 2000 megabytes	order 2	[76,77]	0.1 (CPU)
PPM6		order 6		0.1 (CPU)
PAQ	Context mixing [14,78], gated linear network [79]	weights ~1.7 million, parameters ~3,800	[80,81]	54.5 (CPU)
LSTM	Long short term memory [82]	parameters ~3.93 million	[83]	98.5 (CPU)
TRFsmall	Transformer [84]	parameters ~2.24 million		150.5 (CPU)
TRFmed		parameters ~19.1 million		1,252.5 (CPU) / 21.2 (GPU)
TRFbig		parameters ~279 million		37,759.8 (CPU) / 58.9 (GPU)

419

420 The seven LMs that we investigate are summarised in Table 2. In what follows, further details

421 are given for each language model. Prediction by partial matching (PPM) [75,85] is a

422 dynamic and adaptive variable-order n-gram LM. The algorithm assumes the Markov

423 property: to predict the next symbol, the algorithm uses the last o immediately preceding

424 symbols. For **PPM2**, we set o to 2, i.e., the last 2 symbols are used as context to generate

425 predictions. For **PPM6**, we set o to 6. In both cases, the level of compression is set to

426 maximum and the size of used memory is set to 2,000 megabytes. **PAQ** can be described as a

427 weighted combination of predictions from a large number of models, where the individual

428 models are combined using a gated linear network [14,78,79,81]. The network has a single

19

429   layer with 552 input nodes and 3,080 input weights. The model has a total of ~1.7 million

430   weights, but due to a sparse updating scheme which leads to faster compression and

431   decompression, the effective number of parameters used in training is significantly lower.

432   Only $552 \cdot 7 = 3,864$ weights are updated for each bit of data. We use version PAQ8o and set

433   the compression level to maximum, requiring 1,712 megabytes of memory.

434   NNCP [83] is a lossless data compressor that is based on the Transformer XL model defined

435   in [86]. Modifications to the original Transformer XL model and algorithmic details are

436   provided in [87,88]. As for LSTM, the Adam optimiser is used and we use the "encode only"

437   mode. For **TRFsmall** (version 3.1), we use the default options with four layers and a model

438   dimension of 256, resulting in a total number of ~2.24 million parameters. For **TRFmed**

439   (version 3.2), we set the number of layers to 12 and the model dimension to 512, resulting in a

440   total number of ~19.1 million parameters. For **TRFbig** (version 3.2), we use the available

441   "enwik9" profile that sets the number of layers to 20 and the model dimension to 1,024,

442   resulting in a total number of ~279 million parameters.

443   In addition, NNCP offers compression based on a Long Short-Term Memory deep neural

444   network (**LSTM**) [82]. We use four layers of LSTM cells. The network is trained using

445   truncated-like backpropagation [87,89] and Adam optimisation is used to update network

446   weights [90]. We do not use a text pre-processor or tokeniser and we use the faster "encode

447   only" mode (the output cannot be decompressed, but the compression itself is still lossless).

448   The total number of parameters is ~3.93 million parameters.

449   In addition, when discussing the relevance of the complexity-efficiency trade-off (Sect. 0), we

450   use OpenAI's GPT-2 model [91] with ~1.5 billion parameters, as implemented in the Hugging

451   Face library [92].

20

2.4   Text pre-processing and information encoding units

452

453 Each document is tokenised and Unicode normalised where necessary. All uppercase

454 characters are lowered based on the closest language-specific International Organization for

455 Standardization (ISO) code. Unless otherwise specified in [37], the word-break algorithm of

456 the *International Components for Unicode* library [93; Annex #29] was used to detect word

457 boundaries (in texts without spaces or marks between words, a dictionary lookup method is

458 used by the algorithm [94]). More details regarding each individual text collection can be

459 found in [37].

460 Following [27], we represent a text κ as a random variable that is created by drawing (with

461 replacement) from a set of symbol types $\mathcal{V} = \{s_1, s_2, s_3, \ldots, s_V\}$, where V is the number of

462 symbol types, i.e., $V = |\mathcal{V}|$. Correspondingly, a symbol token is any reoccurrence of a symbol

463 type [27]. In what follows, we estimate the relevant information-theoretic quantities for the

464 following information encoding units/symbol types: (i) (Unicode) characters, (ii) words and

465 (iii) sub-word units. For (iii), we apply byte pair encoding (BPE) [95–97] to split words into

466 one or several units and then train the different LMs over the resulting sequences of sub-word

467 units. We follow [95] and set the number of BPE merges to $0.4 \cdot V$.

468 Note that we neither remove spaces nor punctuation at the character level, nor before applying

469 BPE, since punctuation can provide reliable information regarding stylistic features and

470 differences between authors [98]. Correspondingly, on the word level, each punctuation mark

471 is treated as a separate token, except for the dash, which ensures that dash-separated words

472 (e.g., 'presidency-in-office') are treated as a single word type. After tokenisation into

473 words/sub-word units, each word/BPE type is replaced by one unique Unicode symbol. The

474 different compression algorithms are then used to compress the resulting symbol sequence.

475 On the BPE level, we also compress the mapping of sub-word units to 4-byte Unicode

476 symbols.

21

477 *Limitations & caveats*

478 It is important to acknowledge several caveats when interpreting cross-linguistic results based

479 on the quantitative analysis of written language text at both the character and word levels

480 [27,95,99–101]. First, our analysis is based exclusively on corpora of written language, which

481 begs the question whether it applies to human languages in general. In this context, it is

482 important to note that "written language cannot be regarded as *merely* the transference of

483 spoken language to another medium", as John Lyons put it in his classical introduction to

484 theoretical linguistics [102]. Indeed, there is a long-standing strand of research that recognises

485 written language as a distinct linguistic object worthy of independent study, backed up by a

486 large body of evidence against a mere derivative nature of written language, including

487 evolutionary and biological aspects [103–105]. Obviously, the texts in the PBC (see Sect.

488 2.1.1) are one prime example of texts that are not parasitic on prior oral language production.

489 From this stance, our results clearly apply to written languages, seen as more or less

490 'autonomous' communication systems.

491 While we do not have analysed genuine oral language data due to the lack of sufficiently large

492 and phonemically annotated corpora, it is a valid question whether our results remain valid if

493 we treat the texts in our corpora as representing spoken language, which basically means that

494 we interpret grapheme sequences as standing in for phoneme sequences. A notable challenge

495 arises from the differences in the mapping between phonemes and graphemes across

496 languages [64,99]. For instance, languages with deep orthographies like English have

497 inconsistent phoneme-grapheme correspondences (e.g., "ough" in "thought" vs. "through" vs.

498 "dough" rendering /ɔː/, /uː/, and /əʊ/ phonemes, respectively), while languages with shallow

499 orthographies like Spanish have more systematic correspondences (e.g., "a" as in "casa"

500 always corresponds to /a/) [96]. Another example is the difference in the representation of the

501 phoneme /tʃ/ between Czech and German: Czech uses a single character ("č"), while German

22

uses a four-character sequence ("tsch") to represent the same phoneme [64]. As mentioned in Sect. 2.1.1, certain characters like short vowels are sometimes omitted or represented by diacritics, depending on the orthographic conventions of a given language.

A significant example for a well-known mismatch between a graphemic and an 'underlying' phonemic representation is how lexical tone is represented in tone languages. In some languages, such as Thai and Lao, tone is indicated using a combination of tone marks (diacritics) and consonant characters, while in Hmong, tone is marked with specific characters, like a letter at the end of words. Yet in other tone languages, like Mandarin Chinese, lexical tone is typically not encoded in the written form. This variation mirrors the challenges we previously discussed in the PBC, where tone languages like Chinantec or Mazatec use raised numbers to represent pitch, adding complexity to tokenisation and analysis (see Sect. 2.1.1). In languages where tone is omitted in writing, complexity estimates for the phonemic representation may be understated, while languages that explicitly mark tone could appear more complex. These differences in tonal representation across writing systems must be considered when interpreting cross-linguistic complexity estimates, especially between tone and non-tone languages.

Independent from the distinction between spoken and written language, there are several factors that add further complexity to cross-linguistic word-level analyses. First, differences in writing systems are highly relevant for cross-linguistic comparisons of written text. For example, written Mandarin Chinese uses a logographic system where characters typically represent words or morphemes, while languages like English or Russian use alphabetic systems, where symbols usually represent phonemes.

Secondly, while we use an orthography-based algorithmic definition of "word", which is standard in computational linguistics, it is important to note that there is no universally accepted definition of "word-hood" across languages [100,101], although see Haspelmath

23

527   [106]. However, Geertzen et al. [107] demonstrate that the computational approach to "word-

528   hood" we employ may be sufficiently well-suited within an information-theoretic and

529   compression-based quantitative framework for capturing linguistic regularities in cross-

530   linguistic analyses. Moreover, the substantial effort and typologically informed curation of the

531   PBC adds further reliability to our cross-linguistic analyses (see Sect. 2.1.1). Therefore, if the

532   patterns observed in the other multilingual text collections we investigate align with those

533   from the PBC, this consistency would bolster confidence in the robustness and validity of our

534   quantitative results.

535   Thirdly, languages show considerable variation with respect to the morphological complexity

536   of words. Words in highly synthetic languages may, for example, contain long chains of

537   derivational affixes, such as in Turkish *tan-ış-tır-ıl-a-ma-dık-lar-ın-dan-dır* 'it is because they

538   cannot be introduced to each other' [108]. Verbs in particular may exhibit complex

539   inflectional paradigms featuring incorporated object nouns and multiple tense, mood, and

540   agreement markers, e.g. in Cayuga *t-ę-hęn-atat-hǫna't-a-yę́:thw-ahs* 'they will plant potatoes

541   for each other' [109]. Importantly, the status of such long morpheme sequences as single

542   words is, in general, not controversial. In contrast, analytical languages like modern English

543   express similar grammatical elements with multiple words, showing little or no inflection.

544   This means that the number of grammatically possible words differs by several orders of

545   magnitude between languages. This has a significant impact on word-level quantitative

546   analyses, where any two different words are treated as unanalysed, unrelated entities. In

547   addition, the number of words required to express the same propositional content varies

548   significantly across languages.

549   To further address the aforementioned challenges, we compute our quantitative estimates at

550   multiple levels of linguistic structure: characters, words, and the supra-character, sub-word

551   level using BPE [95,96]. BPE is a sub-word segmentation technique that iteratively merges

24

552   the most frequent pairs of characters or character sequences, creating sub-word units that can

553   capture many meaningful linguistic patterns. This method is crucial in modern language

554   modelling, as it efficiently handles morphological variation and rare words, enhancing model

555   performance across diverse languages. Moreover, BPE's ability to extract language-specific

556   sub-word patterns from raw text makes it particularly valuable in cross-linguistic studies, as it

557   reveals structural differences and encodes features that align with those described in

558   traditional linguistic typology, as recently discussed in-depth by Gutierrez-Vasques et al. [56].

559   By comparing language-specific information-theoretic estimates across symbolic levels,

560   languages, and corpora, we can assess the consistency of results. If these estimates, derived

561   from different symbolic levels and corpora with varying qualities for cross-linguistic studies,

562   point in the same direction, this would strengthen the validity of our quantitative findings.

563   Additionally, as described in Sect. 2.6, we statistically account for potential confounding

564   factors by including the type of writing script as a covariate in our analyses. In the multi-

565   model multilevel analyses (Sect. 2.6.2), we also include random intercepts for macro-family,

566   sub-family and – crucially – the language itself. These controls are expected to help mitigate

567   some of the issues outlined above. Moreover, in Sect. 4, we discuss the results of several

568   additional analyses presented in the Supporting Information, which further explore how

569   robust our results are with respect to the potential sources of influence described above.

570   Nevertheless, we wish to emphasise that, given the scope of this study – encompassing

571   several hundred languages and a wide variety of writing systems, each with its own distinct

572   and often idiosyncratic conventions for representing language in written form – it is difficult,

573   if not impossible, to completely rule out the possibility that these cross-linguistic variations

574   may systematically influence the validity of our quantitative results.

2.5 Entropy estimation

In order to quantify the amount of information contained in κ, we can represent κ as a distribution of symbol frequencies by counting how often each symbol j appears in κ and call the resulting frequency f_j. The Gibbs-Shannon unigram entropy H of this distribution can be computed as [41]:

$$H(\kappa) = - \sum_{j=1}^{V} p(s_j) \cdot \log p(s_j) \tag{1}$$

where $p(s_j) = \frac{f_j}{\sum_{j=1}^{V} f_j}$ is the maximum likelihood estimator of the probability of s_j in κ consisting of $\sum_{j=1}^{V} f_j$ tokens. In what follows, all logs are to the base two, so the quantities are expressed in bits. *H(κ)* can be interpreted as the average number of (yes/no) guesses that are needed to correctly predict the type of a symbol token that is randomly sampled from κ. The entropy rate or per-symbol entropy of a stochastic process can be formally defined as [27,41]:

$$h(\kappa) = \lim_{N \to \infty} \frac{1}{N} H_N(\kappa) = \lim_{N \to \infty} \frac{1}{N} H(t_1^N) \tag{2}$$

where $t_1^N = t_1, t_2, \dots, t_N$ represents a block of consecutive tokens of length N and $H_N(\kappa)$ denotes the so-called block entropy of block size N [27,110].

Following [43], we define F_N as the *prediction complexity* of t_N given $t_1, t_2, \dots, t_{N-1}$ as follows:

$$F_N \equiv H(t_N | t_1^{N-1}) \tag{3}$$

F_N quantifies the average uncertainty of the Nth symbol, given all preceding tokens t_1^{N-1}. Assuming a stationary ergodic stochastic process [27,41,44], F_N reaches the entropy rate h as N tends to infinity [41,43]:

$$h(\kappa) = \lim_{N \to \infty} F_N \tag{4}$$

26

596   In analogy to $H(\kappa)$, the entropy rate $h(\kappa)$ can be informally understood as the average number

597   of guesses that are needed to guess the next symbol of a sequence and thus incorporating the

598   notion that prediction and understanding are intimately related [14,72,110,111]. Information

599   can then be defined as any kind of knowledge that, when in your possession, allows you to

600   make predictions with greater accuracy than mere chance [112,113]. Thus, h encompasses

601   complexity from various linguistic sub-domains, since any form of linguistic (e.g.

602   grammatical, phonological, lexical, pragmatic) or non-linguistic (e.g. world) knowledge will

603   help a reader or listener to predict upcoming linguistic material more accurately and will

604   therefore reduce h [10]. This implies that when using LMs trained on written text to draw

605   conclusions about languages, there is yet another important caveat that must be emphasised:

606   not all information is encoded in written language. Key aspects such as extra-linguistic

607   information, including prosody – i.e., the supra-segmental features of speech such as pitch,

608   loudness, and tempo – are not fully captured. While Wolf et al. [114] show that much of the

609   prosodic information can be inferred from the words and surrounding context, they also

610   demonstrate that prosodic features cannot be entirely predicted from text alone. This suggests

611   that prosody carries information that extends beyond the written word, contributing additional

612   layers of knowledge.

613   Since the probability distribution for any natural language is unknown [14,110], we use the

614   data compression algorithms described above to estimate $h(\kappa)$. Per LM, the entropy rate

615   estimate is computed (roughly speaking) as the number of bits per symbol in the compressed

616   text:

617
$$\hat{h}^{LM}(\kappa) = \frac{K^{LM}(\kappa)}{L(\kappa)} \tag{5}$$

618   where $K^{LM}(\kappa) = R^{LM}(\kappa) - R^{LM}(\kappa_{train})$. Here, $R^{LM}(\kappa)$ denotes the number of bits that are

619   needed by the LM to compress κ, κ_{train} represents the first half of κ, $L(\kappa)$ represents the

27

620 length of the second half of κ in words on the level of words or, both on the character and the

621 BPE level, in Unicode characters. Note that on the BPE levels we also compress the mapping

622 of unique symbols to 4-byte Unicode symbols mentioned above and add the resulting

623 compressed lengths to $R^{LM}(\kappa)$ and $R^{LM}(\kappa_{train})$. Further note that $\hat{h}^{LM}$ is directly related to

624 the quantity perplexity that is often used in NLP to measure the quality of a language model,

625 where perplexity is defined as $2^{\hat{h}^{LM}}$ [42]. We use this relationship to also choose the LM that

626 achieves the lowest perplexity on the test data:

627
$$\hat{h}^{best}(\kappa) = \min_{LM \in \mathcal{L}} \hat{h}^{LM}(\kappa) \tag{6}$$

628 where $\mathcal{L}$ denotes the set of different LMs, i.e., $\mathcal{L} =$ {PPM2, PPM6, PAQ, LSTM, TRFsmall,

629 TRFmed, TRFbig}. In a similar vein, we choose $\hat{K}^{best}(\kappa)$.

2.6 Statistical analysis

2.6.1 Comparing entropy/length distributions across LMs and corpora

632 We now take κ to be a corpus that consists of individual texts κ_i, where i denotes $1,\ldots, I$

633 different languages. For brevity, we omit the superscript and the hat in what follows.

634 However, we compute the correlation coefficients described below for both the best LM

635 ($\hat{h}^{best}$ and $\hat{K}^{best}$) and for each individual LM ($\hat{h}^{LM}$ and $\hat{K}^{LM}$).

636 The entropy estimate for κ_i is denoted as $h_\varsigma(\kappa_i)$, where ς denotes one of three information

637 encoding units (words, characters, BPE), likewise for $K_\varsigma(\kappa_i)$ and $L_\varsigma(\kappa_i)$. $h_\varsigma(\kappa_i)$ and $K_\varsigma(\kappa)$

638 are computed on all three levels, while $L_\varsigma(\kappa)$ is computed on either the word or the character

639 level. Note that for languages with more than one available translation/document in a corpus,

640 all quantities are averaged.

641 To evaluate the (dis-)similarity of entropy/length distributions across corpora and test for a

642 potential trade-off between entropy and length, we first compute the pairwise Pearson

643   correlation $\rho\left[v'_\zeta(\kappa), v''_\zeta(\iota)\right]$ for all corpus pairs (κ, ι) where v'_ζ denotes either h, L or K on the

644   encoding level ζ. Likewise for v''_ζ. In addition, we also consider H computed on the basis of

645   the Crúbadán word frequency information (eq. 1), denoted as H_{Cr} in what follows. Both $v'_\zeta(\kappa)$

646   and $v''_\zeta(\iota)$ are logged. Pearson correlations range from -1 to 1, with higher absolute values

647   indicating stronger associations. Positive values reflect positive associations, while negative

648   values reflect negative associations between distributions. This allows us to assess both

649   similarities and dissimilarities between entropy and length distributions across corpora.

650   To control for potential sources of influence, we fit linear models of the form:

651 $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{7}$$

652   where $\mathbf{y}$ is the $n \times 1$ vector of observed values, either $v'_\zeta(\kappa)$ or $v''_\zeta(\iota)$, n denotes the number of

653   languages that are available in both κ and ι, $\mathbf{X}$ is the $n \times p$ design matrix of p covariates

654   including a $n \times 1$ vector of ones for the intercept, $\boldsymbol{\beta}$ is the corresponding $p \times 1$ vector of

655   coefficients and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of residuals. We assume that ϵ_i are identically distributed

656   with $E(\epsilon_i) = 0$ with variance σ^2. Importantly, we want to rule out potential autocorrelation

657   among the residuals, i.e., we wish to test the following null hypothesis in what follows:

658 $$H_0: E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}] = \sigma^2 \mathbf{I}. \tag{8}$$

659   where $\mathbf{I}$ is the $n \times n$ identity matrix and T denotes the matrix transpose.

660   As potential control variables, we consider the EGIDS level, the (logged) speaker population

661   size, the (logged) number of corpora and the (logged) number of countries in which the

662   language is spoken. We also include a set of indicator variables for the levels (categories) of

663   writing script. To avoid overfitting, scripts that were unique to a single language were

664   grouped into a common category. For H_{Cr}, we additionally control for the number of words

665   and the number of documents (both logged) on which the word frequency list is based, and a

29

666   binary variable indicating whether the word frequency list is truncated to account for

667   differences in the way different Crúbadán word lists were generated. Further information can

668   be found in [37]. To select the relevant control variables from the candidate set, we use the

669   lasso machine learning technique [115]. To choose the optimal value for the penalty

670   parameter for each lasso, we use a heteroskedastic plugin estimator [116]. Languages with

671   missing information on any of the control variables are excluded in each case. Let $\widetilde{\mathbf{X}}$ denote a

672   $n \times \tilde{p}$ matrix of $\tilde{p}$ covariates selected by the lasso. We then regress $\mathbf{y}$ on $\widetilde{\mathbf{X}}$ and compute

673   residuals denoted as $\tilde{\boldsymbol{\epsilon}}$. To test H_0 from above (eq. 8), we compute the modified version of

674   Moran's I [117] suggested by Kelejian and Prucha [118], written as:

$$I = n(\tilde{\boldsymbol{\epsilon}}^{\mathrm{T}} \mathbf{W} \tilde{\boldsymbol{\epsilon}}) \left[(\tilde{\boldsymbol{\epsilon}}^{\mathrm{T}} \tilde{\boldsymbol{\epsilon}}) \sqrt{\mathrm{tr}\{(\mathbf{W}^{\mathrm{T}} + \mathbf{W})\mathbf{W}\}} \right]^{-1} \tag{9}$$

676   where $\mathbf{W}$ denotes an $n \times n$ weighting matrix and tr represents the trace operator. For $\mathbf{W}$, we

677   consider two inverse distance matrices: (i) to test for spatial autocorrelation, we construct an

678   inverse distance matrix $\mathbf{W}_{\mathrm{G}}$ based on longitude and latitude information, (ii) to test for

679   phylogenetic autocorrelation, we construct an inverse distance matrix $\mathbf{W}_{\mathrm{P}}$ based on a

680   phylogenetic similarity matrix provided by [65]. In both cases, matrix elements are equal to

681   the reciprocal of distance that are then normalised using spectral normalisation. We test for

682   autocorrelation with (i) $\mathbf{W}_{\mathrm{G}}$ as input, (ii) $\mathbf{W}_{\mathrm{P}}$ as input and (iii), as explained below, both $\mathbf{W}_{\mathrm{G}}$

683   and $\mathbf{W}_{\mathrm{P}}$ as input. For brevity, we drop the subscript in what follows and describe our

684   algorithmic approach for input matrix $\mathbf{W}$. Since $I^2 \sim X^2(1)$ [118], we test H_0 via a standard X^2-

685   test with one degree of freedom. If $p < 0.05$, we extend our linear regression model (eq. 7) by

686   a semiparametric filter [119–121] as:

$$\mathbf{y} = \mathbf{F}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{10}$$

688   where, in addition to the above, $\mathbf{F}$ is a $n \times q$ matrix of q eigenvectors and $\boldsymbol{\gamma}$ is a $n \times 1$ vector of

689   parameters. $\mathbf{F}$ is computed based on a transformed version of $\mathbf{W}$ defined as [119]:

30

$$690 \qquad \mathbf{M} \equiv \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{n}\right)\mathbf{W}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{n}\right) \qquad (11)$$

691   where $\mathbf{1}$ represents a $n \times 1$ column vector of ones. The eigensystem decomposition of $\mathbf{M}$

692   generates n eigenvalues and n corresponding eigenvectors. The eigenvalues are then sorted in

693   descending order, denoted as $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \dots \lambda_n)$, so that the largest eigenvalue receives the

694   subscript 1, the second largest eigenvalue receives the subscript 2 and so on. The

695   corresponding set of eigenvectors can then be denoted as $\mathbf{E} = (\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \dots \mathbf{E}_n)$. We include

696     $\mathbf{E}_1$ into $\mathbf{F}$ and let the lasso select a subset of control variables from $\mathbf{X}$. We then compute $\tilde{\boldsymbol{\epsilon}}$

697   based on a regression of $\mathbf{y}$ on $\widetilde{\mathbf{X}}$ and $\mathbf{F}$. After that, we perform the X^2-test again. If the test is

698   still significant, we also include $\mathbf{E}_2$ into $\mathbf{F}$ and re-perform estimation. This iterative procedure

699   is repeated until $p \geq .05$. [Note that in scenario (iii), where we simultaneously control for $\mathbf{W}_G$

700   and $\mathbf{W}_P$, we compute two sets of eigenvectors, denoted as $\mathbf{E}_G = (\mathbf{E}_{G,1}, \mathbf{E}_{G,2}, \mathbf{E}_{G,3}, \dots \mathbf{E}_{G,n})$, and

701     $\mathbf{E}_P = (\mathbf{E}_{P,1}, \mathbf{E}_{P,2}, \mathbf{E}_{P,3}, \dots \mathbf{E}_{P,n})$, and alternate the inclusion of eigenvectors into $\mathbf{F}$, i.e., we first

702   include $\mathbf{E}_{G,1}$, then $\mathbf{E}_{P,1}$, and so on. Correspondingly, $I^2 \sim X^2(2)$]. Let $\tilde{\epsilon}_{v_E'}(\kappa)$ denote the

703   resulting residuals for $v_\zeta'(\kappa)$, likewise for $v_\zeta''(\iota)$. This procedure ensures that there is no

704   spatial/phylogenetic correlation among the resulting residuals, i.e., $\tilde{\epsilon}_{v_\zeta'}(\kappa)$ and $\tilde{\epsilon}_{v_\zeta''}(\iota)$. We

705   then compute the Pearson correlation between those residuals and proceed as described above.

706   The correlation coefficients per condition (none, geographical, phylogenetic and both) are

707   denoted as ρ_{none}, ρ_{geo}, ρ_{phylo} and ρ_{both}.

708   **2.6.2   Multi-model multilevel inference**

709   To evaluate if the trade-off is moderated by the social environment in which languages are

710   being used, we run separate multilevel effects models (MLEMs) with (i) h or K as the

711   outcome on all three levels (words/characters/BPE) for all eight LMs (best, PPM2, PPM6,

712   PAQ, LSTM, TRFsmall, TRFmed, TRFbig) and (ii) L as the outcome for words/characters.

713   For $N = 3{,}705$ individual documents, we fit MLEMs of the form [122]:

    31

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\epsilon} \qquad (12)$$

where, in addition to the above, $\mathbf{Z}$ is a matrix of random predictors and $\mathbf{u}$ is a vector of random effects that are assumed to follow a normal distribution, with mean 0 and variance-covariance matrix $\mathbf{G}$. The residual errors $\boldsymbol{\epsilon}$ are assumed to follow a normal distribution, with mean 0 and variance matrix $\sigma^2\mathbf{I}$; $\mathbf{u} \perp \boldsymbol{\epsilon}$. To enhance convergence, the outcome is standardised 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, but we also provide results for log-transformed outcomes (see https://osf.io/93csg/) that can be visualised in our interactive online application (https://www.owid.de/plus/tradeoffvis/). We consider a fixed effect for the estimated speaker population size (logged) as a proxy for population structure [123]. The following control variables are included: (i) fixed effects: corpus type (parallel/comparable), binary indicators for the first four EGIDS levels and the (logged) number of countries; (ii) random intercepts for the following groups: writing script, corpus, macro-area, macro-family, sub-family and language. We cross corpus, macro-area, macro-family and writing script and explicitly nest language within sub-family within macro-family; (iii) random slopes for population size, i.e., we allow the effect of population size to vary across the different groups.

We adopt a multi-model inference approach [124] by sub-setting each full model, i.e., we generate a set of models with all possible control variable subsets, which are then fitted to the data. We fit sub-models per outcome, type and LM. All models were fitted with gradient-based maximization (maximal number of 20 iterations) and via maximum likelihood (ML). Per outcome and per type, we then compute a frequentist model averaging (FMA) estimator over all R candidate models [125,124,126]:

$$\tilde{\beta}_x = \sum_{j=1}^{R} \omega_j \beta_{x,j} \qquad (13)$$

32

737   where $\beta_{x,j}$ denotes the estimated fixed effect of variable x for model j and ω_j is a weight

738   computed as:

$$\omega_j = \frac{e^{\left(-\frac{1}{2}\Delta_j\right)}}{\Omega} \tag{14}$$

739

740   where $\Omega = \sum_{r=1}^{R} e^{\left(-\frac{1}{2}\Delta_r\right)}$ represents the sum of weights for all R models. To compute Δ_j, we

741   use Akaike's information criterion (AIC) [127], where lower values indicate a better model,

742      $\Delta_j = \text{AIC}_j - \text{AIC}_{min}$ where AIC_j denotes the AIC value computed for model j and AIC_{min}

743   represents the minimum AIC value over all R models. Note that in models where x does not

744   appear, $\beta_{x,j} \equiv 0$. On this basis, we compute an FMA estimator of the standard error (SE) as

745   [124]:

$$\text{SE}(\tilde{\beta}_x) = \sum_{j=1}^{R} \omega_j \sqrt{\text{SE}(\beta_{x,j})^2 + (\beta_{x,j} - \tilde{\beta}_x)^2} \tag{15}$$

746

747   where $\text{SE}(\beta_{x,j})$ denotes the estimated standard error of $\beta_{x,j}$ for model j. In models where x

748   does not appear, we set $\text{SE}(\beta_{x,j}) \equiv 0$. To assess statistical significance, we compute a

749   corresponding two-tailed p-value as $p = 2 \cdot \left(1 - \Phi\left(\left|\frac{\tilde{\beta}_x}{\text{SE}(\tilde{\beta}_x)}\right|\right)\right)$ where $\Phi()$ denotes the

750   cumulative standard normal distribution. In a similar vein, we compute a 95% confidence

751   interval (95%-CI) as $\tilde{\beta}_x \pm \Phi^{-1}(0.975) \cdot \text{SE}(\tilde{\beta}_x)$ where $\Phi^{-1}()$ denotes the inverse cumulative

752   standard normal distribution.

753   Note that the Akaike weights ω_j can be "interpreted as approximate probabilities of each

754   model being the actual best model, given the data" [124]. Thus, we can use the ω_j to estimate

755   the relative importance of variable x, computed as [124]:

$$\sigma_x = \sum_{j=1}^{R} \omega_j c_{x,j} \tag{16}$$

756

33

757   where $c_{x,j}$ is a binary indicator that is equal to 1 if x is explicitly in model j and 0 otherwise

758   [124]. The larger σ_x, the more important x. To put the value of σ_x into perspective, we show

759   in Supporting Information S2 that its theoretical minimum is ~0.27.

3 Results

3.1 Comparing entropy/length distributions across LMs and corpora

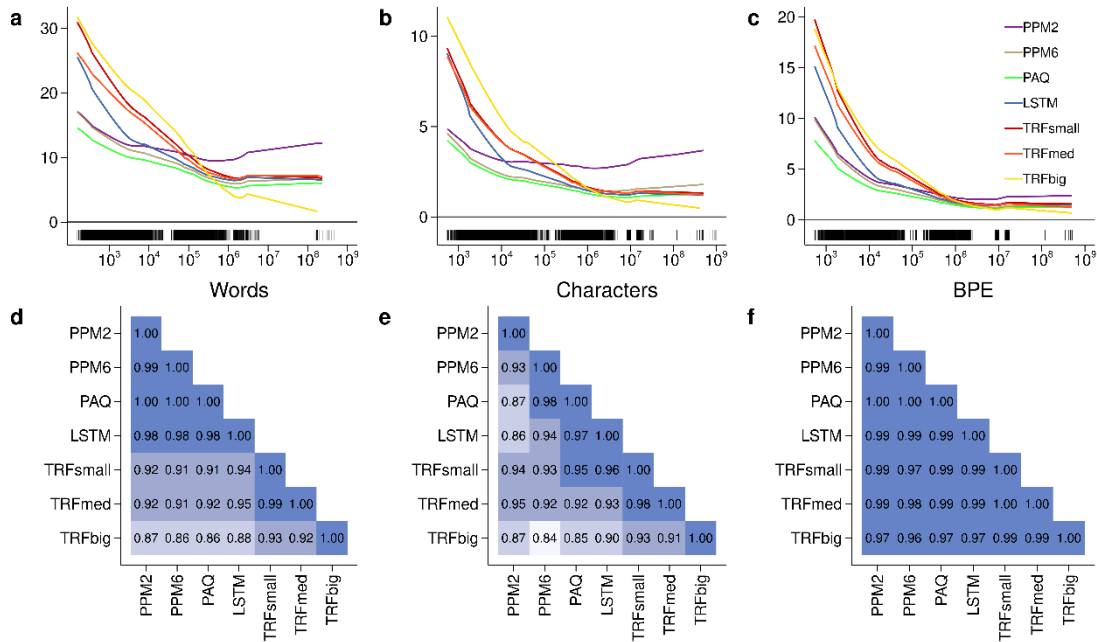#### 3.1.1 Comparing language models

763

Figure 2: Comparing language models. (**a** – **c**) LM-specific entropy rates as a function of text length for different symbolic levels (words, characters, BPE). Each solid line represents a locally weighted scatterplot smoother (bandwidth = 0.3) for the entropy estimates of N_κ = 4,297 different documents that belong to the compiled multilingual database (see Sect. 2.1 for details, rug plots at the bottom of each graph illustrate the length distribution). Note that on the BPE level, $\widehat{h}^{LM}$ is plotted against L in characters. (**d** – **f**) Median unadjusted Pearson correlation, $\widetilde{\rho}_{\mathbf{none}}$, across LMs. These values are calculated by first cross-correlating average entropy rates per language among LMs for each of the 40 full text corpora, followed by computing the median value for each LM pair. Interactive visualisations are available at https://www.owid.de/plus/tradeoffvis/.

772

773   Figure 2a-c summarises the distribution of h as a function of length L for each level and each

774   investigated LM across the 40 full-text multilingual text collections/corpora, totalling N_κ =

775   4,297 different documents (see Sect. 2.1 for details). The plots indicate that for most

34

776 documents PAQ (mint line) turns out to be the best LM, i.e., $\hat{h}^{best}(\kappa) = \hat{h}^{PAQ}(\kappa)$ (see Sect.

777 2.5 for details) for most of the documents. For longer documents, the larger LMs (LSTM and

778 the three Transformer LMs) achieve similar or lower entropy rates. Correspondingly, Table 3

779 shows that PAQ has the lowest h in more than 90% of the documents across all three

780 symbolic levels.

781 **Table 3: Best LM per level.** 1st column: LM. 2nd column: word level. 3rd column: character
782 level. 4th column: BPE level. For each of N_κ = 4,297 different documents, each column lists, for a given
783 symbolic level, the number of documents such that the LM in the given row is the best, i.e., $\hat{h}^{best} = \hat{h}^{LM}$. On all
784 three levels, PAQ is the best LM in more than 90%.

LM	Words	Characters	BPE
PPM2	0 (0.00%)	0 (0.00%)	0 (0.00%)
PPM6	18 (0.42%)	261 (6.07%)	85 (1.98%)
PAQ	4,241 (98.70%)	3,960 (92.16%)	4,182 (97.32%)
LSTM	1 (0.02%)	3 (0.07%)	0 (0.00%)
TRFsmall	0 (0.00%)	0 (0.00%)	0 (0.00%)
TRFmed	3 (0.07%)	0 (0.00%)	1 (0.02%)
TRFbig	34 (0.79%)	73 (1.70%)	29 (0.67%)

785

786 To determine if entropy rate distributions are systematically affected by the choice of LM, we

787 used the entropy estimates for the 40 full text corpora at each symbolic level to compute

788 pairwise correlations ρ_{none} for each pair of LMs.

789 Figure 2d-f presents the resulting pairwise relationships as correlation matrices. Each cell

790 represents the median value of ρ_{none} for a pair of LMs. At each symbolic level, the statistical

791 association between different LMs is remarkably strong. Even the lowest median value,

792 observed at the character level between TRFbig and PPM6, is high, with a value of $\tilde{\rho}_{none} =$

793 0.84. To rule out the possibility that these associations mainly result from either language-

794 and document-specific characteristics or the genealogical and geographic relatedness of

795 languages (see Sect. 2.6.1 for details), Figure 3 visualises all four types of estimated

796 correlation coefficients, i.e., ρ_{none}, and corresponding adjusted partial correlations, i.e., ρ_{geo},

35

797 ρ_{phylo} and ρ_{both} for all LM pairs across corpora and symbolic levels. The results indicate that

798 the adjusted partial correlations point in the same direction as the unadjusted correlations.

799 The results presented in this section demonstrate that although there are differences in the

800 performance of various LMs depending on document length (Figure 2a-c), the resulting

801 entropy rate distributions are remarkably consistent across LMs (Figure 3). This suggests that,

802 from a cross-linguistic perspective, the choice of LM for investigating different languages

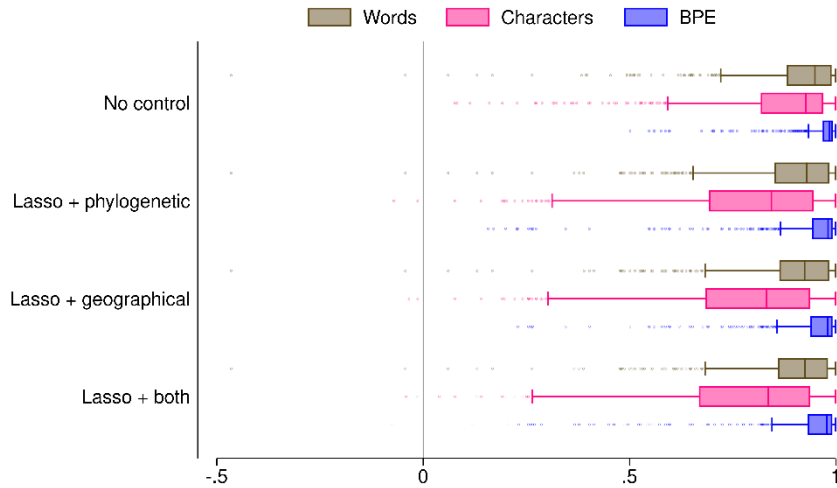803 may have a minimal impact.

804

805

806 **Figure 3: Distribution of pairwise correlations across LMs for each symbolic level.** For each of the 40 full
807 text corpora and per symbolic level, we compute the median value of the pairwise correlation between the
808 estimated entropy rate distributions for each LM pair. We compute both unadjusted correlations, i.e., ρ_{none}, and
809 adjusted partial correlations, i.e., ρ_{geo}, ρ_{phylo} and ρ_{both} (see Sect. 2.6.1 2.1 for details) across LM pairs. Per type of
810 correlation coefficient and symbolic level, we compute $\boldsymbol{N_\rho} = 840$ individual correlation coefficients where each
811 data point represents one LM pair.
812

3.1.2 Comparing languages

814 We proceed by comparing entropy rate distributions across corpora to determine whether a

815 language that tends to be more complex in one corpus also tends to be more complex in

816 another corpus. Given that the results from the previous section clearly indicate stability

817 across different LMs, we will focus, for each corpus κ, on the estimates of its best model, i.e.,

36

818 $\hat{h}^{best}(\kappa)$ (see Sect. 2.5 for details). Interactive visualisations for each LM, i.e., $\hat{h}^{LM}(\kappa)$. are

819 available at https://www.owid.de/plus/tradeoffvis/.
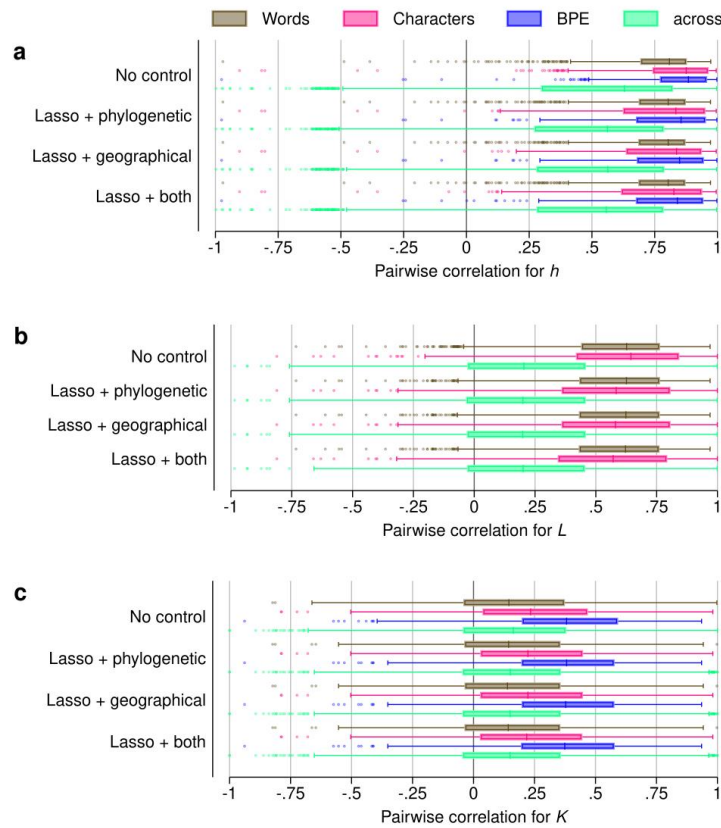
Figure 4: Distribution of pairwise correlations across corpora and symbolic levels. For each corpus pair, we compute both unadjusted correlations, i.e., ρ_{none}, and adjusted partial correlations, i.e., ρ_{geo}, ρ_{phylo} and ρ_{both} (see Sect. 2.6.1 for details). Correlations are computed both per symbolic level, where estimates are derived on the same symbolic level, and across symbolic levels, where, e.g., the entropy rate distribution calculated for words as information encoding units in one corpus is correlated with, e.g., the distribution calculated at either the character or the BPE level in another corpus. (**a**) Similarity of $\hat{h}^{best}$-distributions across corpora, including correlations between $\hat{h}^{best}$ for the 40 full text corpora and H_{Cr} based on the Crúbadán word lists ($N_\rho = 7{,}254$). (**b**) Similarity of L-distributions across corpora ($N_\rho = 3{,}160$). (**c**) Similarity of $\hat{K}^{best}$-distributions across corpora ($N_\rho = 7{,}140$). Interactive LM-specific visualisations are available at https://www.owid.de/plus/tradeoffvis/.

831 As outlined above, we evaluate the similarity of $\hat{h}^{best}$-distributions by computing ρ_{none}, ρ_{geo},

832 ρ_{phylo} and ρ_{both} *across* corpora and *across* symbolic levels. Per correlation type, we compute

833 $N_\rho = 7{,}254$ individual correlations. Figure 4a demonstrates that entropy rate distributions are

834 very similar across corpora and symbolic levels as indicated by a strong positive correlation

835 between corpus pairs. The results remain stable when we control for language- and document-

37

836 specific characteristics, as well as the genealogical and geographic relatedness of languages

837 (see Sect. 2.6.1 for details). Interactive LM-specific visualisations, available at

838 https://www.owid.de/plus/tradeoffvis/ show that highly comparable patterns are obtained

839 when estimating entropy rates for individual LMs.

840 Entropy rates are estimated as the ratio of the number of bits needed to compress the test data,

841 K, to the length of the test data, i.e., L (cf. eq. 5). To further understand the above results, we

842 repeated the analyses for both variables that are part of this ratio. Figure 4b reveals that while

843 the results are more pronounced for h, the distributions for L are also very comparable across

844 corpora and symbolic levels ($N_\rho = 3{,}160$). However, Figure 4c shows that the results are

845 much weaker for distributions of $\widehat{K}^{best}$ ($N_\rho = 7{,}140$). Since K is the product of h and L, this

846 suggests a potential trade-off between h and L.

847 **Table 4: Association between $\widehat{h}^{best}$ and L across corpora.** 1st row: number of individual correlation
848 coefficients, N_ρ. 2nd – 4th column, type of correlation coefficient ρ_{none}, ρ_{phylo}, ρ_{geo} and ρ_{both}. 2nd – 3rd column:
849 associations on each level (words, characters, BPE), listed quantities for ρ are median values per parameter
850 combination. Median values for $\widehat{h}^{best}$(words) include H_{Cr} based on the Crúbadán word lists. Note that for $\widehat{h}^{best}$
851 on the BPE level, distributions are correlated with L in characters.

	$\widehat{h}^{best}$(words)	$\widehat{h}^{best}$(chars)	$\widehat{h}^{best}$(BPE)
N_ρ	1,638	1,600	1,600
$\tilde{\rho}_{none}$	-0.69	-0.69	-0.69
$\tilde{\rho}_{phylo}$	-0.69	-0.62	-0.62
$\tilde{\rho}_{geo}$	-0.69	-0.62	-0.62
$\tilde{\rho}_{both}$	-0.69	-0.60	-0.61

852

853 To investigate this possibility, we compute ρ_{none}, ρ_{geo}, ρ_{phylo} and ρ_{both} between $\widehat{h}^{best}$ and L

854 across corpora on all three symbolic levels. Table 4 demonstrates that on all three symbolic

855 levels and for all four correlation types, there is a pronounced negative statistical association

856 between entropy rate and length distribution. Again, our interactive visualisation tool shows

857 highly comparable patterns for individual LMs. These results indicate the existence of a trade-

858 off between both variables.

38
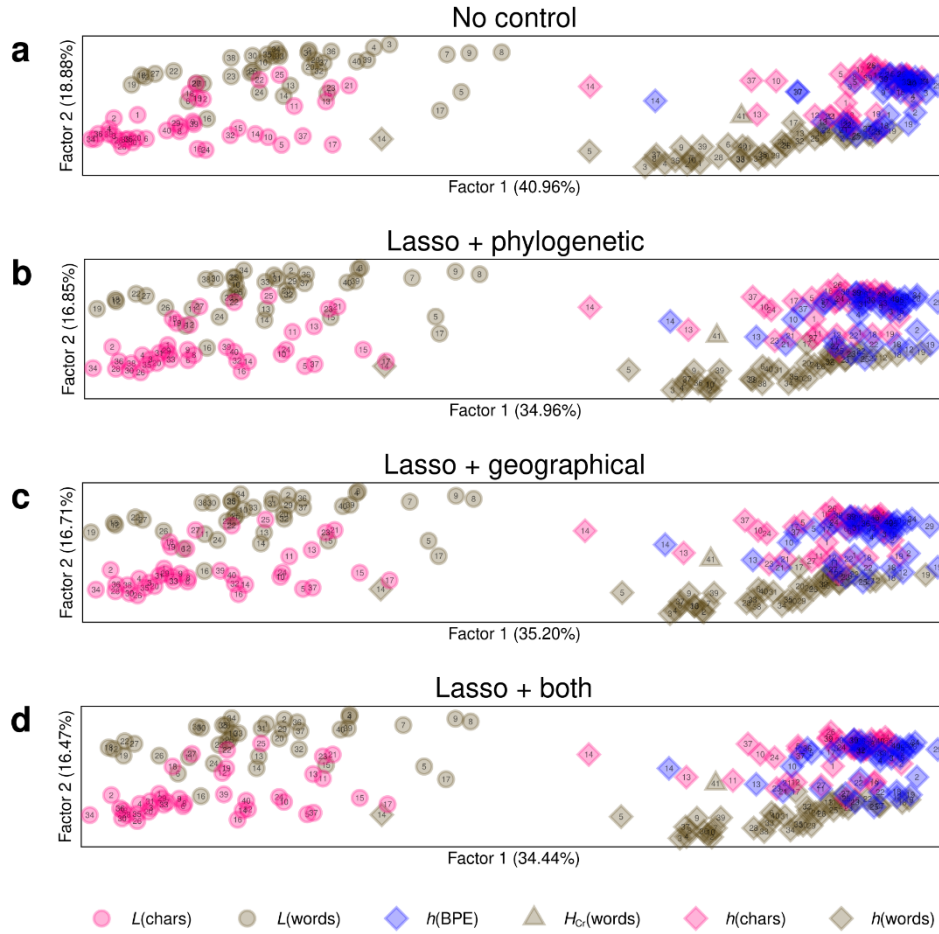
Figure 5: Comparing entropy and length across corpora. For each corpus pair and across symbolic levels, we compute both unadjusted correlations (ρ_{none}, **a**), and adjusted partial correlations (ρ_{phylo}, **b**; ρ_{geo}, **c**; ρ_{both}, **d**), see Sect. 2.6.1 for details. For each correlation type, $N_\rho = 20{,}100$ individual correlations are computed to generate a corresponding correlation matrix. Principal-component factoring reveals that for both unadjusted and adjusted correlations, more than ~50% of the variance in the matrix can be attributed to two factors: one main factor representing the strong negative correlation between length and entropy measures (accounting for 34.44% to 40.96% of the variance), and one factor distinguishing symbol types (accounting for 16.47% to 18.88% of the variance). Each marker label represents a numeric ID for one of the 41 investigated corpora (see Supporting Information S3). Interactive LM-specific visualisations are available at https://www.owid.de/plus/tradeoffvis/.

To establish if the trade-off between entropy and length holds across symbolic levels, we

compute ρ_{none}, ρ_{geo}, ρ_{phylo} and ρ_{both} between $\hat{h}^{best}$ and L for all corpus pairs across all three

symbolic levels. For each correlation type, $N_\rho = 20{,}100$ individual correlations were

calculated to generate a correlation matrix, which was then subjected to principal component

analysis. Figure 5 presents scatterplots of the first two factors that explain most of the

variance in the matrix. For both unadjusted and adjusted partial correlations, more than a third

39

876 of the variance is attributed to the trade-off between entropy and length in each case.

877 Interactive LM-specific visualisations that are available at

878 https://www.owid.de/plus/tradeoffvis/ illustrate that highly comparable patterns are observed

879 when estimating entropy rates for different LMs.

880 Given that, from an information theoretic point of view, message length quantifies efficiency

881 – the shorter the message the higher the efficiency [15] – we arrive at our main empirical

882 result: human languages trade off complexity against efficiency. More explicitly, a higher

883 average amount of choice/uncertainty per produced/received symbol is compensated by a

884 shorter average message length.

885 ### 3.1.3 Summary

886 In the last two sub-sections, we demonstrated that (i) entropy rate distributions are highly

887 consistent across different LMs, suggesting that the choice of LM might have minimal impact

888 on cross-linguistic investigations of the kind we presented here. We also showed that (ii) there

889 is a pronounced trade-off between entropy rate and document length across different corpora,

890 which implies that languages balance efficiency and complexity. This finding highlights a

891 potentially fundamental principle in linguistic structure, where higher uncertainty per symbol

892 is offset by shorter message lengths.

893 To bring these results together, we now compute fully adjusted partial correlations, ρ_{both} for

894 all possible pairwise combinations of the 41 corpora, the three variables ($\hat{h}^{LM}$, H_{Cr}, L), the

895 three symbolic levels (words, characters, BPE), and the seven investigated LMs (PPM2,

896 PPM6, PAQ, LSTM, TRFsmall, TRFmed, TRFbig). In total, $N_\rho = 423,660$ individual

897 correlation coefficients were computed. Figure 6 shows that the findings point in the same

898 direction as previously observed, confirming the consistency of entropy rate distributions and

899 their trade-off with document length.

40

900   To emphasise this point, we extracted the first 80 factors from the factor analysis, which

901   together account for ~90% of the variance in the correlation matrix. For each factor, we

902   conducted separate linear regressions with the factor as the outcome and a binary indicator for

903   the type of variable $(1 = L$ vs. $0 = \hat{h}^{LM}$ or $H_{\mathrm{Cr}})$ as the predictor. For each factor, we then

904   extracted the amount of explained variance (R^2) as a measure of model fit. Among all factors,

905   R^2 is highest for the first factor, with $R^2 = 76.14\%$. We then repeated the analyses using

906   indicator variables for the investigated LMs as predictors. Again, R^2 is highest for the first

907   factor, but with a much smaller value of $R^2 = 6.44\%$). Figure 6 demonstrates that the first

908   factor distinguishes between length and entropic variables but not between LMs. We further

909   visually inspected all remaining factors, none of which separated the LMs, reinforcing the

910   robustness of our findings across different models.

911   These results underscore that the negative statistical association between entropy and length

912   in human languages is consistent across various LMs and corpora, suggesting that the trade-

913   off between complexity and efficiency may reflect a fundamental property of human

914   language.

41
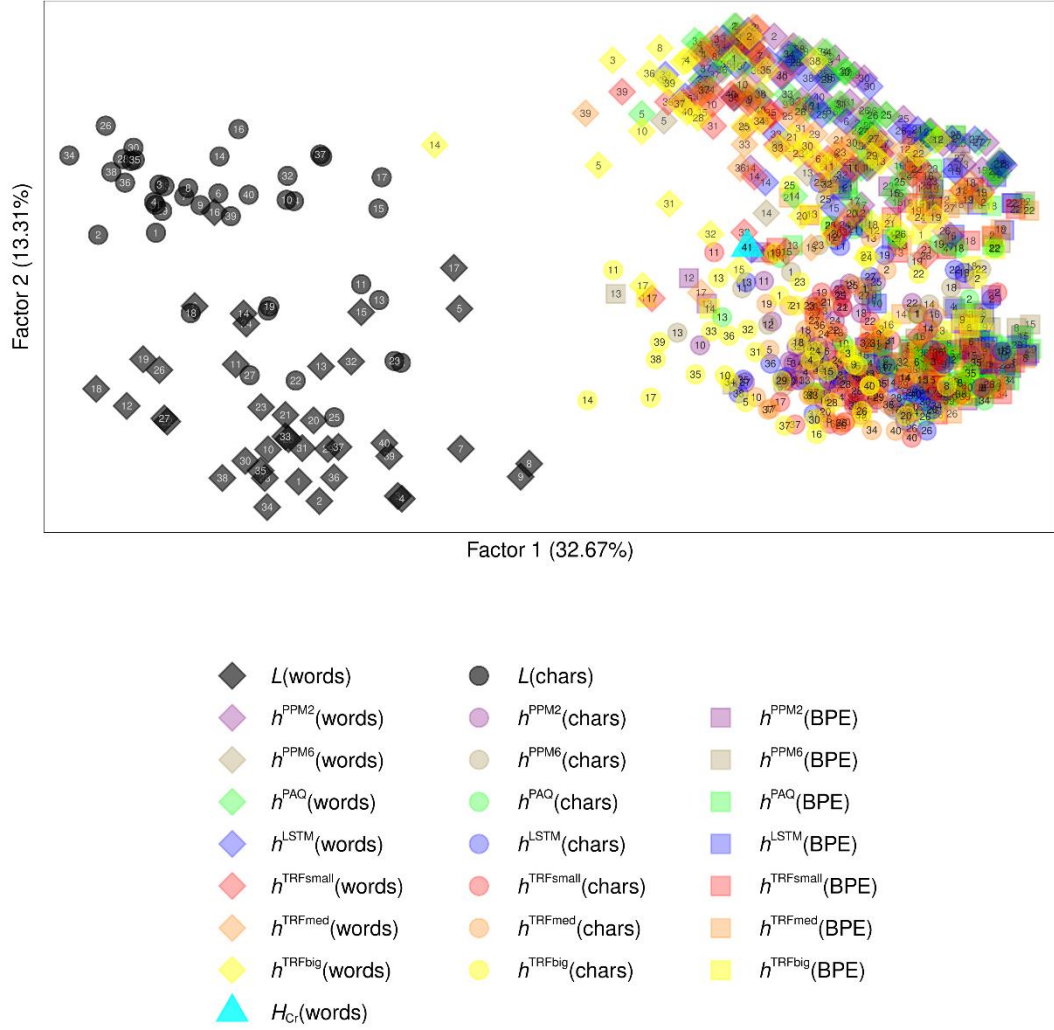
Figure 6: Evidence for a trade-off between complexity and efficiency across corpora and language models. We compute adjusted partial correlations, ρ_{both} (see Sect. 2.6.1 for details), for each combination of the 41 corpora, the three variables $(\hat{h}^{LM}, H_{Cr}, L)$, the three symbolic levels (words, characters, BPE), and the seven investigated LMs (PPM2, PPM6, PAQ, LSTM, TRFsmall, TRFmed, TRFbig), totalling $N_\rho = 423,660$ individual correlations. The resulting correlation matrix is then analyzed with principal-component factoring. The scatterplot demonstrates that (i) different LMs are very similar and (ii) the most important factor, accounting for roughly a third of the variance in the matrix, represents a trade-off between complexity and efficiency: languages that tend to have a higher entropic value tend to need fewer symbols to encode messages. Each marker label represents a numeric ID for one of the 41 investigated corpora (see Supporting Information S3).

42

3.2 Multi-model multilevel inference

3.2.1 Random effects models

To evaluate if the trade-off between complexity and efficiency is influenced by the social environment in which languages are used, we adopt the multi-model inference approach outlined in Sect. 2.6.2.
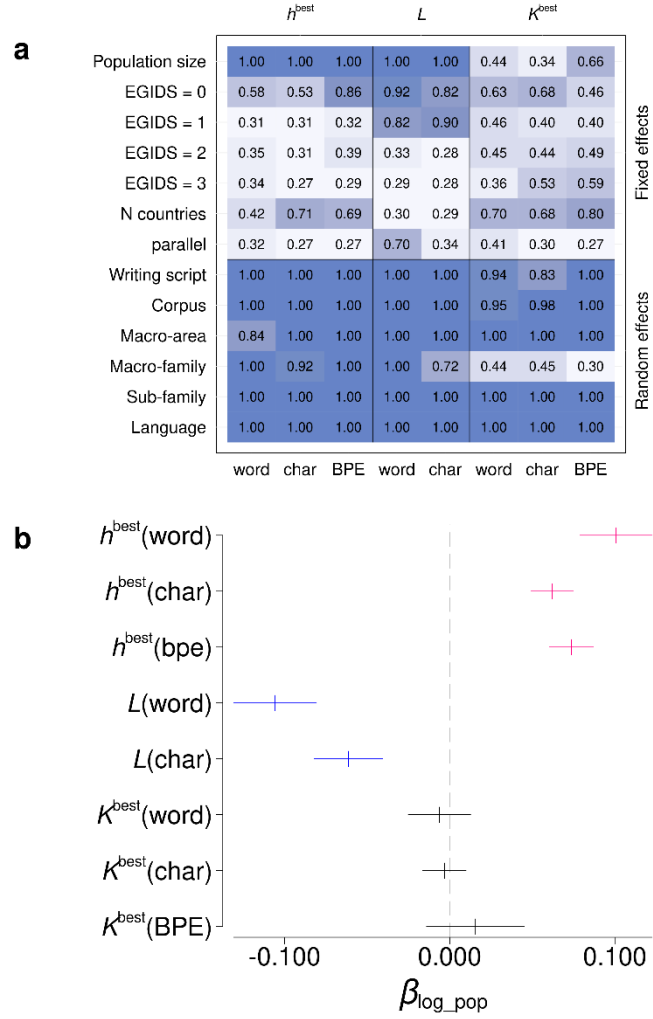

Figure 7: Estimated variable importance and FMA-estimates by outcome and symbolic level ('word' – words, 'char' – characters, 'BPE' – byte pair encoding). For each parameter combination (outcome/level), $R = 8{,}192$ candidate MLEMs that include fixed and random effects were run. (**a**) Estimated variable importance ($\hat{\sigma}_x$) per variable. Higher values indicate greater importance (cf. Sect. 2.6.2 for details), $\hat{\sigma}_x$-values range from 0 (white) to 1 (blue). (**b**) FMA-estimated effect of speaker population size on each outcome per symbolic level. Vertical lines represent the FMA-estimate, $\tilde{\beta}_x$, of population size, here denoted as $\beta_{\log\_pop}$, on $\hat{h}^{best}$, L or $\hat{K}^{best}$. Horizontal lines show corresponding 95%-CIs (cf. Sect. 2.6.2 for details). Lines are coloured in black if the

43

939   95%-CI crosses zero (vertical dashed grey line), whereas blue and pink indicate significant negative and positive
940   effects. Interactive LM-specific visualisations are available at https://www.owid.de/plus/tradeoffvis/.

941

942   By including several fixed and random effects, we control for both (i) language- and

943   document-specific characteristics and (ii) spatial and phylogenetic autocorrelation. For each

944   outcome and symbolic level, we fit $R = 8{,}192$ candidate models. Figure 7a visualises the

945   estimated relative importance, σ_x, for each variable. With the exception of macro-family for

946    $\widehat{K}^{best}$, all considered random effects have high relative importance for all three outcomes and

947   across all three levels. Conversely, there is only weak evidence for the importance of most

948   fixed effects, with the clear exception of speaker population size, which is maximally

949   important – relative to the other variables – for both $\hat{h}^{best}$ and L, but not for $\widehat{K}^{best}$.

950   To investigate if there is also evidence for a trade-off between entropy and length similar to

951   the results presented in the previous section, Figure 7b plots the FMA-estimated effect, $\tilde{\beta}_x$, of

952   speaker population size on each outcome per symbolic level. There is a positive and

953   significant effect of population size on $\hat{h}^{best}$ across all three symbolic levels and a negative

954   significant effect on L for both characters and BPE. For $\widehat{K}^{best}$, there is no significant evidence

955   of an effect of population size on any of the three symbolic levels. Table 5 lists the

956   corresponding estimates for all investigated fixed effects, showing that the only consistent

957   evidence of a noteworthy effect is for speaker population size on either entropy or length.

958   These results substantiate the evidence of a trade-off between entropy and length and indicate

959   that languages with more speakers tend to have higher entropic values, i.e., are more complex,

960   but also tend to produce shorter messages, i.e., are more efficient.

961

962

963

44

	$\widehat{h}^{best}$			*L*		$\widehat{K}^{best}$		
	word	**char**	**BPE**	**word**	**char**	**word**	**char**	**BPE**
Population size	**β = 0.101** **p < .001**	**β = 0.062** **p < .001**	**β = 0.074** **p < .001**	**β = -0.106** **p < .001**	**β = -0.061** **p < .001**	β = -0.006 p = 0.524	β = -0.003 p = 0.638	β = 0.016 p = 0.305
EGIDS = 0	β = -0.239 p = 0.394	β = -0.085 p = 0.436	β = -0.251 p = 0.086	**β = 0.630** **p < .05**	β = 0.356 p = 0.130	β = 0.225 p = 0.343	β = 0.267 p = 0.295	β = 0.147 p = 0.499
EGIDS = 1	β = -0.015 p = 0.706	β = 0.008 p = 0.709	β = 0.009 p = 0.702	β = 0.169 p = 0.134	β = 0.167 p = 0.051	β = 0.048 p = 0.508	β = 0.034 p = 0.575	β = 0.038 p = 0.567
EGIDS = 2	β = -0.029 p = 0.618	β = -0.010 p = 0.680	β = -0.021 p = 0.571	β = 0.025 p = 0.656	β = 0.004 p = 0.897	β = -0.053 p = 0.508	β = -0.050 p = 0.521	β = -0.067 p = 0.476
EGIDS = 3	β = -0.023 p = 0.632	β = -0.001 p = 0.954	β = -0.007 p = 0.740	β = 0.012 p = 0.746	β = -0.001 p = 0.948	β = -0.028 p = 0.605	β = -0.070 p = 0.439	β = -0.091 p = 0.380
N countries	β = -0.024 p = 0.539	β = -0.037 p = 0.249	β = -0.038 p = 0.280	β = -0.006 p = 0.784	β = 0.002 p = 0.912	β = -0.065 p = 0.272	β = -0.063 p = 0.287	β = -0.091 p = 0.156
parallel	β = 0.017 p = 0.657	β = -0.008 p = 0.843	β = -0.001 p = 0.987	β = -0.090 p = 0.253	β = -0.026 p = 0.618	β = -0.022 p = 0.544	β = -0.009 p = 0.714	β = -0.001 p = 0.932

3.2.2 Random effects and slope models

969 As argued in the introduction (Sect. 1), an approach that focuses exclusively on random

970 effects ignores variation within language families and geographical units [49–51,53,128]. We

971 thus proceed by including random slopes, i.e., we allow the effect of population size to vary

972 across different groups, representing deviations from the overall mean linear effect of speaker

973 population size. The methodological question in this context is, which random effects should

974 include random slopes?

975 We do not include a random slope for language since population size does not vary within

976 languages. Due to geographic proximity and phylogenetic non-independence, it makes sense

977 to include random slopes for macro-area, macro-family, and sub-family as geographic and

978 phylogenetic structures can be critical for understanding language diversity and evolution,

979 which justifies including these random slopes to account for shared inheritance and

980 environmental factors [49].

981 There are, however, no a priori reasons for whether or not random slopes for writing script

982 and corpus also make sense since both are not necessarily tied to population size or linguistic

983 features in a way that would suggest significant variability in the effect of population size

984 across different scripts. Including random slopes for both without clear justification could lead

45

985 to overfitting, adding unnecessary complexity to the model and especially reducing the power

986 to detect true effects: as noted by [129], a too-complex MLEM, by including excessive

987 random slopes, may inflate Type I error rates and reduce the ability to identify significant

988 predictors due to overfitting. This underscores the importance of balancing model complexity

989 with the need to capture meaningful variability.

990 We thus opted for a two-stage estimation process. We first include random slopes for macro-

991 area, macro-family, and sub-family only in our multi-model multilevel approach ($R =17,920$).

992 As a second step, we then additionally include random slopes for writing script and corpus (R

993 $= 35,200$). Figure 8 visualises the results.

994 With respect to the relative importance of the fixed and the random effects, both Figure 8a

995 and Figure 8b largely point in the same direction as the random-effects-only approach (see

996 Figure 7a). A noteworthy exception in both cases is that speaker population size is not only

997 maximally important in predicting both $\hat{h}^{best}$ and L, but also very to maximally important in

998 predicting $\hat{K}^{best}$. Regarding the relative importance of the variables for which we include

999 random slopes in both scenarios, there is significant agreement for the two slopes included for

1000 phylogenetic non-independence (macro- and sub-family): neither seems to play an important

1001 role in predicting either L or $\hat{K}^{best}$.

46

Figure panel a — Estimated variable importance (first scenario):

	H^{best} word	H^{best} char	H^{best} BPE	L word	L char	K^{best} word	K^{best} char	K^{best} BPE	
Population size	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.81	F effects
EGIDS = 0	0.40	0.60	0.85	0.36	0.48	0.30	0.35	0.45	
EGIDS = 1	0.30	0.33	0.33	0.38	0.63	0.28	0.29	0.38	
EGIDS = 2	0.47	0.30	0.40	0.33	0.28	0.52	0.41	0.51	
EGIDS = 3	0.45	0.27	0.30	0.36	0.27	0.28	0.38	0.61	
N countries	0.46	0.66	0.64	0.29	0.28	0.89	0.82	0.83	
parallel	0.32	0.27	0.27	0.74	0.35	0.48	0.33	0.27	
Writing script	1.00	1.00	1.00	1.00	1.00	0.77	0.54	1.00	R effects
Corpus	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	
Macro-area	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Macro-family	1.00	1.00	1.00	1.00	0.73	0.48	0.50	0.36	
Sub-family	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Language	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Macro-area	0.26	0.27	0.27	0.97	0.76	1.00	1.00	0.24	R slopes
Macro-family	0.97	0.27	0.47	0.34	0.37	0.13	0.13	0.08	
Sub-family	0.40	0.99	0.86	0.27	0.28	0.27	0.27	0.22	

Figure panel b — Estimated variable importance (second scenario):

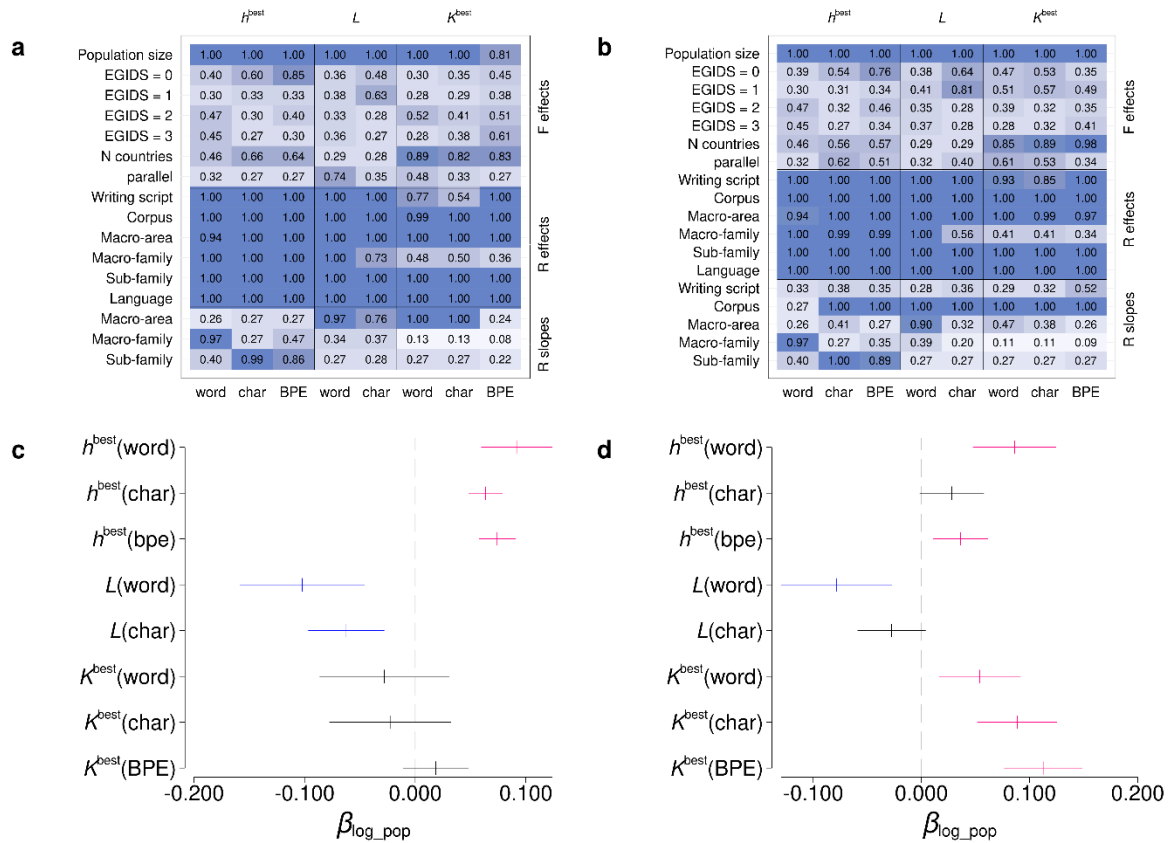|	H^{best} word	H^{best} char	H^{best} BPE	L word	L char	K^{best} word	K^{best} char	K^{best} BPE	
Population size	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	F effects
EGIDS = 0	0.39	0.54	0.76	0.38	0.64	0.47	0.53	0.35	
EGIDS = 1	0.30	0.31	0.34	0.41	0.81	0.51	0.57	0.49	
EGIDS = 2	0.47	0.32	0.46	0.35	0.28	0.39	0.32	0.35	
EGIDS = 3	0.45	0.27	0.34	0.28	0.28	0.32	0.32	0.41	
N countries	0.46	0.56	0.57	0.29	0.29	0.85	0.89	0.98	
parallel	0.32	0.62	0.51	0.32	0.40	0.61	0.53	0.34	
Writing script	1.00	1.00	1.00	1.00	1.00	1.00	0.85	1.00	R effects
Corpus	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Macro-area	0.94	1.00	1.00	1.00	1.00	1.00	0.99	0.97	
Macro-family	1.00	0.99	0.99	1.00	0.56	0.41	0.41	0.34	
Sub-family	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Language	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Writing script	0.33	0.38	0.35	0.28	0.36	0.29	0.32	0.52	R slopes
Corpus	0.27	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Macro-area	0.26	0.41	0.27	0.90	0.32	0.47	0.38	0.26	
Macro-family	0.97	0.27	0.35	0.39	0.20	0.11	0.11	0.09	
Sub-family	0.40	1.00	0.89	0.27	0.27	0.27	0.27	0.27	

Figure 8: Estimated variable importance and FMA-estimates by outcome and symbolic level. (**a** and **c**) For each parameter combination, $R = 17{,}920$ candidate MLEMs that include fixed effects, random effects and random slopes for macro-area, macro-family and sub-family were run. (**b** and **d**) For each parameter combination, $R = 35{,}200$ candidate MLEMs that include fixed effects, random effects and random slopes for writing script, corpus, macro-area, macro-family and sub-family were run. (**a** and **b**) Estimated variable importance ($\hat{\sigma}_x$) per variable. Higher values indicate greater importance (cf. Sect. 2.6.2 for details), $\hat{\sigma}_x$-values range from 0 (white) to 1 (blue). (**c** and **d**) FMA-estimated effect of speaker population size on each outcome per symbolic level. Vertical lines represent the FMA-estimate, $\tilde{\beta}_x$, of population size, here denoted as $\beta_{\log\_pop}$, on $\hat{h}^{best}$, L or $\widehat{K}^{best}$. Horizontal lines show corresponding 95%-CIs (cf. Sect. 2.6.2 for details). Lines are coloured in black if the 95%-CI crosses zero (vertical dashed grey line), whereas blue and pink indicate significant negative and positive effects. Analogous LM-specific interactive visualisations and numeric results are available at https://www.owid.de/plus/tradeoffvis/.

For $\hat{h}^{best}$, random interactions between population size and either macro-family or sub-family is very important. Interestingly, the variable included to account for geographic proximity as a random slope (macro-area) only plays an important role in predicting either L or $\widehat{K}^{best}$ in the first scenario. It seems that, to a large extent, this influence might be absorbed by the inclusion of random slopes for writing script and corpus in the second scenario, especially for L on the character level and for $\widehat{K}^{best}$ on both the word and the character level. The inclusion

47

1021   of writing script as a random slope does not seem to be very important. However, including

1022   corpus seems to make a difference for all seven parameter combinations, except for $\hat{h}^{best}$ on

1023   the word level.

1024   Regarding the FMA-estimated effect ($\tilde{\beta}_x$) of speaker population size, Figure 8c shows that in

1025   the first scenario, the results for each outcome and symbolic level are qualitatively identical to

1026   the random-effects-only approach (see Figure 7b): (i) a significant positive effect of

1027   population size on $\hat{h}^{best}$, (ii) a significant negative effect on L, and (iii) no significant

1028   evidence for any effect on $\hat{K}^{best}$. This again suggests a trade-off between entropy and length.

1029   In the second scenario, while there remains stable evidence for an entropy-length trade-off for

1030   words as symbols, at the character level, neither the positive effect on $\hat{h}^{best}$ nor the negative

1031   effect on L reaches significance at the 5% level ($p = 0.064$ for $\hat{h}^{best}$ and $p = 0.086$ for L).

1032   Unexpectedly, for $\hat{K}^{best}$, there is a significant positive effect of population size across all

1033   three levels. Future work could determine whether this indicates a true effect or if this result

1034   arises from an overly complex model structure that is unable to detect true effects accurately.

1035   To foster such endeavours, we provide a dataset on our OSF repository (https://osf.io/93csg/)

1036   that contains all information needed to replicate our findings and conduct follow-up

1037   investigations. This dataset includes estimates for both (i) $\hat{h}^{best}$ and K^{best} and (ii) LM-

1038   specific estimates, $\hat{h}^{LM}$ and $\hat{K}^{LM}$ for the seven investigated LMs across all three symbolic

1039   levels and two different outcome transformations (standardised vs. logged, see Sect. 2.6.2 for

1040   details), totalling information for $R = 3{,}520{,}000$ individual models.

1041   Further note that as mentioned above (see Sect. 2.6.2), we chose AIC_j as the criterion to weigh

1042   models. AIC is computed by balancing goodness-of-fit with model complexity, i.e., $-2\hat{\mathcal{L}}_j +$

1043 $2(k_j^f + k_j^r)$, where $\hat{\mathcal{L}}_j$ denotes the maximized log-likelihood of model j and k_j^f and k_j^r are the

48

1044 numbers of estimated fixed effects and random effects (including random slopes) parameters,

1045 respectively. However, AIC is not the only potential choice as a criterion. For example, we

1046 can choose a variant of the Bayesian Information Criterion (BIC) [130–132] that imposes a

1047 more substantial penalty on the inclusion of additional parameters than AIC, computed as

1048 $-2\hat{\mathcal{L}}_j + 2(\log(N - k_j^f) k_j^r)$. If we use this criterion to compute results for the second

1049 scenario, all obtained FMA-estimated effects of speaker population size, i.e., positive effects

1050 for h and negative effects for L, reach significance at $p < 0.05$ on all three symbolic levels. We

1051 invite interested readers to further explore this using our interactive visualization tool

1052 available at https://www.owid.de/plus/tradeoffvis/, where we offer results for a total of four

1053 different information criteria.

4 Discussion

1055 As written above, Baroni argues that LMs should be seen as distinct algorithmic linguistic

1056 theories rather than "blank slates," as they inherently encode structural biases that shape their

1057 linguistic capabilities. Each LM thus represents a "general theory defining a space of possible

1058 grammars." [48]. Put differently, an LM can be seen as a model of an idealised language

1059 learner [10,133,134]. Hence, we can think of an LM that is trained on language-specific data

1060 "as a *grammar*, that is, a computation system that, given an input, can predict whether the

1061 sequence is acceptable to an idealized speaker of the language" [48]. In this paper, we

1062 investigated very different types of LMs, belonging to different classes: statistical, machine

1063 learning, and deep learning models. Each class exemplifies unique learning capabilities and

1064 limitations:

1065 PPM can be seen as an idealised learner focused on identifying and refining local word

1066 sequence patterns. By memorising partial matches with the last few symbols (e.g., 2 for

1067 PPM2, 6 for PPM6), PPM constructs a probabilistic grammar of language that adjusts

1068 predictions based on immediate context. This model is particularly effective at learning the

49

1069 structure of short-term dependencies and frequent n-grams within sentences based on very

1070 little input, making it adept at capturing local grammatical rules but limited in handling long-

1071 term dependencies.

1072 PAQ functions as an idealised learner that integrates insights from multiple models to form a

1073 coherent representation of grammar. Each model within PAQ contributes specialised

1074 'knowledge', and a gated linear network moderates these contributions to balance and refine

1075 various perspectives. This process allows PAQ to learn from multiple strings simultaneously,

1076 adapting to a wide range of linguistic patterns and refining predictions through consensus.

1077 However, the complexity of integrating multiple models may limit its ability to focus on

1078 highly specific patterns or rare linguistic structures.

1079 LSTM networks represent an idealised learner adept at preserving and utilizing temporal

1080 sequences. By using memory cells and gates, LSTMs maintain continuity and context over

1081 longer sequences, enabling them to learn long-term dependencies and sequential information.

1082 This capability allows LSTMs to represent the progression of language over time, capturing

1083 narrative coherence and integrating both long-term and short-term adjustments. However,

1084 LSTMs may struggle with complex hierarchical structures due to their sequential processing

1085 nature and consequently require extensive training time.

1086 Lastly, the Transformer-XL model [86], which the NNCP compressor used by us is based on

1087 [83], can be seen as an idealised learner that is proficient at mapping global context through

1088 its self-attention mechanism. This mechanism enables each symbol to dynamically relate to

1089 others, constructing a comprehensive web of interactions between symbols. Transformers are

1090 highly capable of learning complex dependencies and contextual relationships within entire

1091 sentences or documents. Their ability to process all preceding tokens simultaneously allows

1092 them to represent language in a broad and interconnected manner, making them particularly

1093 adept at generating coherent and contextually relevant text. However, this ability also comes

50

1094   at a cost, as Transformer models need to be trained on huge amounts of data to achieve this

1095   level of performance (see, e.g., Figure 2a-c).

1096   Our first main result (Sect. 3.1.1) indicates that the choice of the LM has very little impact on

1097   the obtained results. Given the far-reaching architectural differences between the investigated

1098   LMs, we think this is a surprising result. For instance, a PPM2 model, by design, lacks the

1099   memory to store long-term dependencies, yet Figure 2d-f shows that the results are highly

1100   comparable across LMs. This trend holds even for our largest corpus, UNPC ($\tilde{L}_W =$

1101   341,723,872, see Sect. 2.1.3), which contains information for six languages. For example, the

1102   median value of ρ_{both} between $\hat{h}^{PPM2}$ and $\hat{h}^{TRFmed}$, the former estimate being based on an LM

1103   with a context window of exactly two symbols and the latter based on a transformer model

1104   with ~19.1 million parameters (see Table 2), across symbolic levels ($N_\rho = 9$) is $\tilde{\rho}_{\text{both}} = 0.90$.

1105   Similarly, take H_{Cr}, which is based on an even simpler LM, i.e., a 1-gram LM that does not

1106   consider any relationships between words but only their frequency of occurrence. Yet the

1107   median of ρ_{both} between H_{Cr} and $\hat{h}^{LM}$ for the seven considered LMs across levels and corpora

1108   ($N_\rho = 840$) is $\tilde{\rho}_{\text{both}} = 0.42$. This consistency across LMs is an important observation for

1109   several reasons. Firstly, larger LMs are notably more expensive to train, requiring

1110   substantially more computational resources as compared to smaller models like PAQ, as

1111   outlined in Table 2. Additionally, larger models need a lot of training data to achieve optimal

1112   performance (see Figure 2). This cost-effectiveness makes smaller LMs particularly attractive

1113   for cross-linguistic studies, especially when computational resources are scarce. Furthermore,

1114   as written above, the available electronic data for many languages, particularly those spoken

1115   by smaller populations, is very limited [36]. Our results indicate that training smaller LMs in

1116   such endeavours seems to be a viable option.

1117   It is important to point out that our demonstration of consistency across LMs occurs under a

1118   specific scenario: we examined whether the information-theoretic complexity of languages

51

1119 relative to each other remains consistent regardless of the LM used. In other words, if

1120 language A is deemed more complex than language B when analysed with one particular LM,

1121 this relationship persists even when a completely different type of LM is employed. Future

1122 cross-linguistic research should explore whether this agreement across LMs extends to other

1123 types of analyses and questions, particularly those that are not strictly based on information-

1124 theoretic measures.

1125 As our second main result (Sect. 3.1.2), we re-evaluated the hypothesis that all languages are

1126 equally complex. To address this, we developed a statistical approach that integrates machine

1127 learning with spatial filtering methods. This methodology, detailed in Sect. 2.6.1, was

1128 designed to control for language- and document-specific characteristics, as well as the

1129 phylogenetic and geographic relationships between languages. We used this approach to

1130 compare entropy estimates across corpora and showed that for different LMs, different types

1131 of symbols as information encoding units, and under control of potential sources of influence,

1132 a language with a high/low entropy rate in one corpus also tends to be more/less complex in

1133 another corpus. While entropy rate measures the difficulty in predicting text after the

1134 statistical structure of the input language has been learned [135], we can further explore the

1135 learning process itself by examining how difficult it is for LMs to learn these predictions. In

1136 our prior work [37], we showed that languages which are harder to predict often tend to be

1137 easier/faster to learn for PPM models. Building on this, recent findings [54] provide evidence

1138 of a relationship between learning difficulty and speaker population size, suggesting –

1139 contrary to expectations derived from previous research – that languages with larger speaker

1140 populations tend to be harder for LMs to learn. These results, along with the analyses

1141 presented here, offer information-theoretic evidence that challenges the equi-complexity

1142 hypothesis at the level of written language.

1143   This result inevitably leads to the question: as higher complexity in language results in more

1144   demanding processing efforts, why should there be a trend towards increased complexity in

1145   certain languages? We provide a potential answer to this question as our third main result

1146   (Sect. 3.1.2): we showed that there is a trade-off between the distributions of estimated

1147   entropy rates and length across corpora and across LMs. Given that, from an information-

1148   theoretic perspective, message length quantifies efficiency [15,136] we argue that this result

1149   suggests that higher complexity is compensated by higher efficiency.

1150   In discussions about the obtained trade-off between L and the average amount of information

1151   per symbol, i.e., h, a recurring objection was that this result seems trivial. As some colleagues

1152   pointed out, if the same message is encoded in two languages, each symbol in the language

1153   with the shorter message length must transmit more information, almost by definition. In a

1154   similar vein, in a recent publication a large-scale quantitative information-theoretic analysis of

1155   parallel corpus data in almost 1,000 languages was presented to show that there are apparently

1156   strong associations between the way languages encode information into words and patterns of

1157   communication, e.g. the configuration of semantic information [137]. This publication was

1158   criticised by [138] who demonstrated that the results presented by [137] are systematically

1159   biased by varying text lengths, which is a very well-known fact in quantitative linguistics as

1160   most, if not all, quantities in the context of word frequency distributions vary systematically

1161   with text length [139–141]. The authors of [137] responded that what they call "information

1162   density" and text length are "two sides of the same coin" and that the Gibbs-Shannon entropy

1163   and text length, conditional on the same content, measure the same underlying construct

1164   [142]: "because the information content of the parallel translations is the same across

1165   comparison languages, we can infer that the more words present [sic] within a document

1166   covering the same material, the less information is encoded in each." Another recent response

1167   to [137] made a similar argument: "if it takes a language more words to convey the 'same'

53

1168   message, then each word conveys less information." [143]. Both arguments are incorrect.

1169   First, let us note that a trade-off between h and L does not only occur for parallel corpora, but

1170   it is also observed (i) for comparable corpora (the median adjusted correlation, ρ_{both}, between

1171 $\hat{h}^{best}$ and L for the seven comparable corpora in our database (see Sect. 2.1 and Supporting

1172   Information S3) amounts to $\tilde{\rho}_{\text{both}} = $ -0.47 for words, $\tilde{\rho}_{\text{both}} = $ -0.59 for characters and $\tilde{\rho}_{\text{both}} = $ -

1173   0.59 for BPE, $N_\rho = 49$), and (ii) across corpora – in other words: if we know the entropy

1174   distributions in one multilingual text collection (parallel or not), we can predict the length in

1175   another corpus. For example, the adjusted pairwise correlation ($N = 73$) between $\hat{h}^{best}$ on the

1176   BPE level for the UDHR (parallel) corpus and L on the character level for the LCC news

1177   (comparable) corpus is $\rho_{\text{both}} = $ -0.79. Secondly, eq. 1 clearly demonstrates that the Gibbs-

1178   Shannon unigram entropy is not simply equivalent to text length, but rather a diversity index

1179   [45] that measures the amount of "freedom in the combination of symbols" [144], as it is a

1180   function of the number of different symbols and how evenly those symbols are distributed.

1181   Thirdly, the main problem with such arguments is that the information-theoretic concept of

1182   information is fully agnostic about the content of messages [15]. As Shannon, the founding

1183   father of information theory, puts it [136]:

1184   "The fundamental problem of communication is that of reproducing at one point either

1185   exactly or approximately a message selected at another point. Frequently the messages have

1186   *meaning*; that is they refer to or are correlated according to some system with certain physical

1187   or conceptual entities. These semantic aspects of communication are irrelevant to the

1188   engineering problem. The significant aspect is that the actual message is one selected from a

1189   set of possible messages."

1190   In fact, the use of "information" as a label (e.g., "amount of information" or "information

1191   content") has been a "continuous source of misunderstanding" since the very inception of

1192   information theory, as pointed out by Bar-Hillel [145], because "it is psychologically almost

1193   impossible not to make the shift from the one sense of information, […], i.e., information =

1194   signal sequence, to the other sense, information = what is expressed by the signal sequence".

1195   In information theory, messages are only treated as signal sequences but not as "content-

1196   bearing entities" [145]. Bar-Hillel shows that only under very special circumstances, which do

1197   not typically apply to human language, can we infer the amount of information conveyed

1198   merely by the length of a message. There is no logical connection between the concept of

1199   semantic information (i.e., what is "expressed" by a transmitted signal sequence) and the

1200   rarity or improbability of the signal, which is measured by h. Put differently, information in

1201   the information-theoretic sense has nothing to do with meaning, but only measures the

1202   amount of uncertainty that is removed when receiving a message or the average amount of

1203   information learned about upcoming symbols when observing a symbol of the message.

1204   To illustrate this, consider the sentence "Rain occurs most frequently in spring." Using LCC

1205   word frequency information (cf. Sect. 0), the predictive uncertainty or self-information, I, of

1206   the sentence, calculated based on a unigram LM (see eq. 1), is $I_{word}^{1gram} = 12.531$ bits per

1207   symbol (bps). Now compare this with the sentence "Liquid precipitation occurs most

1208   frequently in spring." If [142,143] were correct, we should expect that this sentence has lower

1209   per-symbol information content, as it contains the same message but uses 7 instead of 6

1210   words. However, the self-information for this sentence is $I_{word}^{1gram} = 13.646$ bps and therefore

1211   higher. This is due to the fact that "rain" is much more frequent in our example data ($f_{LCC} =$

1212   1,594) than "liquid" ($f_{LCC} = 324$) or "precipitation" ($f_{LCC} = 96$). Since a unigram LM does not

1213   consider any contextual information, it makes sense to additionally use a large LM to compute

1214   self-information. Using a GPT-2 model with ~1.5 billion parameters, the self-information of

1215   the first sentence amounts to $I_{word}^{GPT-2} = 0.952$ bps. Again, the self-information of the second

1216   sentence is higher with $I_{word}^{GPT-2} = 0.965$ bps.

55

1217   As a further illustration, take the following two sentences: "I watched TV." and "I watched

1218   television." From the point of view of propositional content, both sentences contain the same

1219   message. Following [142,143], we thus should expect that the self-information is the same in

1220   both cases. However, both I_{word}^{1gram} and I_{word}^{GPT-2} are lower for the first sentence (I_{word}^{1gram} =

1221   11.915, I_{word}^{GPT-2} = 3.068) compared to the second one (I_{word}^{1gram} = 12.397, I_{word}^{GPT-2} = 3.210)

1222   again because "television" is much less frequent (f_{LCC} = 1,568) than "TV" (f_{LCC} = 4,269).

1223   Finally, consider the sentence pair "They will attack at dawn." versus "They will attack at

1224   5am." From the point of view of propositional content, the "total amount of information" of

1225   the second sentence should be higher because it specifies a precise time. Based on a unigram

1226   LM, the total amount of information is I_{total}^{1gram} = 53.389 bits for the first sentence and I_{total}^{1gram}

1227   = 55.884 bits for the second one. However, based on the arguably much better GPT-2 LM, we

1228   obtain the opposite result with I_{total}^{GPT-2} = 7.000 bits for the first sentence and I_{total}^{GPT-2} = 6.620

1229   bits for the second one.

1230   These illustrations demonstrate that both H and h reflect the statistical structure of "linguistic

1231   sequences independently of the specific information that [is] being encoded" [146]. As such,

1232   it is important to point out that information theory cannot be used to measure something like

1233   the total amount of semantic information/propositional content of a sentence. Instead,

1234   information theory provides a framework for quantifying the efficiency of symbol sequences

1235   in transmitting data, focusing on the probabilistic structure and redundancy of the language

1236   rather than its semantic content.

1237   To elaborate further, it should be noted that translation from one language into another is

1238   fundamentally different from what is meant by encoding in the information-theoretic sense. In

1239   early works on machine translation, such as Warren Weaver's famous memorandum [147],

1240   translation was sometimes understood to be similar to, e.g., cryptographic decoding. But

1241   encoding schemes are basically just biunique mappings of an information source to symbol

1242   sequences that permit exact recovery of the original symbols, whereas there is no information-

1243   theoretically well-defined sense in which one (of possibly very many) translation of a text into

1244   another human language conveys the "same content". The difference is similar to that

1245   between sign language and fingerspelling. A written English text can be translated into a sign

1246   language, such as American Sign Language (ASL). Since ASL is a full-blown natural

1247   language of its own (with no relation whatsoever to spoken English), there are usually many

1248   possible translations none of which allows unambiguous reconstruction of the source text

1249   from it. On the other hand, fingerspelling (which is a part of ASL) enables users to

1250   mechanically render e.g. English words through sequences of signs representing letters and

1251   can thus, in principle, be back-translated losslessly to the oral language original.

1252   The biuniqueness of encoding schemes is the reason why text length and entropy rate are

1253   indeed trivially inversely related to each other when comparing different encodings of a

1254   source text, since different encodings of the same source can always be compressed to the

1255   exact same outcome and the length of that outcome is then used to estimate the entropy rate.

1256   For translations between human languages no such argument is available. Indeed, as we will

1257   demonstrate in what follows, shorter translations may in principle come with a lower entropy

1258   rate instead of a higher one.

1259   Before further interpreting the complexity-efficiency trade-off, we will discuss the results,

1260   presented in the Supporting Information, from several additional analyses we conducted to

1261   test the reliability and validity of this trade-off. First, practicing what we preached above, we

1262   rule out the possibility that the association between h and L is simply the result of a well-

1263   known systematic text-length bias (Supporting Information S4). Secondly, we demonstrate

1264   that there is clear evidence for an entropy-length trade-off only *between* languages, but not

1265   *within* languages (Supporting Information S5). Thirdly, we show in Supporting Information

57

1266   S6 both theoretically and empirically that the trade-off is indeed not trivial, as one can define

1267   processes that increase both entropy and length at the same time. Fourthly, we control for

1268   cross-linguistic differences in the number of different words and characters that might stem

1269   from, amongst others, morphological typology, different writing systems, varying

1270   phonological constraints etc. (see Supporting Information S7). When using the number of

1271   different wordform types and the number of different characters (both log-transformed

1272   averages over all translations per language in the *BibleNT* collection) as additional control

1273   variables when assessing the (dis)similarity of entropy and length distributions across corpora.

1274   Fifthly, to address the issue of varying levels of orthographic depth (see Sect. 2.4), we

1275   incorporated data from ref. [148], who uses a transformer LM to estimate the percentage of

1276   correct predictions in phoneme-to-grapheme and grapheme-to-phoneme "translation" tasks

1277   across 16 languages. These estimates, which serve as measures of phonemic transparency,

1278   provide a quantitative assessment of the ease with which written text can be converted to or

1279   from its spoken form, i.e., how directly characters map to phonemes. In Supporting

1280   Information S8, we used this information as an additional covariate to control for

1281   orthographic depth in our analyses. In both Supporting Information S7 and Supporting

1282   Information S8 our results remain robust: (i) languages with high/low entropy rates in one

1283   corpus also tend to be more/less complex in another corpus, (ii) an analogous pattern holds for

1284   cross-language length distributions, and (iii) there is a trade-off between the distributions of

1285   estimated entropy rates and text length across corpora. Sixthly, we show in Supporting

1286   Information S9 that the patterns observed in the other investigate multilingual text collection

1287   align with those from the PBC. As written above (see Sect. 2.4), we believe that this

1288   consistency bolsters confidence in the robustness and validity of our quantitative results, due

1289   to the substantial effort and typologically informed curation of the PBC (see Sect. 2.1.1).

1290   Seventhly, our study focuses exclusively on written language. To explore a potential

1291   connection to spoken language, we utilised data from the *VoxClamantis* corpus [149], which

58

is based on audio recordings of the New Testament of the Bible. In Supporting Information S10, we outline how phoneme sequences were prepared and used to train PAQ in order to compute an entropy estimate at the phonemic level, denoted as $\hat{h}^{\varphi}$. For $N_L = 26$ languages from 8 different language families, we have estimates at both the phonemic level from the *VoxClamantis* corpus and the written level from the *BibleNT* corpus. Using these estimates, we calculated the Pearson correlation between the logarithms of $\hat{h}^{\varphi}$ and $\hat{h}^{best}(\kappa)$. At the word level, the correlation is $\rho_{words} = 0.32$, while at the character and sub-word levels, the correlations are stronger, with $\rho_{characters} = 0.58$ and $\rho_{BPE} = 0.54$, respectively. The results also suggest an inverse relationship between $\hat{h}^{\varphi}$ and the logged text length $L(\kappa)$, with a more pronounced effect at the character level ($\rho_{characters} = $ -0.52) compared to the word level ($\rho_{characters} = $ -0.26). While these findings are preliminary, due to the limited sample size, the varying quality and quantity of Bible readings per language, and other caveats [149], they suggest that the complexity-efficiency trade-off observed in written language may also extend to spoken language. We believe this indicates the potential for further exploration of spoken language as a valuable direction for future research.

With these results in mind, we conclude this paper by discussing potential reasons for a trade-off between complexity and message length. This discussion is based on our fourth main result: using a multi-model multilevel approach detailed in Sect. 2.6.2, we presented findings in Sect. 3.2 indicating that the trade-off is influenced by the social environment in which languages are learned and used. Specifically, languages with more speakers tend to be more complex. At first glance, this result contrasts with previous research suggesting that languages spoken in larger communities tend to be less complex [150–156], as larger communities are assumed to favour simple and predictable language structures. At the same time, our results indicate that languages with more speakers tend to produce shorter messages, i.e., are more efficient.

1317 Let us speculate that in large societies, institutionalised education potentially makes greater

1318 linguistic complexity possible by providing systematic and formalised language learning,

1319 which, in turn, can support the acquisition and use of more complex linguistic structures. In

1320 line with this, a recent large-scale study found a positive statistical correlation between

1321 grammatical complexity and speaker population size [53]. At the same time, the importance

1322 of written communication in larger societies might create a natural pressure towards shorter

1323 messages, as it saves costs for producing, storing, and transmitting written texts (e.g., book

1324 paper, storage space, bandwidth). This dual influence – educational systems enabling

1325 complexity and the practical need for efficiency in written communication – could help

1326 explain why languages in larger communities might evolve to balance these pressures,

1327 resulting in shorter but more complex messages. Testing this hypothesis is an important

1328 avenue for future research.

1329

1342 **Authors' contributions**

1343 **AK**: conceptualisation (lead), data curation, resources (lead), formal analysis, investigation,
1344 methodology, software (lead), validation (lead), visualisation (lead), writing – original draft, writing –
1345 review & editing (equal). **SW**: software (supporting), visualisation (supporting), validation

(supporting), writing – review & editing (equal). **JOR**: resources (supporting), writing – review & editing (supporting). **PM**: conceptualisation (supporting), writing – review & editing (equal).

References

1. Jurafsky D, Martin JH. Speech and Language processing: an introduction to natural language processing, computational Linguistics, and speech recognition. Upper Saddle River: Pearson Education (US); 2009.

2. Piantadosi ST. Modern language models refute Chomsky's approach to language. 2023. doi:lingbuzz/007180

3. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024 [cited 3 Jun 2024]. doi:10.1038/s41586-024-07487-w

4. Gruver N, Finzi M, Qiu S, Wilson AG. Large Language Models Are Zero-Shot Time Series Forecasters. arXiv; 2023. doi:10.48550/ARXIV.2310.07820

5. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37: 1038–1040. doi:10.1038/s41587-019-0224-x

6. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559: 547–555. doi:10.1038/s41586-018-0337-2

7. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 2015;12: 931–934. doi:10.1038/nmeth.3547

8. Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling Climate Change with Machine Learning. ACM Comput Surv. 2023;55: 1–96. doi:10.1145/3485128

9. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2020. pp. 1877–1901. Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

10. Chater N, Vitányi P. 'Ideal learning' of natural language: Positive results about learning from positive evidence. Journal of Mathematical Psychology. 2007;51: 135–163. doi:10.1016/j.jmp.2006.10.002

11. Contreras Kallens P, Kristensen-McLachlan RD, Christiansen MH. Large Language Models Demonstrate the Potential of Statistical Learning in Language. Cognitive Science. 2023;47: e13256. doi:10.1111/cogs.13256

12. Grindrod J. Modelling Language. arXiv; 2024. doi:10.48550/ARXIV.2404.09579

1384   13.   Abramski K, Citraro S, Lombardi L, Rossetti G, Stella M. Cognitive Network Science
1385   Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-
1386   School Students. BDCC. 2023;7: 124. doi:10.3390/bdcc7030124

1387   14.   Mahoney M. Data Compression Explained. Dell Inc.; 2013. Available:
1388   http://mattmahoney.net/dc/dce.html

1389   15.   Gibson E, Futrell R, Piandadosi ST, Dautriche I, Mahowald K, Bergen L, et al. How
1390   Efficiency Shapes Human Language. TRENDS in Cognitive Science. 2019;23: 389–
1391   407. doi:10.1016/j.tics.2019.02.003

1392   16.   Cysouw M, Wälchli B. Parallel texts: using translational equivalents in linguistic
1393   typology. Language Typology and Universals. 2007;60: 95–99.
1394   doi:10.1524/stuf.2007.60.2.95

1395   17.   Bentz C, Ferrer-i-Cancho R. Zipf's law of abbreviation as a language universal. In:
1396   Bentz C, Jäger G, Yanovich I, editors. Proceedings of the Leiden Workshop on
1397   Capturing Phylogenetic Algorithms for Linguistics. Tübingen: University of Tübingen;
1398   2016.

1399   18.   Bentz C. The Low-complexity-belt: Evidence For Large-scale Language Contact In
1400   Human Prehistory? In: Roberts SG, Cuskley C, McCrohon L, Barceló-Coblijn L, Fehér
1401   O, Verhoef T, editors. The Evolution of Language: Proceedings of the 11th
1402   International Conference (EVOLANGX11). Online at
1403   http://evolang.org/neworleans/papers/93.html; 2016.

1404   19.   Koplenig A. Language structure is influenced by the number of speakers but seemingly
1405   not by the proportion of non-native speakers. Royal Society Open Science. 2019;6:
1406   181274. doi:10.1098/rsos.181274

1407   20.   Koplenig A. Quantifying the efficiency of written language. Linguistics Vanguard.
1408   2021;7: 20190057. doi:10.1515/lingvan-2019-0057

1409   21.   Kauhanen H, Einhaus S, Walkden G. Language structure is influenced by the
1410   proportion of non-native speakers: A reply to Koplenig (2019). Journal of Language
1411   Evolution. 2023; lzad005. doi:10.1093/jole/lzad005

1412   22.   Koplenig A. Still no evidence for an effect of the proportion of non-native speakers on
1413   language complexity -- A response to Kauhanen, Einhaus & Walkden (2023). 2023
1414   [cited 8 May 2023]. doi:10.48550/ARXIV.2305.00217

1415   23.   Greenberg JH. A Quantitative Approach to the Morphological Typology of Language.
1416   International Journal of American Linguistics. 1960;26: 178–194.

1417   24.   Bentz C, Ruzsics T, Koplenig A, Samardzic T. A Comparison Between Morphological
1418   Complexity Measures: Typological Data vs. Language Corpora. Proceedings of the
1419   Workshop on Computational Linguistics for Linguistic Complexity (CL4LC). Osaka,
1420   Japan; 2016.

1421   25.   Wälchli B, Cysouw M. Lexical typology through similarity semantics: Toward a
1422   semantic map of motion verbs. Linguistics. 2012;50. doi:10.1515/ling-2012-0021

1423   26.   Östling R. Word Order Typology through Multilingual Word Alignment. Proceedings
1424   of the 53rd Annual Meeting of the Association for Computational Linguistics and the
1425   7th International Joint Conference on Natural Language Processing (Volume 2: Short
1426   Papers). Beijing, China: Association for Computational Linguistics; 2015. pp. 205–211.
1427   doi:10.3115/v1/P15-2034

1428   27.   Bentz C, Alikaniotis D, Cysouw M, Ferrer-i-Cancho R. The Entropy of Words—
1429   Learnability and Expressivity across More than 1000 Languages. Entropy. 2017;19:
1430   275. doi:10.3390/e19060275

1431   28.   Koplenig A, Meyer P, Wolfer S, Müller-Spitzer C. The statistical trade-off between
1432   word order and word structure – Large-scale evidence for the principle of least effort.
1433   Smith K, editor. PLOS ONE. 2017;12: e0173614. doi:10.1371/journal.pone.0173614

1434   29.   Pimentel T, Meister C, Salesky E, Teufel S, Blasi D, Cotterell R. A surprisal–duration
1435   trade-off across and within the world's languages. Proceedings of the 2021 Conference
1436   on Empirical Methods in Natural Language Processing. Online and Punta Cana,
1437   Dominican Republic: Association for Computational Linguistics; 2021. pp. 949–962.
1438   doi:10.18653/v1/2021.emnlp-main.73

1439   30.   de Vries LJ. Some remarks on the use of Bible translations as parallel texts in linguistic
1440   research. Sprachtypologie und Universalienforschung. 2007;60: 148–157.
1441   doi:10.1524/stuf.2007.60.2.148

1442   31.   Wälchli B. Advantages and disadvantages of using parallel texts in typological
1443   investigations. Language Typology and Universals. 2007;60: 118–134.
1444   doi:doi:10.1524/stuf.2007.60.2.118

1445   32.   Tiedemann J. Parallel Data, Tools and Interfaces in OPUS. LREC'12 Proceedings.
1446   Istanbul, Turkey: ELRA; 2012. pp. 2214–2218.

1447   33.   Mayer T, Cysouw M. Creating a Massively Parallel Bible Corpus. In: Chair) NC
1448   (Conference, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, et al.,
1449   editors. Proceedings of the Ninth International Conference on Language Resources and
1450   Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association
1451   (ELRA); 2014.

1452   34.   Levshina N. Verbs of letting in Germanic and Romance languages: A quantitative
1453   investigation based on a parallel corpus of film subtitles. LiC. 2016;16: 84–117.
1454   doi:10.1075/lic.16.1.04lev

1455   35.   Goldhahn D, Eckart T, Quasthoff U. Building Large Monolingual Dictionaries at the
1456   Leipzig Corpora Collection: From 100 to 200 Languages. Proceedings of the Eighth
1457   International Conference on Language Resources and Evaluation (LREC'12). Istanbul,
1458   Turkey: European Language Resources Association (ELRA); 2012. pp. 759–765.
1459   Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf

1460   36.   Scannell KP. The Crúbadán Project: Corpus building for under-resourced languages.
1461   Proceedings of the 3rd Web as Corpus Workshop: Building and Exploring Web
1462   Corpora. 2007. pp. 5–15. Available: http://cs.slu.edu/~scannell/pub/wac3.pdf

1463 37. Koplenig A, Wolfer S, Meyer P. A large quantitative analysis of written language
1464   challenges the idea that all languages are equally complex. Sci Rep. 2023;13: 15351.
1465   doi:10.1038/s41598-023-42327-3

1466 38. Ehret K, Blumenthal-Dramé A, Bentz C, Berdicevskis A. Meaning and Measures:
1467   Interpreting and Evaluating Complexity Metrics. Front Commun. 2021;6: 640510.
1468   doi:10.3389/fcomm.2021.640510

1469 39. Sampson G. A linguistic axiom challenged. In: Sampson G, Gil D, Trudgill P, editors.
1470   Language complexity as an evolving variable. Oxford: Oxford University Press; 2009.
1471   pp. 1–18.

1472 40. Ehret K, Berdicevskis A, Bentz C, Blumenthal-Dramé A. Measuring language
1473   complexity: challenges and opportunities. Linguistics Vanguard. 2023;0.
1474   doi:10.1515/lingvan-2022-0133

1475 41. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Hoboken, N.J: Wiley-
1476   Interscience; 2006.

1477 42. Futrell R, Hahn M. Information Theory as a Bridge Between Language Function and
1478   Language Form. Front Commun. 2022;7: 657725. doi:10.3389/fcomm.2022.657725

1479 43. Ren G, Takahashi S, Tanaka-Ishii K. Entropy Rate Estimation for English via a Large
1480   Cognitive Experiment Using Mechanical Turk. Entropy. 2019;21: 1201.
1481   doi:10.3390/e21121201

1482 44. Meister C, Pimentel T, Wiher G, Cotterell R. Locally Typical Sampling. arXiv; 2022.
1483   doi:10.48550/ARXIV.2202.00666

1484 45. Kolmogorov AN. Three approaches to the quantitative definition of information.
1485   International Journal of Computer Mathematics. 1968;2: 157–168.
1486   doi:10.1080/00207166808803030

1487 46. Kontoyiannis I. The Complexity and Entropy of Literary Styles. NSF Technical Report,
1488   Department of Statistics, Stanford University. 1996;97.

1489 47. Cover TM. Kolmogorov Complexity, Data Compression, and Inference. In:
1490   Skwirzynski JK, editor. The Impact of Processing Techniques on Communications.
1491   Dordrecht: Springer Netherlands; 1985. pp. 23–33. doi:10.1007/978-94-009-5113-6_2

1492 48. Baroni M. On the proper role of linguistically-oriented deep net analysis in linguistic
1493   theorizing. arXiv; 2021. doi:10.48550/ARXIV.2106.08694

1494 49. Bromham L. Solving Galton's problem: practical solutions for analysing language
1495   diversity and evolution. PsyArXiv; 2022 May. doi:10.31234/osf.io/c8v9r

1496 50. Bromham L, Yaxley KJ. Neighbours and relatives: accounting for spatial distribution
1497   when testing causal hypotheses in cultural evolution. Evolut Hum Sci. 2023;5: e27.
1498   doi:10.1017/ehs.2023.23

1499 51. Guzmán Naranjo M, Becker L. Statistical bias control in typology. Linguistic
1500   Typology. 2022;26: 605–670. doi:10.1515/lingty-2021-0002

1501   52.   Claessens S, Kyritsis T, Atkinson QD. Cross-national analyses require additional
1502   controls to account for the non-independence of nations. Nat Commun. 2023;14: 5776.
1503   doi:10.1038/s41467-023-41486-1

1504   53.   Shcherbakova O, Michaelis SM, Haynie HJ, Passmore S, Gast V, Gray RD, et al.
1505   Societies of strangers do not speak less complex languages. Science Advances. 2023;9:
1506   eadf7704. doi:10.1126/sciadv.adf7704

1507   54.   Koplenig A, Wolfer S. Languages with more speakers tend to be harder to (machine-
1508   )learn. Sci Rep. 2023;13: 18521. doi:10.1038/s41598-023-45373-z

1509   55.   Hall S, Moskovitz C, Pemberton M, Text Recycling Research Project. Understanding
1510   text recycling. A guide for researchers V.1. 2021. Available:
1511   https://textrecycling.org/files/2021/06/Understanding-Text-Recycling_A-Guide-for-
1512   Researchers-V.1.pdf

1513   56.   Gutierrez-Vasques X, Bentz C, Samardžić T. Languages Through the Looking Glass of
1514   BPE Compression. Computational Linguistics. 2023;49: 943–1001.
1515   doi:10.1162/coli_a_00489

1516   57.   Simons GF, Fennig CD. Global Dataset Ethnologue: Languages of the World,
1517   Twentieth edition. SIL International; 2017. Available: http://www.ethnologue.com.

1518   58.   Bromham L, Dinnage R, Skirgård H, Ritchie A, Cardillo M, Meakins F, et al. Global
1519   predictors of language endangerment and the future of linguistic diversity. Nat Ecol
1520   Evol. 2022;6: 163–173. doi:10.1038/s41559-021-01604-y

1521   59.   Hammarström H, Forkel R, Haspelmath M, Bank S. glottolog/glottolog: Glottolog
1522   database 4.8. Zenodo; 2023. doi:10.5281/ZENODO.8131084

1523   60.   Simons GF, Fennig CD, editors. Ethnologue: languages of Africa and Europe.
1524   Twentieth edition. Dallas, TX: SIL; 2017.

1525   61.   Simons, Gary F, Fennig CD. Ethnologue: Languages of the World. Dallas, Texas: SIL
1526   International; 2017.

1527   62.   Lewis MP, Simons GF. Assessing Endangerment: Expanding Fishman's GIDS. Revue
1528   Roumaine de Linguistique. 2010;55: 103–120.

1529   63.   Baker M. Corpus Linguistics and Translation Studies — Implications and Applications.
1530   In: Baker M, Francis G, Tognini-Bonelli E, editors. Text and Technology. Amsterdam:
1531   John Benjamins Publishing Company; 1993. p. 233. doi:10.1075/z.64.15bak

1532   64.   Cotterell R, Mielke SJ, Eisner J, Roark B. Are All Languages Equally Hard to
1533   Language-Model? Proceedings of the 2018 Conference of the North American Chapter
1534   of the Association for Computational Linguistics: Human Language Technologies,
1535   Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational
1536   Linguistics; 2018. pp. 536–541. doi:10.18653/v1/N18-2085

1537   65.   Jäger G. Global-scale phylogenetic linguistic inference from lexical resources.
1538   Scientific Data. 2018;5: 180189. doi:10.1038/sdata.2018.189

65

1539   66.   Wichmann S, Holman EW, Brown CH, Forkel R, Tresoldi T. CLDF dataset derived
1540   from Wichmann et al.'s "ASJP Database" v17 from 2016. Zenodo; 2016.
1541   doi:10.5281/ZENODO.3835942

1542   67.   Stewart WA. A Sociolinguistic Typology for Describing National Multilingualism. In:
1543   Fishman JA, editor. Readings in the Sociology of Language. DE GRUYTER; 1968. pp.
1544   531–545. doi:10.1515/9783110805376.531

1545   68.   Bentz C, Dediu D, Verkerk A, Jäger G. The evolution of language families is shaped by
1546   the environment beyond neutral drift. Nature Human Behaviour. 2018;2: 816–821.
1547   doi:10.1038/s41562-018-0457-6

1548   69.   Jaeger TF, Graff P, Croft W, Pontillo D. Mixed effect models for genetic and areal
1549   dependencies in linguistic typology. Linguistic Typology. 2011;15.

1550   70.   Shannon CE. Prediction and Entropy of Printed English. Bell System Technical
1551   Journal. 1951;30: 50–64. doi:10.1002/j.1538-7305.1951.tb01366.x

1552   71.   Chaitin GJ. On the intelligibility of the universe and the notions of simplicity,
1553   complexity and irreducibility. arXiv:math/0210035. 2002 [cited 25 Mar 2021].
1554   Available: http://arxiv.org/abs/math/0210035

1555   72.   Delétang G, Ruoss A, Duquenne P-A, Catt E, Genewein T, Mattern C, et al. Language
1556   Modeling Is Compression. arXiv; 2023. doi:10.48550/ARXIV.2309.10668

1557   73.   Rissanen JJ. Generalized Kraft Inequality and Arithmetic Coding. IBM Journal of
1558   Research and Development. 1976;20: 198–203. doi:10.1147/rd.203.0198

1559   74.   Jurafsky D, Martin JH. Speech and Language Processing. 3rd ed. 2021. Available:
1560   https://web.stanford.edu/~jurafsky/slp3/

1561   75.   Cleary J, Witten I. Data Compression Using Adaptive Coding and Partial String
1562   Matching. IEEE Transactions on Communications. 1984;32: 396–402.
1563   doi:10.1109/TCOM.1984.1096090

1564   76.   Shkarin D. PPM: one step to practicality. Proceedings DCC 2002 Data Compression
1565   Conference. Snowbird, UT, USA: IEEE Comput. Soc; 2002. pp. 202–211.
1566   doi:10.1109/DCC.2002.999958

1567   77.   Pavlov I. 7-zip. 2023. Available: https://7-zip.org/

1568   78.   Knoll B, Freitas N de. A Machine Learning Perspective on Predictive Coding with
1569   PAQ8. 2012 Data Compression Conference. Snowbird, UT, USA: IEEE; 2012. pp.
1570   377–386. doi:10.1109/DCC.2012.44

1571   79.   Veness J, Lattimore T, Budden D, Bhoopchand A, Mattern C, Grabska-Barwinska A, et
1572   al. Gated Linear Networks. 2019 [cited 27 Mar 2023].
1573   doi:10.48550/ARXIV.1910.01526

1574   80.   Mahoney M. PAQ8. 2007. Available: http://mattmahoney.net/dc/paq8l.zip

1575   81.   Mahoney M. Adaptive weighing of context models for lossless data compression.
1576   Florida Tech.; 2005. Available: http://hdl.handle.net/11141/154

1577 82. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9:
1578   1735–1780.

1579 83. Bellard F. NNCP v3.1: Lossless Data Compression with Transformer. 2021. Available:
1580   https://bellard.org/nncp/

1581 84. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is
1582   All You Need. Proceedings of the 31st International Conference on Neural Information
1583   Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. pp. 6000–
1584   6010.

1585 85. Witten IH, Bell TC. The zero-frequency problem: estimating the probabilities of novel
1586   events in adaptive text compression. IEEE Transactions on Information Theory.
1587   1991;37: 1085–1094. doi:10.1109/18.87000

1588 86. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL:
1589   Attentive Language Models Beyond a Fixed-Length Context. arXiv:190102860 [cs,
1590   stat]. 2019 [cited 22 Oct 2020]. Available: http://arxiv.org/abs/1901.02860

1591 87. Bellard F. Lossless Data Compression with Neural Networks. 2019. Available:
1592   https://bellard.org/nncp/nncp.pdf

1593 88. Bellard F. NNCP v2: Lossless Data Compression with Transformer. 2021. Available:
1594   https://bellard.org/nncp/nncp_v2.1.pdf

1595 89. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating
1596   errors. Nature. 1986;323: 533–536. doi:10.1038/323533a0

1597 90. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014 [cited 15 May
1598   2023]. doi:10.48550/ARXIV.1412.6980

1599 91. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are
1600   Unsupervised Multitask Learners. 2019. Available:
1601   https://d4mucfpksywv.cloudfront.net/better-language-
1602   models/language_models_are_unsupervised_multitask_learners.pdf

1603 92. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's
1604   Transformers: State-of-the-art Natural Language Processing. arXiv; 2019.
1605   doi:10.48550/ARXIV.1910.03771

1606 93. Unicode Consortium. Unicode Text Segmentation. In: Unicode® Standard Annex #29
1607   [Internet]. 2019 [cited 23 Jul 2019]. Available:
1608   http://www.unicode.org/reports/tr29/#Word_Boundaries

1609 94. Koplenig A. Stata tip 129: Efficiently processing textual data with Stata's new Unicode
1610   features. Stata Journal. 2018;18: 287–289.

1611 95. Mielke SJ, Cotterell R, Gorman K, Roark B, Eisner J. What Kind of Language Is Hard
1612   to Language-Model? Proceedings of the 57th Annual Meeting of the Association for
1613   Computational Linguistics. Florence, Italy: Association for Computational Linguistics;
1614   2019. pp. 4975–4989. doi:10.18653/v1/P19-1491

67

1615   96.   Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with
1616   Subword Units. Proceedings of the 54th Annual Meeting of the Association for
1617   Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for
1618   Computational Linguistics; 2016. pp. 1715–1725. doi:10.18653/v1/P16-1162

1619   97.   Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural
1620   Language Processing, Computational Linguistics, and Speech Recognition with
1621   Language Models. 3rd ed. 2024. Available: https://web.stanford.edu/~jurafsky/slp3/

1622   98.   Darmon ANM, Bazzi M, Howison SD, Porter MA. Pull out all the stops: Textual
1623   analysis via punctuation sequences. Eur J Appl Math. 2021;32: 1069–1105.
1624   doi:10.1017/S0956792520000157

1625   99.   Moran S, Cysouw M. The Unicode Cookbook For Linguists: Managing Writing
1626   Systems Using Orthography Profiles. Berlin: Language Science Press; 2018.
1627   doi:10.5281/zenodo.1296780

1628   100.   Haspelmath M. The indeterminacy of word segmentation and the nature of morphology
1629   and syntax. Folia Linguistica. 2011;45. doi:10.1515/flin.2011.002

1630   101.   Jacobs J. Grammatik ohne Wörter? In: Engelberg S, Holler A, Proost K, editors.
1631   Sprachliches Wissen zwischen Lexikon und Grammatik. Berlin, Boston: DE
1632   GRUYTER; 2011. Available:
1633   http://www.degruyter.com/view/books/9783110262339/9783110262339.345/97831102
1634   62339.345.xml

1635   102.   Lyons J. Introduction to Theoretical Linguistics. 1st ed. Cambridge University Press;
1636   1968. doi:10.1017/CBO9781139165570

1637   103.   Aaron PG, Joshi RM. Written Language Is as Natural as Spoken language: A
1638   Biolinguistic Perspective. Reading Psychology. 2006;27: 263–311.
1639   doi:10.1080/02702710600846803

1640   104.   Berg K. Graphemic Analysis and the Spoken Language Bias. Front Psychol. 2016;7.
1641   doi:10.3389/fpsyg.2016.00388

1642   105.   Chafe W, Tannen D. The Relation Between Written and Spoken Language. Annu Rev
1643   Anthropol. 1987;16: 383–407. doi:10.1146/annurev.an.16.100187.002123

1644   106.   Haspelmath M. Defining the word. WORD. 2023;69: 283–297.
1645   doi:10.1080/00437956.2023.2237272

1646   107.   Geertzen J, Blevins J, Milin P. Informativeness of linguistic unit boundaries. Apollo -
1647   University of Cambridge Repository. 2016 [cited 23 Jul 2019]. doi:10.17863/cam.69

1648   108.   Bickel B, Nichols J. Inflectional morphology. 2nd ed. In: Shopen T, editor. Language
1649   Typology and Syntactic Description. 2nd ed. Cambridge University Press; 2007. pp.
1650   169–240. doi:10.1017/CBO9780511618437.003

1651   109.   Evans N, Sasse H-J. Introduction: problems of polysynthesis. In: Evans N, Sasse H-J,
1652   editors. Problems of Polysynthesis. Berlin: AKADEMIE VERLAG; 2002. pp. 1–14.
1653   doi:10.1524/9783050080956.1

1654 110. Schürmann T, Grassberger P. Entropy estimation of symbol sequences. Chaos: An
1655   Interdisciplinary Journal of Nonlinear Science. 1996;6: 414. doi:10.1063/1.166191

1656 111. MacKay DJC. Information theory, inference, and learning algorithms. Cambridge, UK ;
1657   New York: Cambridge University Press; 2003.

1658 112. Yaglom AM, Yaglom IM. Probability and information. Dordrecht, Holland ; Boston :
1659   Hingham, MA: D. Reidel ; Sold and distributed in the U.S.A. by Kluwer Boston; 1983.

1660 113. Adami C. What is information? Philosophical Transactions of the Royal Society A:
1661   Mathematical, Physical and Engineering Sciences. 2016;374: 20150230.
1662   doi:10.1098/rsta.2015.0230

1663 114. Wolf L, Pimentel T, Fedorenko E, Cotterell R, Warstadt A, Wilcox E, et al.
1664   Quantifying the redundancy between prosody and text. arXiv; 2023.
1665   doi:10.48550/ARXIV.2311.17233

1666 115. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal
1667   Statistical Society Series B (Methodological). 1996;58: 267–288.

1668 116. Belloni A, Chernozhukov V, Wei Y. Post-Selection Inference for Generalized Linear
1669   Models With Many Controls. Journal of Business & Economic Statistics. 2016;34:
1670   606–619. doi:10.1080/07350015.2016.1166116

1671 117. Moran PAP. Notes on Continuous Stochastic Phenomena. Biometrika. 1950;37: 17.
1672   doi:10.2307/2332142

1673 118. Kelejian HH, Prucha IR. On the asymptotic distribution of the Moran I test statistic
1674   with applications. Journal of Econometrics. 2001;104: 219–257. doi:10.1016/S0304-
1675   4076(01)00064-1

1676 119. Griffith DA. A Spatial Filtering Specification for the Autologistic Model. Environ Plan
1677   A. 2004;36: 1791–1811. doi:10.1068/a36247

1678 120. Tiefelsdorf M, Griffith DA. Semiparametric Filtering of Spatial Autocorrelation: The
1679   Eigenvector Approach. Environment and Planning A. 2007;39: 1193–1221.
1680   doi:10.1068/a37378

1681 121. Oberdabernig DA, Humer S, Crespo Cuaresma J. Democracy, Geography and Model
1682   Uncertainty. Scottish J Political Eco. 2018;65: 154–185. doi:10.1111/sjpe.12140

1683 122. Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random
1684   effects for subjects and items. Journal of Memory and Language. 2008;59: 390–412.
1685   doi:10.1016/j.jml.2007.12.005

1686 123. Nettle D. Social scale and structural complexity in human languages. Philosophical
1687   Transactions of the Royal Society B: Biological Sciences. 2012;367: 1829–1836.
1688   doi:10.1098/rstb.2011.0216

1689 124. Burnham KP, Anderson DR, editors. Model Selection and Multimodel Inference. New
1690   York, NY: Springer New York; 2004. doi:10.1007/b97636

1691   125. Buckland ST, Burnham KP, Augustin NH. Model Selection: An Integral Part of
1692   Inference. Biometrics. 1997;53: 603. doi:10.2307/2533961

1693   126. Steel MFJ. Model Averaging and Its Use in Economics. Journal of Economic
1694   Literature. 2020;58: 644–719. doi:10.1257/jel.20191385

1695   127. Akaike H. A new look at the statistical model identification. IEEE Transactions on
1696   Automatic Control. 1974;19: 716–723. doi:10.1109/TAC.1974.1100705

1697   128. Claessens S, Atkinson Q. The non-independence of nations and why it matters.
1698   PsyArXiv; 2022 Apr. doi:10.31234/osf.io/m6bsn

1699   129. Matuschek H, Kliegl R, Vasishth S, Baayen H, Bates D. Balancing Type I error and
1700   power in linear mixed models. Journal of Memory and Language. 2017;94: 305–315.
1701   doi:10.1016/j.jml.2017.01.001

1702   130. Schwarz G. Estimating the Dimension of a Model. Ann Statist. 1978;6.
1703   doi:10.1214/aos/1176344136

1704   131. Vonesh E, Chinchilli VM. Linear and Nonlinear Models for the Analysis of Repeated
1705   Measurements. 0 ed. CRC Press; 1996. doi:10.1201/9781482293272

1706   132. Gurka MJ. Selecting the Best Linear Mixed Model Under REML. The American
1707   Statistician. 2006;60: 19–26. doi:10.1198/000313006X90396

1708   133. Yang Y, Piantadosi ST. One model for the learning of language. Proc Natl Acad Sci
1709   USA. 2022;119: e2021865119. doi:10.1073/pnas.2021865119

1710   134. Wolff JG. Language acquisition, data compression and generalization. Language &
1711   Communication. 1982;2: 57–89. doi:10.1016/0271-5309(82)90035-0

1712   135. Takahira R, Tanaka-Ishii K, Dębowski Ł. Entropy Rate Estimates for Natural
1713   Language—A New Extrapolation of Compressed Large-Scale Corpora. Entropy.
1714   2016;18: 364. doi:10.3390/e18100364

1715   136. Shannon CE. A Mathematical Theory of Communication. Bell System Technical
1716   Journal. 1948;27: 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

1717   137. Aceves P, Evans JA. Human languages with greater information density have higher
1718   communication speed but lower conversation breadth. Nat Hum Behav. 2024 [cited 20
1719   Feb 2024]. doi:10.1038/s41562-024-01815-w

1720   138. Koplenig A. Corpus size strongly matters when analysing word frequency distributions.
1721   2024. doi:10.31219/osf.io/p5nhd

1722   139. Baayen RH. Word Frequency Distributions. Dordrecht: Kluwer Academic Publishers;
1723   2001.

1724   140. Tweedie FJ, Baayen RH. How Variable May a Constant be? Measures of Lexical
1725   Richness in Perspective. Computers and the Humanities. 1998;32: 323–352.

1726   141. Koplenig A, Wolfer S, Müller-Spitzer C. Studying Lexical Dynamics and Language
1727   Change via Generalized Entropies: The Problem of Sample Size. Entropy. 2019;21.
1728   doi:10.3390/e21050464

1729   142. Aceves P, Evans J. Conditional Word Count and Huffman Code Size are Two Sides of
1730   the Same Coin: Response to Koplenig. 2024. doi:10.31219/osf.io/4b7mc

1731   143. Lupyan G, Contreras Kallens P, Dale R. Information density as a predictor of
1732   communication dynamics. Trends in Cognitive Sciences. 2024;28: 489–491.
1733   doi:10.1016/j.tics.2024.03.012

1734   144. Schmitt AO, Herzel H. Estimating the Entropy of DNA Sequences. Journal of
1735   Theoretical Biology. 1997;188: 369–377. doi:10.1006/jtbi.1997.0493

1736   145. Bar-Hillel Y. An Examination of Information Theory. Philosophy of Science. 1955;22:
1737   86–105.

1738   146. Montemurro MA, Zanette DH. Towards the quantification of the semantic information
1739   encoded in written language. Advances in Complex Systems. 2010;13: 135–153.
1740   doi:10.1142/S0219525910002530

1741   147. Weaver W. Translation. In: Locke WN, Boothe AD, editors. Machine Translation of
1742   Languages. Cambridge, MA: MIT Press; 1949. pp. 15–23.

1743   148. Marjou X. OTEANN: Estimating the Transparency of Orthographies with an Artificial
1744   Neural Network. In: Vylomova E, Salesky E, Mielke S, Lapesa G, Kumar R,
1745   Hammarström H, et al., editors. Proceedings of the Third Workshop on Computational
1746   Typology and Multilingual NLP. Online: Association for Computational Linguistics;
1747   2021. pp. 1–9. doi:10.18653/v1/2021.sigtyp-1.1

1748   149. Salesky E, Chodroff E, Pimentel T, Wiesner M, Cotterell R, Black AW, et al. A Corpus
1749   for Large-Scale Phonetic Typology. Proceedings of the 58th Annual Meeting of the
1750   Association for Computational Linguistics. Online: Association for Computational
1751   Linguistics; 2020. pp. 4526–4546. doi:10.18653/v1/2020.acl-main.415

1752   150. Wray A, Grace GW. The consequences of talking to strangers: Evolutionary corollaries
1753   of socio-cultural influences on linguistic form. Lingua. 2007;117: 543–578.
1754   doi:10.1016/j.lingua.2005.05.005

1755   151. Lupyan G, Dale R. Why Are There Different Languages? The Role of Adaptation in
1756   Linguistic Diversity. TRENDS in Cognitive Science. 2016;20: 649–660.
1757   doi:10.1016/j.tics.2016.07.005

1758   152. Lupyan G, Dale R. Language Structure Is Partly Determined by Social Structure.
1759   O'Rourke D, editor. PLoS ONE. 2010;5: e8559. doi:10.1371/journal.pone.0008559

1760   153. Atkinson M, Kirby S, Smith K. Speaker Input Variability Does Not Explain Why
1761   Larger Populations Have Simpler Languages. Caldwell CA, editor. PLOS ONE.
1762   2015;10: e0129463. doi:10.1371/journal.pone.0129463

1763   154. Raviv L, Meyer A, Lev-Ari S. Larger communities create more systematic languages.
1764   Proceedings of the Royal Society B: Biological Sciences. 2019;286: 20191262.
1765   doi:10.1098/rspb.2019.1262

71

1766 155. Frank S, Smith K. Natural Population Growth Can Cause Language Simplification. In:
1767   Ravignani A, Barbieri C, Martins M, Flaherty M, Jadoul Y, Lattenkamp E, et al.,
1768   editors. The Evolution of Language: Proceedings of the 13th International Conference
1769   (EvoLang13). 2020. doi:10.17617/2.3190925

1770 156. Raviv L, Peckre LR, Boeckx C. What is simple is actually quite complex: A critical
1771   note on terminology in the domain of language and communication. Journal of
1772   Comparative Psychology. 2022 [cited 7 Dec 2022]. doi:10.1037/com0000328

1773

Supporting Information Captions

1775 S1 Text. Language modelling details

1776 S2 Text. Scaling σ_x

1777 S3 Text. Numeric IDs

1778 S4 Text. Testing for a potential systematic length bias

1779 S5 Text. Associations between and within languages

1780 S6 Text. The entropy-length trade-off is not trivial

1781 S7 Text. Controlling for the number of different words and characters

1782 S8 Text. Controlling for orthographic transparency/depth

1783 S9 Text. Alignment with the PBC

1784 S10 Text. Testing the trade-off on the phonemic level