PLOS COMPLEX SYSTEMS

RESEARCH ARTICLE

Human languages trade off complexity against efficiency

Alexander Koplenig*, Sascha Wolfer, Jan Oliver Rüdiger, Peter Meyer

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

* koplenig@ids-mannheim.de

Abstract

From a cross-linguistic perspective, language models are interesting 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. In this paper, we train different language models, from simple statistical models to advanced neural networks, on a database of 41 multilingual text collections comprising a wide variety of text types, which together include nearly 3 billion words across more than 6,500 documents in over 2,000 languages. We use the trained models to estimate entropy rates, a complexity measure derived from information theory. To compare entropy rates across both models and languages, we develop a quantitative approach that combines machine learning with semiparametric spatial filtering methods to account for both language- and document-specific characteristics, as well as phylogenetic and geographical language relationships. We first establish that entropy rate distributions are highly consistent across different language models, suggesting that the choice of model may have minimal impact on cross-linguistic investigations. On the basis of a much broader range of language models than in previous studies, we confirm results showing systematic differences in entropy rates, i.e. text complexity, across languages. These results challenge the long-held notion that all languages are equally complex. We then show that higher entropy rate tends to co-occur with shorter text length, and argue that this inverse relationship between complexity and length implies a compensatory mechanism whereby increased complexity is offset by increased efficiency. Finally, we introduce a multi-model multilevel inference approach to show that this complexity-efficiency trade-off is partly influenced by the social environment in which languages are used: languages spoken by larger communities tend to have higher entropy rates while using fewer symbols to encode messages.

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

OPEN ACCESS

Citation: Koplenig A, Wolfer S, Rüdiger JO, Meyer P (2025) Human languages trade off complexity against efficiency. PLOS Complex Syst 2(2): e0000032. https://doi.org/10.1371/journal.pcsy.0000032

Editor: Giulio Rossetti, Alessandro Faedo Institute of Science and Technology National Research Council: Consiglio Nazionale delle Ricerche Istituto di Scienza e Tecnologie dell'Informazione Alessandro Faedo, ITALY

Received: August 28, 2024

Accepted: December 3, 2024

Published: February 3, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcsy.0000032

Data Availability Statement: All data and code (Stata v18.0 and Python v3.6.8) needed to replicate our analyses are available at https://osf.io/xdwjc/. In addition, interactive results and visualisations are available online at https://www.owid.de/plus/tradeoffvis/.

Funding: We acknowledge support for computing resources from the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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.

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.

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 inverse relationship between complexity and length implies a compensatory mechanism, whereby increased complexity is offset by greater efficiency.
- 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.

1.2. Background and related work

Parallel corpora are valuable in a cross-linguistic context because they enable systematic comparisons of language processing across different languages [16]. By providing translations of the same text in multiple languages, parallel corpora allow researchers to study how LMs handle similar content across varying grammatical structures and lexicons. These datasets also facilitate the testing and understanding of linguistic laws, i.e., statistical patterns shared across human languages [17], or the examination of whether languages adapt to the geographical or sociodemographic environments in which they are learned and used [18–22]. Yet another idea–dating back to the work of Greenberg [23]–that was revived with the development of large parallel corpora [24] is to use parallel texts to classify and compare languages [16]. Examples of such cross-linguistic studies are [25–29]. However, the majority of the aforementioned studies are based on very peculiar text types, especially translations of the Bible and there are several important challenges that the use of the Bible as a parallel text source pose [20,30,31].

To address this limitation, we leveraged available corpora and multilingual text collections [32–35] and compiled a database of parallel texts comprising a large variety of different text types, e.g. religious texts, legalese texts, subtitles for various movies and talks, and machine translations. In addition, we added comparable corpora, i.e., texts that are not parallel but come from comparable sources and are therefore similar in content, again comprising very different text types/genres, e.g. newspaper texts, web crawls, Wikipedia articles, translation tables for system messages in the Ubuntu operating system, or translated example sentences from a free collaborative online database. Furthermore, we added information from the Crúbadán project [36] that aims at creating text corpora for a large number of (especially under-resourced) languages. In total, the compiled database contains 41 different multilingual corpora, comprising nearly 3 billion words or nearly 9 billion Unicode characters across more than 6,500 documents and covering over 2,000 languages. These languages are spoken as a native language by more than 90% of the world's population and represent almost half of all languages with a standardised written representation.

In a recent paper [37], we presented the first results based on this database. In this study, our primary focus was a cross-linguistic examination of language complexity, a topic that has garnered significant attention in linguistics and related fields over the past two decades–for an overview, see [38]. In our study, we quantitatively evaluated the so called equi-complexity hypothesis that suggests that all human languages, despite their diverse and varied nature, have the same level of overall complexity [39]. To overcome the difficulty of measuring overall language complexity [40], we leveraged information theory, an area of mathematics that links probability and communication [41] and provides notions of complexity that are both objective and theory-neutral [42]. To this end, we trained a simple statistical LM on each of our documents and statistically analysed the training process to infer the average per-symbol information content or entropy rate of each document, which can be interpreted as a measure of complexity [43,44]: the harder it is, on average, to predict upcoming text, the higher the entropy rate, the greater is the complexity of the text as a whole [10,45–47]. We argued the entropy rate can thus also be used to compare the complexity of different languages. We then statistically compared complexity rankings across different corpora by calculating correlation coefficients between the entropy rates across all possible pairs of multilingual corpora. For example, we correlated the entropy rate rankings derived from a corpus of movie subtitles in various languages with those from a similarly diverse corpus of Wikipedia sentences. This approach, applied comprehensively to all pairs among our 41 different multilingual corpora, makes it possible to assess the consistency of complexity rankings across various types of linguistic data. From an information-theoretic point of view, we showed that our results constitute evidence against the equi-complexity hypothesis: a language with high/low entropy rate in one corpus also tends to be more/less complex in another corpus.

1.3. Research Question and Scope

As higher complexity in language results in more demanding processing efforts, encompassing both language production and comprehension, our study [37] naturally leads to the question: *Why is there a trend towards increased complexity in certain languages*? In the present study we pursue this issue by providing evidence suggesting that languages with high entropy rates, which indicate greater complexity, require fewer symbols to encode messages. This points to a compensatory mechanism: higher complexity is balanced by increased efficiency, in terms of shorter message lengths.

The main purpose of this study is to present, discuss and evaluate evidence for such a complexity-efficiency trade-off, while also striving to enhance the following empirical and methodological aspects of our prior work.

First, in our previous paper we trained a rather simple statistical LM. However, as Baroni [48] pointed out, different LMs have uniquely structured internal architectures, and thus cannot be viewed as "blank slates". Instead, they should be regarded as algorithmic linguistic theories "encoding non-trivial structural priors facilitating language acquisition and processing" [48]. These architectural differences affect how LMs acquire and process language. For instance, n-gram models primarily rely on local context and assume fixed-length dependencies. To estimate the probability of the next symbol, n-gram models build a conditional probability distribution based on a limited context window that only takes into account the preceding few words (e.g. n = 2, 3, 4, 5) [2]. This restriction on context can lead to a loss in the model's ability to capture long-range dependencies that are essential for grammatical complexity in human languages. More advanced models, such as transformers and other neural network architectures, are equipped with mechanisms like self-attention that allow them to capture complex, long-range dependencies and richer contextual information from the input data [2]. These architectural differences enable different models to capture linguistic complexity at various levels, from basic statistical patterns to more nuanced syntactic and semantic relationships.

To extend our previous work, we thus train various types of LMs on our data, ranging from simple statistical n-gram models to state-of-the-art transformer models. This approach allows us to compare how models with varying levels of sophistication interpret and represent language complexity, providing a more comprehensive understanding of the phenomenon across models. Surprisingly, we show that, from a cross-linguistic perspective, the type of LM has relatively little impact on the obtained results.

Secondly, we improve and extend our prior work methodologically by developing a machine learning method that fully accounts for the relatedness of languages: when comparing languages, statistical challenges arise due to the fact that closely related languages often exhibit more similarities among themselves in various aspects than they do with more distantly related languages. Additionally, languages originating from the same regions often tend to be influenced by common factors, further complicating the analysis [49–51]. While we have included language family, macro-area and country as factors to account for the genealogical and geographic relatedness of languages in our prior paper, this approach ignores variation within language families and geographical units as pointed out in several recent studies [49–53]. To address this issue, we develop two quantitative approaches: (i) a semiparametric machine learning estimation method capable of simultaneously controlling for document- and language-specific characteristics while directly modelling potential effects due to phylogenetic relatedness and geographic proximity; (ii) a multi-model multilevel inference approach designed to test whether cross-linguistic outcomes are statistically associated with sociodemographic factors, while accounting for phylogenetic and spatial autocorrelation via the inclusion of random effects and slopes.

1.4. Structure of the paper

The structure of this paper is as follows: the next section introduces the multilingual database and details the procedures for compiling the text data (Sect. 2.1). This is followed by a description of the sociodemographic and linguistic variables considered in this study (Sect. 2.2). We then introduce the investigated LMs (Sect. 2.3) and describe how the textual data was pre-processed (Sect. 2.4). The methodology for estimating entropy is presented in Sect. 2.5. Sect. 2.6 is devoted to statistical methods. We first present a novel method for evaluating the similarity of entropy rates and length distributions across different corpora (Sect. 2.6.1), followed by a description of the multi-model inference approach used to examine the impact of the number language users on the entropy-length trade-off (Sect. 2.6.2).

In Sect. 3, we present our findings. The paper concludes with the "Discussion" section, where we evaluate potential limitations of our approach, examine the relevance of the complexity-efficiency trade-off, and outline directions for future research (Sect. 4).

All data and code (Stata v18.0 and Python v3.6.8) needed to replicate our analyses are available at https://osf.io/xdwjc/. In addition, interactive results and visualisations are available online at https://www.owid.de/plus/tradeoffvis/.

2. Materials and methods

Some material in this section is recycled from our prior publications [37,54], in accordance 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 glottal stop [33]. Another example is tone marking, which in languages such as Chinantec, Mazatec, or Nambikuára is represented by raised numbers indicating tones. For instance, in the Sochiapan Chinantec translation of the sentence "I am the God of Abraham," the text reads: "Jná^{13} bíh^{1} la^{32} Dió^{32} Juo^{13} Há^{2}bran^{21}" Automatically tokenising this sentence into words (see Sect. 2.4) would mistakenly split words into separate parts, e.g., "Há bran." For further information, see Bentz et al. [27], Appendix A.

Another distinctive feature of the PBC, compared to the other multilingual text collections in the corpus, is the usage of full vocalisation using tashkīl diacritics, as highlighted by Gutierrez-Vasques et al. [56]. In Arabic, short vowels are usually not written–most registers omit them, while some use diacritical marks to indicate their presence. For example, in Egyptian Arabic, the PBC encodes the word for "peace" as "salām" (سَلَام), marking the short vowels with diacritics. In other text collections, only the 'skeleton' letters "sl'm" (سلام) are provided, leaving the short vowels implied. As discussed further in Sect. 2.4, this lack of vowel representation complicates computational and quantitative analyses. The quality and consistency of the PBC in encoding such features enhances the reliability of our cross-linguistic results. For further information regarding text pre-processing, see Sect. 2.4.

The two parallel text collections *WatchtowerV1* and *WatchtowerV2* are also part of the Parallel Bible Corpus, both containing translations of different introductory texts of the Jehovah's Witnesses' official web site [16] (*WatchtowerV1*: N_D = 142; N_L = 140; $\tilde{L}_W = 129{,}008$; $\tilde{L}_C = 659{,}563$; *WatchtowerV2*: N_D = 265; N_L = 260; $\tilde{L}_W = 7{,}194$; $\tilde{L}_C = 35{,}608$). The *Quran* collection consists of parallel translations of the central text of the Islam downloaded from http://tanzil.net/trans/ (accessed 4/30/20, N_D = 43; N_L = 43; $\tilde{L}_W = 182{,}950$; $\tilde{L}_C = 860{,}590$).

Web crawls. The *GlobalVoices* comparable collection consists 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; $\tilde{L}_W = 196{,}899$; $\tilde{L}_C = 1{,}119{,}551$),

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

https://doi.org/10.1371/journal.pcsy.0000032.t001

(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 text. 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 ($N_D = 6$; $N_L = 6$; $\tilde{L}_W = 341{,}723{,}872$; $\tilde{L}_C = 879{,}903{,}168$).

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.

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 09/04/19) to machine translate a short passage excerpt from the book "Aldono al la Dua Libro de l' Lingvo Internacia" in the constructed language Esperanto, written by its inventor L.L. Zamenhof (downloaded from http://esperanto.davidgsimpson.com/librejo/index.html on 09/04/19) into 102 languages (see [37] for the translated passage; $N_D = 102$; $N_L = 102$; $\tilde{L}_W = 716$; $\tilde{L}_C = 3{,}875$). The two comparable corpora *Tatoeba V1* and *Tatoeba V2* are compiled based on *Tatoeba*, a collaborative online platform that makes available sentences translated into different languages. Raw data were downloaded from https://tatoeba.org/deu/downloads (accessed 4/30/20; *V1*: $N_D = 183$; $N_L = 183$; $\tilde{L}_W = 656$; $\tilde{L}_C = 3{,}034$; *V2*: $N_D = 123$; $N_L = 123$; $\tilde{L}_W = 3{,}438$; $\tilde{L}_C = 15{,}710$).

Word list. To generate the word list data, we downloaded all available lists from the Crúbadán project [36] from http://crubadan.org/files/ (accessed 4/30/20). In total, we arrived at 2,216 word frequency lists for a total of $N_L = 1{,}943$ different languages ($\tilde{L}_W = 101{,}079$).

2.1.2. Overview of the database. Fig 1 displays a map highlighting the geographical distribution of languages for the compiled multilingual database. The figure reveals an imbalance at the language level within the database: over 100 languages have at least 10 documents, but approximately 75% of languages have fewer than four documents. This scarcity reflects the limited electronic availability of documents in languages spoken by smaller populations [36]. This is exemplified by the contrast in median speaker numbers: while the median for all non-extinct languages documented by Ethnologue stands at 8,000 [57], the median for languages represented with at least one document in our database is significantly higher, at 30,000. In addition, the majority of our text collections contain a comparatively small number of individual documents, with a median of 40 documents per corpus. This limited size can be attributed to specific reasons in certain cases; for example, the *EUconst* collection is naturally restricted to translations of the European Constitution into the official languages of the European Union. In contrast, for other collections such as the subtitle corpora, translations into further

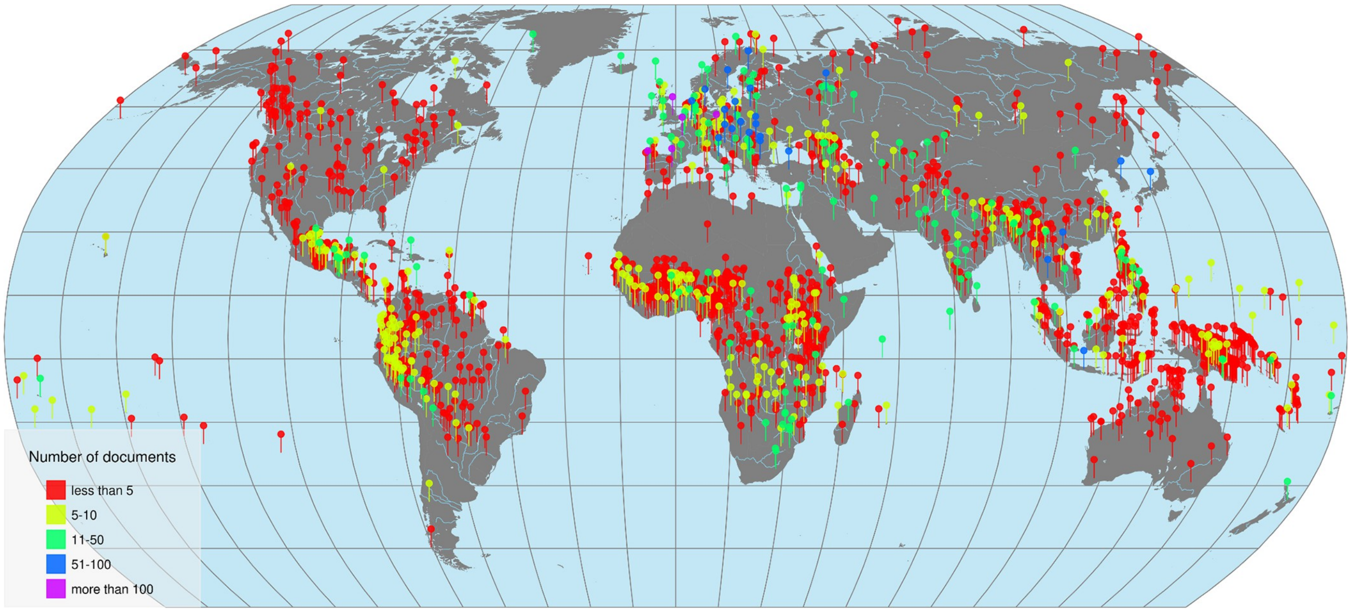


Fig 1. Global distribution of collected documents per language. Approximately 76% of languages have fewer than five documents. On the other side of this spectrum, over 160 languages have more than 10 documents. This imbalance reflects the limited electronic availability of documents in languages spoken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

https://doi.org/10.1371/journal.pcsy.0000032.g001

languages were not available when we compiled the database. On the other side of the spectrum, we have 11 multilingual text collections that consist of more than 100 different documents. As described above, documents are rather short, e.g. 25% of the documents are below 14,575 characters or 3,181 words. However, 200 documents are longer than 1 million characters, 49 documents are longer than 10 million characters and the longest documents are several hundred million words and more than a billion characters long.

In what follows we statistically compare the structure found in smaller corpora (i.e., those consisting of shorter documents and/or a limited number of available documents) with the structure found in larger corpora (i.e., those consisting of longer documents and/or data points for many languages). The idea is that if the results from both smaller and larger corpora align, this strengthens the claim that these results are not merely artefacts resulting from database bias. Additionally, we include control covariates, such as the number of corpora per language, to account for the unbalanced nature of our database, as described in the next section.

2.2. Sociodemographic and linguistic variables

Information on speaker population size, corpus, language family, language (identified by its *ISO*-639-3 code), macro-area, writing script, speaker population size, longitude and latitude are taken from [37]. Writing script refers to the writing system in which the corresponding document is written in. In total, we have information for $N_D = 6{,}513$ different documents using 51 different writing scripts. The vast majority ($N_D = 5{,}183$ or 79.58%) of our documents are written in Latin script, followed by Cyrillic ($N_D = 463$ or 7.11%), Arabic ($N_D = 156$ or 2.40%) and Devanagari ($N_D = 98$ or 1.50%). Expanded Graded Intergenerational Disruption Scale (EGIDS) level information was initially sourced from [58], which is reported in Glottolog [59] (v4.2.1). Country is defined by Ethnologue [60,61] as the primary country/country of origin of the language in question [57]. To ensure completeness, we manually supplemented missing data from [58] by cross-referencing with Glottolog and Ethnologue. The EGIDS level serves as a measure of a language's endangerment status [62]. We use the EGIDS level as a covariate to control for potential translation effects [63,64], as languages with lower EGIDS levels are presumably more likely to be used as source languages, while languages with higher EGIDS levels are presumably more likely to be used as target languages. For example, an EGIDS level of 0 (labelled "International") pertains to the six official United Nations languages: Arabic, Chinese, English, French, Russian, and Spanish. On the other hand, languages with values of five and above pertain to languages that are not used in formal education, mass media or by the government, and they may consequently be more susceptible to (more) pronounced "translationese" influences [64]. With a similar logic in mind and to account for the unbalancedness of our database (see Sect. 2.1.2), we also consider the number of corpora with at least one available document per language as an additional control variable in what follows.

Further information regarding the classification of languages into macro-family and sub-family are taken from [65]. We manually added information for languages that was missing by using publicly available genealogical classifications (see the script 'prepare_language_info.do' available at https://osf.io/tkgph/ for details). Classifications in [65] are given as comma-separated values. We define the first value as the macro-family and the second one as the sub-family, e.g. for the language "Ghotuo" the classification is "Niger-Congo, Atlantic-Congo, Volta-Congo, Benue-Congo, Edoid, North-Central, Ghotuo-Uneme-Yekhee", so the macro-family is "Niger-Congo" and the sub-family is "Atlantic-Congo". Additionally, we use a phylogenetic similarity matrix also provided by [65] that is based on word lists from the Automated Similarity Judgment Program (ASJP) [66]. Information on the number of countries in which each language is spoken was sourced from Glottolog (v4.2.1). We manually supplemented missing data by cross-referencing with Ethnologue [60,61]. The rationale behind considering this variable as a potential covariate is to account for the varying degrees of pluricentrism [67]. For instance, languages such as Chinese or Spanish are spoken in several countries and may therefore have different codified standard forms.

For further information and a discussion of potential caveats and problems regarding the assignment of environmental variables to individual languages in order to reflect local grouping structure, see [68,69].

2.3. Language models

We use general-purpose data compression algorithms, taking advantage of the fact that language modelling and lossless compression are essentially equivalent [70–72]. All data compression algorithms consist of a model and a coder [14]. Our focus is on the class of (lossless) compressors where the algorithm uses training data to estimate a model, i.e., a conditional probability distribution that can be used to generate predictions about upcoming symbols. To perform compression, the predicted probabilities are then used to encode symbols using a technique called arithmetic encoding [73].

The seven LMs that we investigate are summarised in Table 2. In what follows, further details are given for each language model. Prediction by partial matching (PPM) [75,85] is a dynamic and adaptive variable-order n-gram LM. The algorithm assumes the Markov property: to predict the next symbol, the algorithm uses the last *o* immediately preceding symbols. For **PPM2**, we set *o* to 2, i.e., the last 2 symbols are used as context to generate predictions. For **PPM6**, we set *o* to 6. In both cases, the level of compression is set to maximum and the size of used memory is set to 2,000 megabytes. **PAQ** can be described as a weighted combination of predictions from a large number of models, where the individual models are combined using a gated linear network [14,78,79,81]. The network has a single layer with 552 input nodes and 3,080 input weights. The model has a total of ~1.7 million weights, but due to a sparse updating scheme which leads to faster compression and decompression, the effective number of parameters used in training is significantly lower. Only 552·7 = 3,864 weights are updated for each bit of data. We use version PAQ8o and set the compression level to maximum, requiring 1,712 megabytes of memory.

NNCP [83] is a lossless data compressor that is based on the Transformer XL model defined in [86]. Modifications to the original Transformer XL model and algorithmic details are provided in [87,88]. As for LSTM, the Adam optimiser is used and we use the "encode only"

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

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)

https://doi.org/10.1371/journal.pcsy.0000032.t002

mode. For **TRFsmall** (version 3.1), we use the default options with four layers and a model dimension of 256, resulting in a total number of ~2.24 million parameters. For **TRFmed** (version 3.2), we set the number of layers to 12 and the model dimension to 512, resulting in a total number of ~19.1 million parameters. For **TRFbig** (version 3.2), we use the available "enwik9" profile that sets the number of layers to 20 and the model dimension to 1,024, resulting in a total number of ~279 million parameters.

In addition, NNCP offers compression based on a Long Short-Term Memory deep neural network (**LSTM**) [82]. We use four layers of LSTM cells. The network is trained using truncated-like backpropagation [87,89] and Adam optimisation is used to update network weights [90]. We do not use a text pre-processor or tokeniser and we use the faster "encode only" mode (the output cannot be decompressed, but the compression itself is still lossless). The total number of parameters is ~3.93 million parameters.

In addition, when discussing the relevance of the complexity-efficiency trade-off (Sect. 4), we use OpenAI's GPT-2 model [91] with ~1.5 billion parameters, as implemented in the Hugging Face library [92].

2.4. Text pre-processing and information encoding units

Each document is tokenised and Unicode normalised where necessary. All uppercase characters are lowered based on the closest language-specific International Organization for Standardization (ISO) code. Unless otherwise specified in [37], the word-break algorithm of the *International Components for Unicode* library [93; Annex #29] was used to detect word boundaries (in texts without spaces or marks between words, a dictionary lookup method is used by the algorithm [94]). More details regarding each individual text collection can be found in [37].

Following [27], we represent a text κ as a random variable that is created by drawing (with replacement) from a set of symbol types $\mathcal{V} = \{s_1, s_2, s_3, \ldots, s_V\}$, where V is the number of symbol types, i.e., $V = |\mathcal{V}|$. Correspondingly, a symbol token is any reoccurrence of a symbol type [27]. In what follows, we estimate the relevant information-theoretic quantities for the following information encoding units/symbol types: (i) (Unicode) characters, (ii) words and (iii) sub-word units. For (iii), we apply byte pair encoding (BPE) [95–97] to split words into one or several units and then train the different LMs over the resulting sequences of sub-word units. We follow [95] and set the number of BPE merges to $0.4 \cdot V$.

Note that we neither remove spaces nor punctuation at the character level, nor before applying BPE, since punctuation can provide reliable information regarding stylistic features and differences between authors [98]. Correspondingly, on the word level, each punctuation mark is treated as a separate token, except for the dash, which ensures that dash-separated words (e.g., 'presidency-in-office') are treated as a single word type. After tokenisation into words/sub-word units, each word/BPE type is replaced by one unique Unicode symbol. The different compression algorithms are then used to compress the resulting symbol sequence. On the BPE level, we also compress the mapping of sub-word units to 4-byte Unicode symbols.

Limitations & caveats. It is important to acknowledge several caveats when interpreting cross-linguistic results based on the quantitative analysis of written language text at both the character and word levels [27,95,99–101]. First, our analysis is based exclusively on corpora of written language, which begs the question whether it applies to human languages in general. In this context, it is important to note that "written language cannot be regarded as *merely* the transference of spoken language to another medium", as John Lyons put it in his classical introduction to theoretical linguistics [102]. Indeed, there is a long-standing strand of research that recognises written language as a distinct linguistic object worthy of independent study, backed up by a large body of evidence against a mere derivative nature of written language, including evolutionary and biological aspects [103–105]. Obviously, the texts in the PBC (see Sect. 2.1.1) are one prime example of texts that are not parasitic on prior oral language production. From this stance, our results clearly apply to written languages, seen as more or less 'autonomous' communication systems.

While we do not have analysed genuine oral language data due to the lack of sufficiently large and phonemically annotated corpora, it is a valid question whether our results remain valid if we treat the texts in our corpora as representing spoken language, which basically means that we interpret grapheme sequences as standing in for phoneme sequences. A notable challenge arises from the differences in the mapping between phonemes and graphemes across languages [64,99]. For instance, languages with deep orthographies like English have inconsistent phoneme-grapheme correspondences (e.g., "ough" in "thought" vs. "through" vs. "dough" rendering /ɔː/, /uː/, and /əʊ/ phonemes, respectively), while languages with shallow orthographies like Spanish have more systematic correspondences (e.g., "a" as in "casa" always corresponds to /a/) [96]. Another example is the difference in the representation of the phoneme /tʃ/ between Czech and German: Czech uses a single character ("č"), while German 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 [106]. However, Geertzen et al. [107] demonstrate that the computational approach to "word-hood" we employ may be sufficiently well-suited within an information-theoretic and compression-based quantitative framework for capturing linguistic regularities in cross-linguistic analyses. Moreover, the substantial effort and typologically informed curation of the PBC adds further reliability to our cross-linguistic analyses (see Sect. 2.1.1). Therefore, if the patterns observed in the other multilingual text collections we investigate align with those from the PBC, this consistency would bolster confidence in the robustness and validity of our quantitative results.

Thirdly, languages show considerable variation with respect to the morphological complexity of words. Words in highly synthetic languages may, for example, contain long chains of derivational affixes, such as in Turkish *tan-ış-tır-ıl-a-ma-dık-lar-ın-dan-dır* ‘it is because they cannot be introduced to each other’ [108]. Verbs in particular may exhibit complex inflectional paradigms featuring incorporated object nouns and multiple tense, mood, and agreement markers, e.g. in Cayuga *t-ę-hęn-atat-hǫna’t-a-yę́:thw-ahs* ‘they will plant potatoes for each other’ [109]. Importantly, the status of such long morpheme sequences as single words is, in general, not controversial. In contrast, analytical languages like modern English express similar grammatical elements with multiple words, showing little or no inflection. This means that the number of grammatically possible words differs by several orders of magnitude between languages. This has a significant impact on word-level quantitative analyses, where any two different words are treated as unanalysed, unrelated entities. In addition, the number of words required to express the same propositional content varies significantly across languages.

To further address the aforementioned challenges, we compute our quantitative estimates at multiple levels of linguistic structure: characters, words, and the supra-character, sub-word level using BPE [95,96]. BPE is a sub-word segmentation technique that iteratively merges the most frequent pairs of characters or character sequences, creating sub-word units that can capture many meaningful linguistic patterns. This method is crucial in modern language modelling, as it efficiently handles morphological variation and rare words, enhancing model performance across diverse languages. Moreover, BPE’s ability to extract language-specific sub-word patterns from raw text makes it particularly valuable in cross-linguistic studies, as it reveals structural differences and encodes features that align with those described in traditional linguistic typology, as recently discussed in-depth by Gutierrez-Vasques et al. [56]. By comparing language-specific information-theoretic estimates across symbolic levels, languages, and corpora, we can assess the consistency of results. If these estimates, derived from different symbolic levels and corpora with varying qualities for cross-linguistic studies, point in the same direction, this would strengthen the validity of our quantitative findings.

Additionally, as described in Sect. 2.6, we statistically account for potential confounding factors by including the type of writing script as a covariate in our analyses. In the multi-model multilevel analyses (Sect. 2.6.2), we also include random intercepts for macro-family, sub-family and–crucially–the language itself. These controls are expected to help mitigate some of the issues outlined above. Moreover, in Sect. 4, we discuss the results of several additional analyses presented in the Supporting Information, which further explore how robust our results are with respect to the potential sources of influence described above.

Nevertheless, we wish to emphasise that, given the scope of this study–encompassing several hundred languages and a wide variety of writing systems, each with its own distinct and often idiosyncratic conventions for representing language in written form–it is difficult, if not impossible, to completely rule out the possibility that these cross-linguistic variations 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]:

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

where $p\left(s_j\right) = \frac{f_j}{\sum_{j=1}^{V} f_j}$ 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]:

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

where $t_1^N = t_1, t_2, \ldots, t_N$ 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 $t_1,t_2,\ldots,t_{N-1}$ 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]:

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

In analogy to *H(κ)*, the entropy rate $h(\kappa)$ can be informally understood as the average number of guesses that are needed to guess the next symbol of a sequence and thus incorporating the notion that prediction and understanding are intimately related [14,72,110,111]. Information can then be defined as any kind of knowledge that, when in your possession, allows you to make predictions with greater accuracy than mere chance [112,113]. Thus, h encompasses complexity from various linguistic sub-domains, since any form of linguistic (e.g. grammatical, phonological, lexical, pragmatic) or non-linguistic (e.g. world) knowledge will help a reader or listener to predict upcoming linguistic material more accurately and will therefore reduce h [10]. This implies that when using LMs trained on written text to draw conclusions about languages, there is yet another important caveat that must be emphasised: not all information is encoded in written language. Key aspects such as extra-linguistic information, including prosody–i.e., the supra-segmental features of speech such as pitch, loudness, and tempo–are not fully captured. While Wolf et al. [114] show that much of the prosodic information can be inferred from the words and surrounding context, they also demonstrate that prosodic features cannot be entirely predicted from text alone. This suggests that prosody carries information that extends beyond the written word, contributing additional layers of knowledge.

Since the probability distribution for any natural language is unknown [14,110], we use the data compression algorithms described above to estimate $h(\kappa)$. Per LM, the entropy rate estimate is computed (roughly speaking) as the number of bits per symbol in the compressed text:

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

where $K^{LM}(\kappa) = R^{LM}(\kappa) – R^{LM}(\kappa_{train})$. Here, $R^{LM}(\kappa)$ denotes the number of bits that are needed by the LM to compress κ, κ_{train} represents the first half of κ, $L(\kappa)$ represents the length of the second half of κ in words on the level of words or, both on the character and the BPE level, in Unicode characters. Note that on the BPE level we also compress the mapping of unique symbols to 4-byte Unicode symbols mentioned above and add the resulting compressed lengths to $R^{LM}(\kappa)$ and $R^{LM}(\kappa_{train})$. Further note that $\hat{h}^{LM}$ is directly related to the quantity perplexity that is often used in NLP to measure the quality of a language model, where perplexity is defined as $2^{\hat{h}^{LM}}$ [42]. We use this relationship to also choose the LM that achieves the lowest perplexity on the test data:

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

where $\mathcal{L}$ denotes the set of different LMs, i.e., $\mathcal{L}$ = {PPM2, PPM6, PAQ, LSTM, TRFsmall, 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. We now take κ to be a corpus that consists of individual texts κ_i, where i denotes 1,. . ., I different languages. (For brevity, we omit the superscript and the hat in what follows.) However, we compute the correlation coefficients described below for both the best LM ($\hat{h}^{best}$ and $\hat{K}^{best}$) and for each individual LM ($\hat{h}^{LM}$ and $\hat{K}^{LM}$).

The entropy estimate for κ_i is denoted as $h_{\varsigma}(\kappa_i)$, where ς denotes one of three information encoding units (words, characters, BPE), likewise for $K_{\varsigma}(\kappa_i)$ and $L_{\varsigma}(\kappa_i)$. $h_{\varsigma}(\kappa_i)$ and $K_{\varsigma}(\kappa)$ are computed on all three levels, while $L_{\varsigma}(\kappa)$ is computed on either the word or the character level. Note that for languages with more than one available translation/document in a corpus, all quantities are averaged.

To evaluate the (dis-)similarity of entropy/length distributions across corpora and test for a potential trade-off between entropy and length, we first compute the pairwise Pearson correlation $\rho[\nu'_{\varsigma}(\kappa), \nu''_{\varsigma}(\iota)]$ for all corpus pairs (κ,ι) where ν'_{ς} denotes either h, L or K on the encoding level ς. Likewise for ν''_{ς}. In addition, we also consider H computed on the basis of the Crúbadán word frequency information (Eq 1), denoted as H_{Cr} in what follows. Both $\nu'_{\varsigma}(\kappa)$ and $\nu''_{\varsigma}(\iota)$ are logged. Pearson correlations range from -1 to 1, with higher absolute values indicating stronger associations. Positive values reflect positive associations, while negative values reflect negative associations between distributions. This allows us to assess both similarities and dissimilarities between entropy and length distributions across corpora.

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

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

where $\mathbf{y}$ is the $n \times 1$ vector of observed values, either $\nu'_{\varsigma}(\kappa)$ or $\nu''_{\varsigma}(\iota)$, n denotes the number of languages that are available in both κ and ι, $\mathbf{X}$ is the $n \times p$ design matrix of p covariates including a $n \times 1$ vector of ones for the intercept, $\boldsymbol{\beta}$ is the corresponding $p \times 1$ vector of coefficients and ϵ is the $n \times 1$ vector of residuals. We assume that ϵ_i are identically distributed with $E(\epsilon_i)$ = 0 with variance σ^2. Importantly, we want to rule out potential autocorrelation among the residuals, i.e., we wish to test the following null hypothesis in what follows:

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

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

As potential control variables, we consider the EGIDS level, the (logged) speaker population size, the (logged) number of corpora and the (logged) number of countries in which the language is spoken. We also include a set of indicator variables for the levels (categories) of writing script. To avoid overfitting, scripts that were unique to a single language were grouped into a common category. For H_{Cr}, we additionally control for the number of words and the number of documents (both logged) on which the word frequency list is based, and a binary variable indicating whether the word frequency list is truncated to account for differences in the way different Crúbadán word lists were generated. Further information can be found in [37]. To select the relevant control variables from the candidate set, we use the lasso machine learning technique [115]. To choose the optimal value for the penalty parameter for each lasso, we use a heteroskedastic plugin estimator [116]. Languages with missing information on any of the control variables are excluded in each case. Let $\tilde{\mathbf{X}}$ denote a $n \times \tilde{p}$ matrix of $\tilde{p}$ covariates selected by the lasso. We then regress $\mathbf{y}$ on $\tilde{\mathbf{X}}$ and compute residuals denoted as $\tilde{\epsilon}$. To test H_0 from above (Eq 8), we compute the modified version of Moran's I [117] suggested by Kelejian and Prucha [118], written as:

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

where $\mathbf{W}$ denotes an $n \times n$ weighting matrix and tr represents the trace operator. For $\mathbf{W}$, we consider two inverse distance matrices: (i) to test for spatial autocorrelation, we construct an inverse distance matrix $\mathbf{W}_{\mathrm{G}}$ based on longitude and latitude information, (ii) to test for phylogenetic autocorrelation, we construct an inverse distance matrix $\mathbf{W}_{\mathrm{P}}$ based on a phylogenetic similarity matrix provided by [65]. In both cases, matrix elements are equal to the reciprocal of distance that are then normalised using spectral normalisation. We test for autocorrelation with (i) $\mathbf{W}_{\mathrm{G}}$ as input, (ii) $\mathbf{W}_{\mathrm{P}}$ as input and (iii), as explained below, both $\mathbf{W}_{\mathrm{G}}$ and $\mathbf{W}_{\mathrm{P}}$ as input. For brevity, we drop the subscript in what follows and describe our algorithmic approach for input matrix $\mathbf{W}$. Since $I^2 \sim X^2(1)$ [118], we test H_0 via a standard X^2-test with one degree of freedom. If $p < 0.05$, we extend our linear regression model (Eq 7) by a semiparametric filter [119–121] as:

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

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

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

where $\mathbf{1}$ represents a $n \times 1$ column vector of ones. The eigensystem decomposition of $\mathbf{M}$ generates n eigenvalues and n corresponding eigenvectors. The eigenvalues are then sorted in descending order, denoted as $\boldsymbol{\lambda} = (\lambda_1,\lambda_2,\lambda_3,...\lambda_n)$, so that the largest eigenvalue receives the subscript 1, the second largest eigenvalue receives the subscript 2 and so on. The corresponding set of eigenvectors can then be denoted as $\mathbf{E} = (\mathbf{E_1},\mathbf{E_2},\mathbf{E_3},...\mathbf{E}_n)$. We include $\mathbf{E_1}$ into $\mathbf{F}$ and let the lasso select a subset of control variables from $\mathbf{X}$. We then compute $\tilde{\epsilon}$ based on a regression of $\mathbf{y}$ on $\tilde{\mathbf{X}}$ and $\mathbf{F}$. After that, we perform the X^2-test again. If the test is still significant, we also include $\mathbf{E_2}$ into $\mathbf{F}$ and re-perform estimation. This iterative procedure is repeated until $p \geq .05$. [Note that in scenario (iii), where we simultaneously control for $\mathbf{W}_G$ and $\mathbf{W}_P$, we compute two sets of eigenvectors, denoted as $\mathbf{E}_{\boldsymbol{G}} = (\mathbf{E}_{\mathrm{G},1},\mathbf{E}_{\mathrm{G},2},\mathbf{E}_{\mathrm{G},3},...\mathbf{E}_{G,n})$, and $\mathbf{E}_{\boldsymbol{P}} = (\mathbf{E}_{\mathrm{P},1},\mathbf{E}_{\mathrm{P},2},\mathbf{E}_{\mathrm{P},3},...\mathbf{E}_{P,n})$, and alternate the inclusion of eigenvectors into $\mathbf{F}$, i.e., we first include $\mathbf{E}_{\mathrm{G},1}$, then $\mathbf{E}_{\mathrm{P},1}$, and so on. Correspondingly, $I^2 \sim X^2(2)$]. Let $\tilde{\epsilon}_{\nu'_E}(\kappa)$ denote the resulting residuals for $\nu'_{\varsigma}(\kappa)$, likewise for $\nu''_{\varsigma}(\iota)$. This procedure ensures that there is no spatial/phylogenetic correlation among the resulting residuals, i.e., $\tilde{\epsilon}_{\nu'_{\varsigma}}(\kappa)$ and $\tilde{\epsilon}_{\nu''_{\varsigma}}(\iota)$. We then compute the Pearson correlation between those residuals and proceed as described above. The correlation coefficients per condition (none, geographical, phylogenetic and both) are denoted as ρ_{none}, ρ_{geo}, ρ_{phylo} and ρ_{both}.

2.6.2. Multi-model multilevel inference. To evaluate if the trade-off is moderated by the social environment in which languages are being used, we run separate multilevel effects models (MLEMs) with (i) h or K as the outcome on all three levels (words/characters/BPE) for all eight LMs (best, PPM2, PPM6, PAQ, LSTM, TRFsmall, TRFmed, TRFbig) and (ii) L as the outcome for words/characters. For N = 3,705 individual documents, we fit MLEMs of the form [122]:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \epsilon \tag{12}$$

where, in addition to the above, **Z** is a matrix of random predictors and **u** is a vector of random effects that are assumed to follow a normal distribution, with mean 0 and variance-covariance matrix **G**. The residual errors ϵ are assumed to follow a normal distribution, with mean 0 and variance matrix $\sigma^2\mathbf{I}$; $\mathbf{u} \perp \epsilon$. 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]:

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

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

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

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 use Akaike's information criterion (AIC) [127], where lower values indicate a better model, Δ_j = AIC_j–AIC_{min} where AIC_j denotes the AIC value computed for model j and AIC_{min} represents the minimum AIC value over all R models. Note that in models where x does not appear, $\beta_{x,j} \equiv 0$. On this basis, we compute an FMA estimator of the standard error (SE) as [124]:

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

where $\text{SE}(\beta_{x,j})$ denotes the estimated standard error of $\beta_{x,j}$ for model j. In models where x does not appear, we set $\text{SE}(\beta_{x,j}) \equiv 0$. To assess statistical significance, we compute a 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 Φ() denotes the cumulative standard normal distribution. In a similar vein, we compute a 95% confidence interval (95%-CI) as $\tilde{\beta}_x \pm \Phi^{-1}(0.975) \cdot \mathrm{SE}(\tilde{\beta}_x)$ where $\Phi^{-1}()$ denotes the inverse cumulative standard normal distribution.

Note that the Akaike weights ω_j can be "interpreted as approximate probabilities of each model being the actual best model, given the data" [124]. Thus, we can use the ω_j to estimate the relative importance of variable x, computed as [124]:

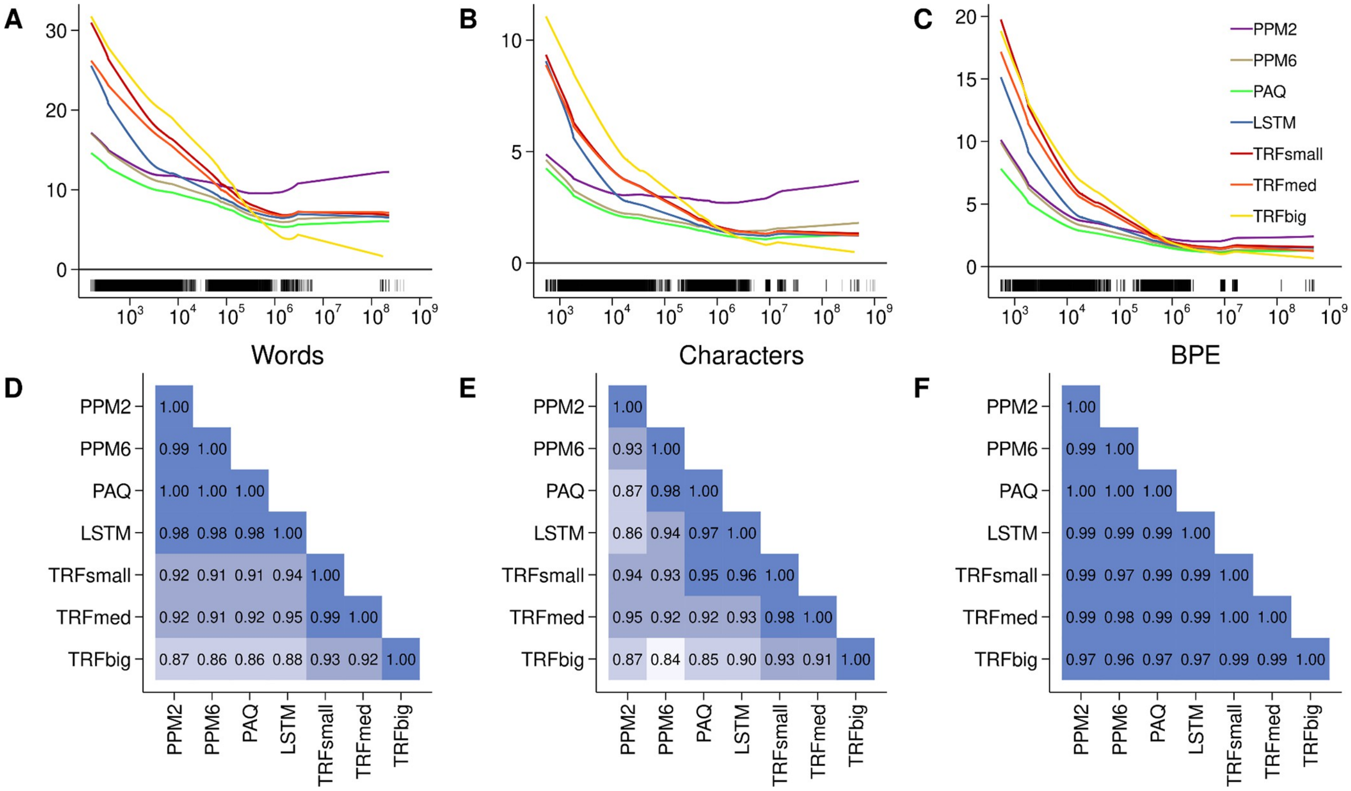
$$\sigma_x = \sum\nolimits_{j=1}^{R} \omega_j c_{x,j} \qquad (16)$$

where $c_{x,j}$ is a binary indicator that is equal to 1 if x is explicitly in model j and 0 otherwise [124]. The larger σ_x, the more important x. To put the value of σ_x into perspective, we show in S2 Text 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. Fig 2A–2C summarises the distribution of h as a function of length L for each level and each investigated LM across the 40 full-text multilingual text collections/corpora, totalling N_κ = 4,297 different documents (see Sect. 2.1 for details). The plots indicate that for most documents PAQ (mint line) turns out to be the best LM, i.e.,



D (Words)

	PPM2	PPM6	PAQ	LSTM	TRFsmall	TRFmed	TRFbig
PPM2	1.00						
PPM6	0.99	1.00					
PAQ	1.00	1.00	1.00				
LSTM	0.98	0.98	0.98	1.00			
TRFsmall	0.92	0.91	0.91	0.94	1.00		
TRFmed	0.92	0.91	0.92	0.95	0.99	1.00	
TRFbig	0.87	0.86	0.86	0.88	0.93	0.92	1.00

E (Characters)

	PPM2	PPM6	PAQ	LSTM	TRFsmall	TRFmed	TRFbig
PPM2	1.00						
PPM6	0.93	1.00					
PAQ	0.87	0.98	1.00				
LSTM	0.86	0.94	0.97	1.00			
TRFsmall	0.94	0.93	0.95	0.96	1.00		
TRFmed	0.95	0.92	0.92	0.93	0.98	1.00	
TRFbig	0.87	0.84	0.85	0.90	0.93	0.91	1.00

F (BPE)

	PPM2	PPM6	PAQ	LSTM	TRFsmall	TRFmed	TRFbig
PPM2	1.00						
PPM6	0.99	1.00					
PAQ	1.00	1.00	1.00				
LSTM	0.99	0.99	0.99	1.00			
TRFsmall	0.99	0.97	0.99	0.99	1.00		
TRFmed	0.99	0.98	0.99	0.99	1.00	1.00	
TRFbig	0.97	0.96	0.97	0.97	0.99	0.99	1.00

Fig 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, $\hat{h}^{LM}$ is plotted against L in characters. (**D**–**F**) Median unadjusted Pearson correlation, $\tilde{\rho}_{\text{none}}$, 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/.

https://doi.org/10.1371/journal.pcsy.0000032.g002

Table 3. Best LM per level. 1st column: LM. 2nd column: word level. 3rd column: character level. 4th column: BPE level. For each of N_κ = 4,297 different documents, each column lists, for a given 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 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%)

https://doi.org/10.1371/journal.pcsy.0000032.t003

$\hat{h}^{best}(\kappa) = \hat{h}^{PAQ}(\kappa)$ (see Sect. 2.5 for details) for most of the documents. For longer documents, the larger LMs (LSTM and the three Transformer LMs) achieve similar or lower entropy rates. Correspondingly, Table 3 shows that PAQ has the lowest h in more than 90% of the documents across all three symbolic levels.

To determine if entropy rate distributions are systematically affected by the choice of LM, we used the entropy estimates for the 40 full text corpora at each symbolic level to compute pairwise correlations ρ_{none} for each pair of LMs.

Fig 2D–2F presents the resulting pairwise relationships as correlation matrices. Each cell represents the median value of ρ_{none} for a pair of LMs. At each symbolic level, the statistical association between different LMs is remarkably strong. Even the lowest median value, observed at the character level between TRFbig and PPM6, is high, with a value of $\tilde{\rho}_{\text{none}} = 0.84$. To rule out the possibility that these associations mainly result from either language- and document-specific characteristics or the genealogical and geographic relatedness of languages (see Sect. 2.6.1 for details), Fig 3 visualises all four types of estimated correlation coefficients, i.e., ρ_{none}, and corresponding adjusted partial correlations, i.e., ρ_{geo}, ρ_{phylo} and ρ_{both} for all LM pairs across corpora and symbolic levels. The results indicate that the adjusted partial correlations point in the same direction as the unadjusted correlations.

The results presented in this section demonstrate that although there are differences in the performance of various LMs depending on document length (Fig 2A–2C), the resulting entropy rate distributions are remarkably consistent across LMs (Fig 3). This suggests that, from a cross-linguistic perspective, the choice of LM for investigating different languages may have a minimal impact.

3.1.2. Comparing languages. We proceed by comparing entropy rate distributions across corpora to determine whether a language that tends to be more complex in one corpus also tends to be more complex in another corpus. Given that the results from the previous section clearly indicate stability across different LMs, we will focus, for each corpus κ, on the estimates of its best model, i.e., $\hat{h}^{best}(\kappa)$ (see Sect. 2.5 for details). Interactive visualisations for each LM, i.e., $\hat{h}^{LM}(\kappa)$. are available at https://www.owid.de/plus/tradeoffvis/.

As outlined above, we evaluate the similarity of $\hat{h}^{best}$-distributions by computing ρ_{none}, ρ_{geo}, ρ_{phylo} and ρ_{both} *across* corpora and *across* symbolic levels. Per correlation type, we compute N_ρ = 7,254 individual correlations. Fig 4A demonstrates that entropy rate distributions are very similar across corpora and symbolic levels as indicated by a strong positive correlation between corpus pairs. The results remain stable when we control for language- and document-specific characteristics, as well as the genealogical and geographic relatedness of languages (see Sect. 2.6.1 for details).

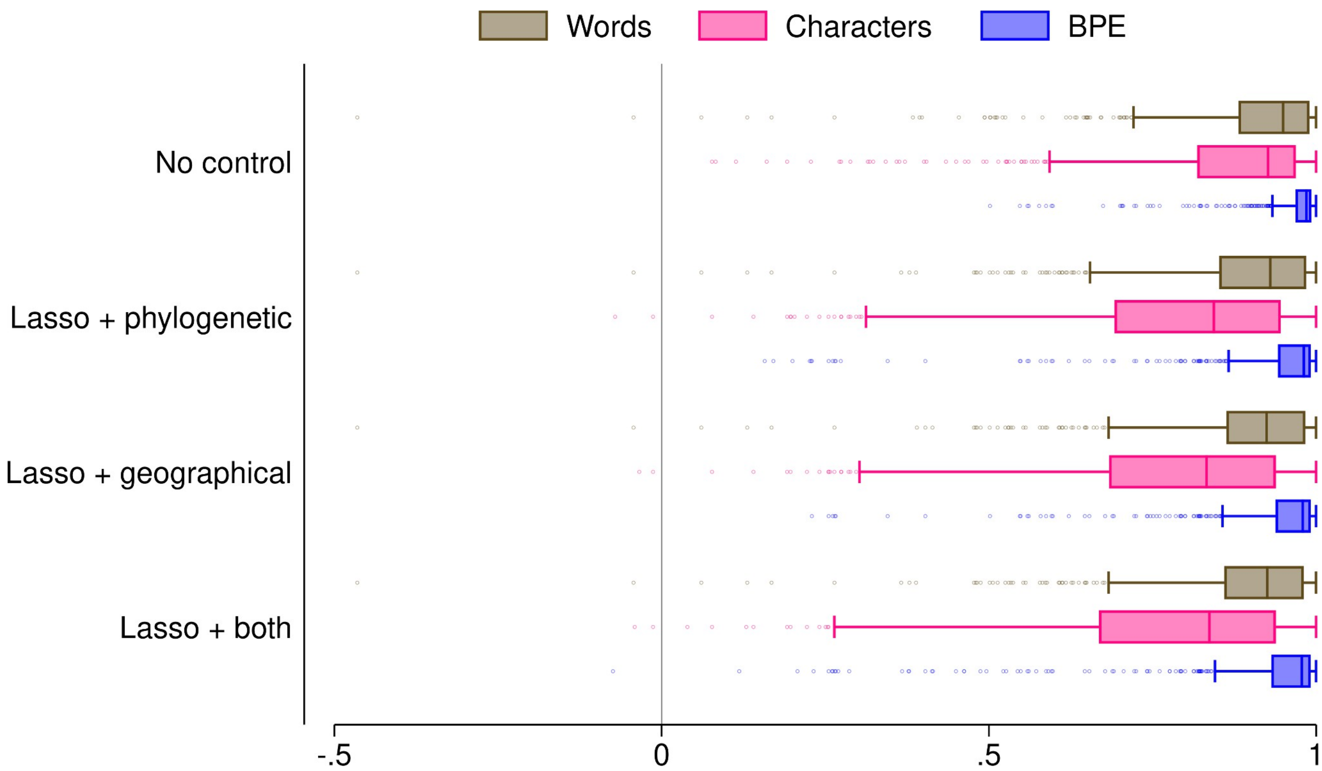


Fig 3. Distribution of pairwise correlations across LMs for each symbolic level. For each of the 40 full text corpora and per symbolic level, we compute the median value of the pairwise correlation between the estimated entropy rate distributions for each LM pair. We compute both unadjusted correlations,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

https://doi.org/10.1371/journal.pcsy.0000032.g003

Interactive LM-specific visualisations, available at https://www.owid.de/plus/tradeoffvis/ show that highly comparable patterns are obtained when estimating entropy rates for individual LMs.

Entropy rates are estimated as the ratio of the number of bits needed to compress the test data, K, to the length of the test data, i.e., L (cf. Eq 5). To further understand the above results, we repeated the analyses for both variables that are part of this ratio. Fig 4B reveals that while the results are more pronounced for h, the distributions for L are also very comparable across corpora and symbolic levels ($N_\rho = 3{,}160$). However, Fig 4C shows that the results are much weaker for distributions of $\hat{K}^{best}$ ($N_\rho = 7{,}140$). Since K is the product of h and L, this suggests a potential trade-off between h and L.

To investigate this possibility, we compute ρ_{none}, ρ_{geo}, ρ_{phylo} and ρ_{both} between $\hat{h}^{best}$ and L across corpora on all three symbolic levels. Table 4 demonstrates that on all three symbolic levels and for all four correlation types, there is a pronounced negative statistical association between entropy rate and length distribution. Again, our interactive visualisation tool shows highly comparable patterns for individual LMs. These results indicate the existence of a trade-off between both variables.

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. Fig 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 of the

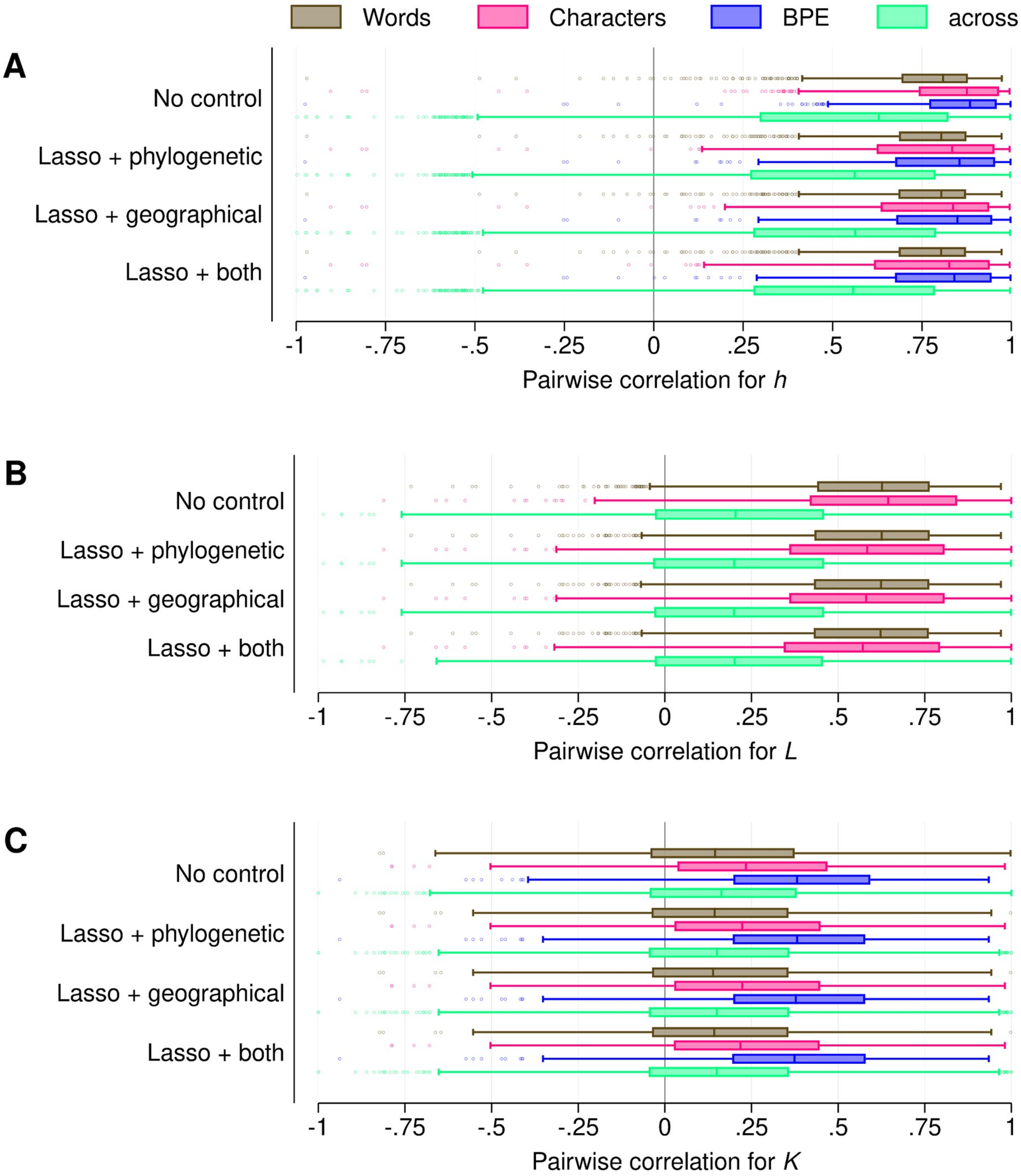


Fig 4. Distribution of pairwise correlations across corpora and symbolic levels. For each corpus pair, we compute both unadjusted correlations, i.e., ρ_{none}, and adjusted partial correlations, i.e., ρ_{geo}, ρ_{phylo} and ρ_{both} (see Sect. 2.6.1 for details). Correlations are computed both per symbolic 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 ($\boldsymbol{N_\rho} = 3{,}160$). (**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/.

https://doi.org/10.1371/journal.pcsy.0000032.g004

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

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

https://doi.org/10.1371/journal.pcsy.0000032.t004

variance is attributed to the trade-off between entropy and length in each case. Interactive LM-specific visualisations that are available at https://www.owid.de/plus/tradeoffvis/ illustrate that highly comparable patterns are observed when estimating entropy rates for different LMs.

Given that, from an information theoretic point of view, message length quantifies efficiency–the shorter the message the higher the efficiency [15]–we arrive at our main empirical result: human languages trade off complexity against efficiency. More explicitly, a higher average amount of choice/uncertainty per produced/received symbol is compensated by a shorter average message length.

3.1.3. Summary. In the last two sub-sections, we demonstrated that (i) entropy rate distributions are highly consistent across different LMs, suggesting that the choice of LM might have minimal impact on cross-linguistic investigations of the kind we presented here. We also showed that (ii) there is a pronounced trade-off between entropy rate and document length across different corpora, which implies that languages balance efficiency and complexity. This finding highlights a potentially fundamental principle in linguistic structure, where higher uncertainty per symbol is offset by shorter message lengths.

To bring these results together, we now compute fully adjusted partial correlations, ρ_{both} for all possible pairwise combinations 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). In total, N_{ρ} = 423,660 individual correlation coefficients were computed. Fig 6 shows that the findings point in the same direction as previously observed, confirming the consistency of entropy rate distributions and their trade-off with document length.

To emphasise this point, we extracted the first 80 factors from the factor analysis, which together account for ~90% of the variance in the correlation matrix. For each factor, we conducted separate linear regressions with the factor as the outcome and a binary indicator for the type of variable (1 = L vs. 0 = $\hat{h}^{LM}$ or H_{Cr}) as the predictor. For each factor, we then extracted the amount of explained variance (R^2) as a measure of model fit. Among all factors, R^2 is highest for the first factor, with R^2 = 76.14%. We then repeated the analyses using indicator variables for the investigated LMs as predictors. Again, R^2 is highest for the first factor, but with a much smaller value of R^2 = 6.44%). Fig 6 demonstrates that the first factor distinguishes between length and entropic variables but not between LMs. We further visually inspected all remaining factors, none of which separated the LMs, reinforcing the robustness of our findings across different models.

These results underscore that the negative statistical association between entropy and length in human languages is consistent across various LMs and corpora, suggesting that the trade-off between complexity and efficiency may reflect a fundamental property of human language.

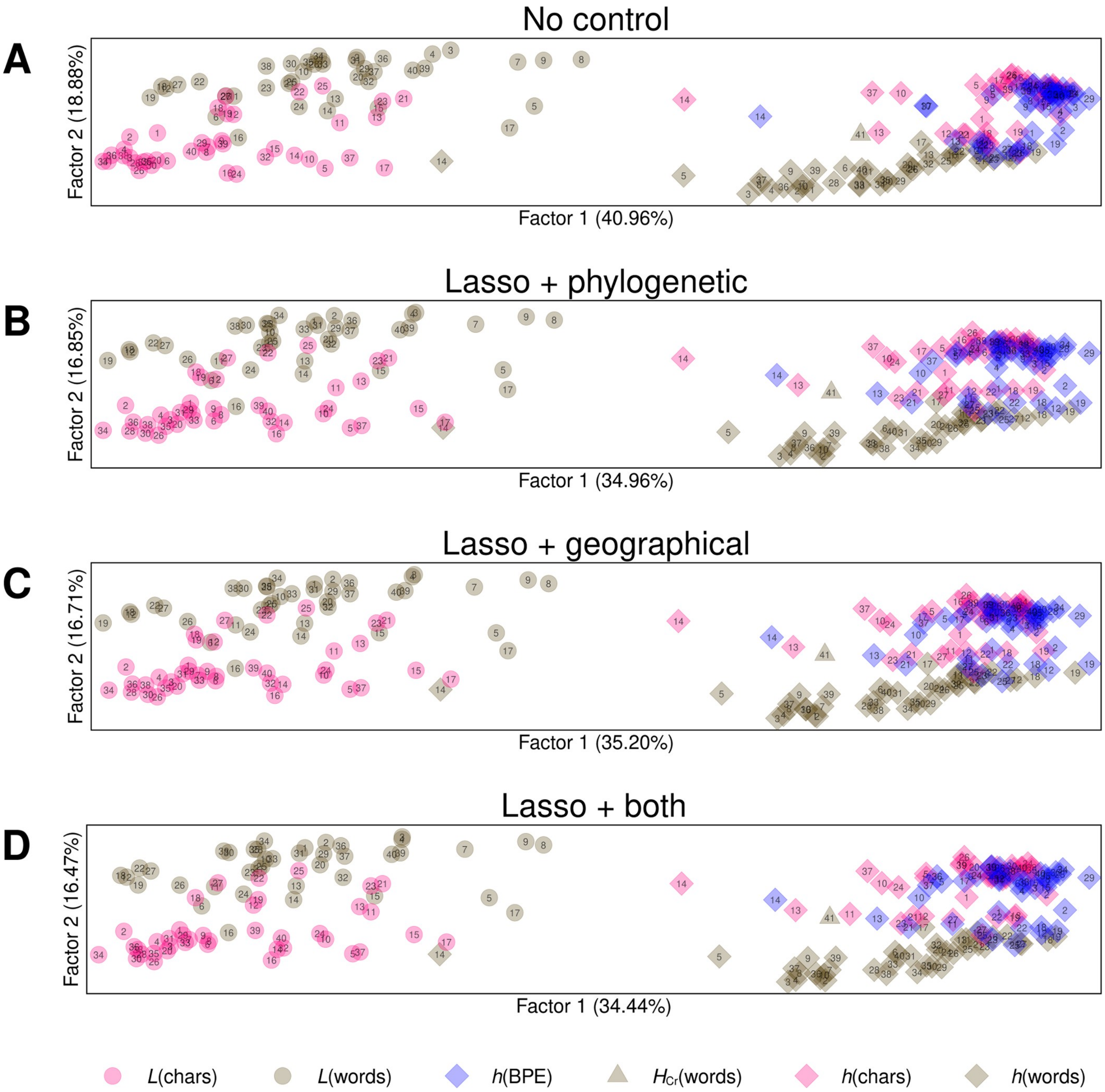


Fig 5. Comparing entropy and length across corpora. For each corpus pair and across symbolic levels, we compute both unadjusted correlations (ρ_{none}, **A**), and adjusted partial correlations (ρ_{phylo}, **B**; ρ_{geo}, **C**; ρ_{both}, **D**), see Sect. 2.6.1 for details. For each correlation type, N_ρ = 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 S3 Text). Interactive LM-specific visualisations are available at https://www.owid.de/plus/tradeoffvis/.

https://doi.org/10.1371/journal.pcsy.0000032.g005

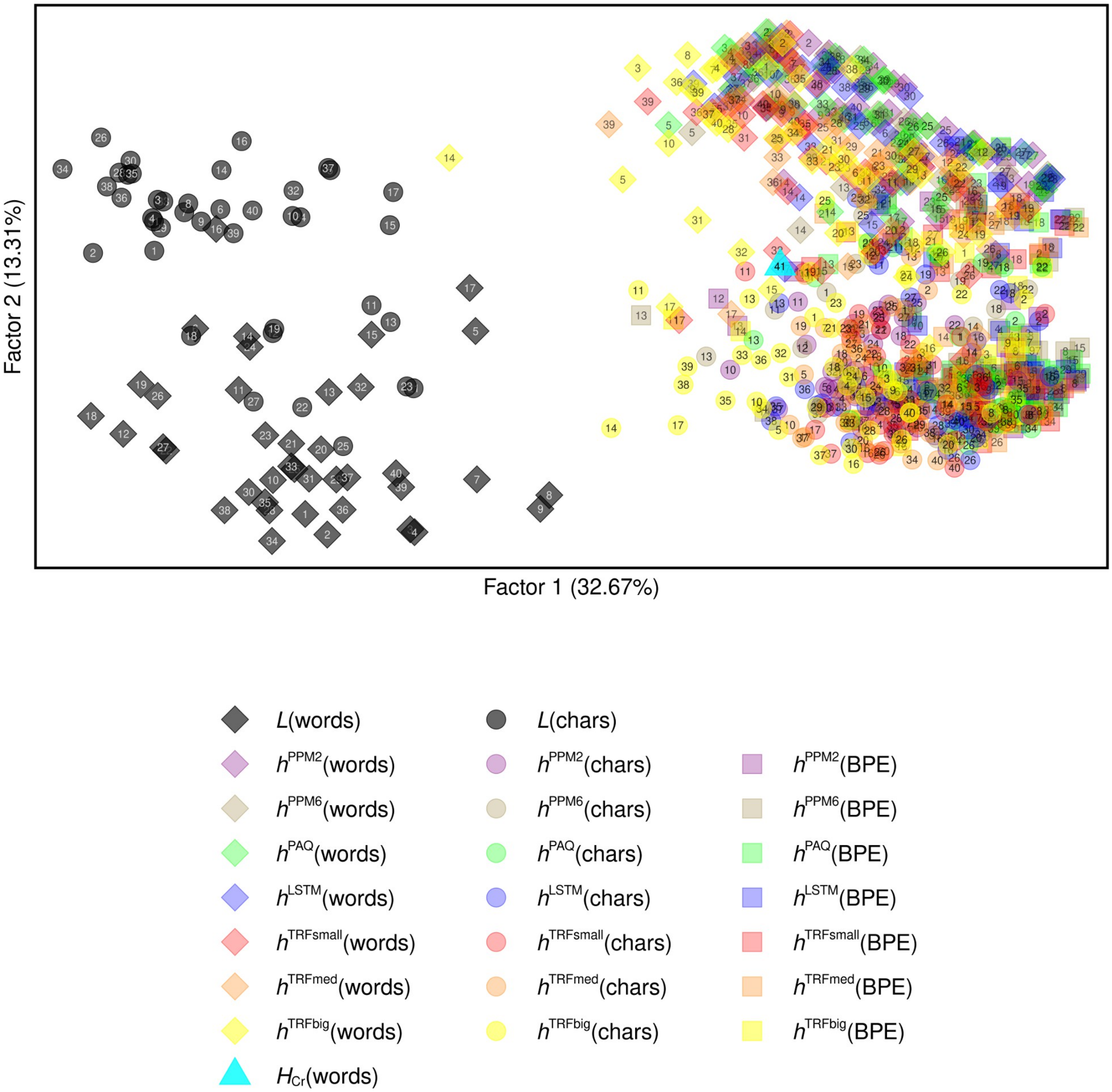


Fig 6. Evidence for a trade-off between complexity and efficiency across corpora and language models. We compute adjusted partial correlations, ρ_{both} (see Sect. 2.6.1 for details), for each combination of the 41 corpora, the three variables ($\hat{h}^{LM}$, H_{Cr}, L), the three symbolic levels (words, 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 analysed 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 S3 Text).

https://doi.org/10.1371/journal.pcsy.0000032.g006

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.

By including several fixed and random effects, we control for both (i) language- and document-specific characteristics and (ii) spatial and phylogenetic autocorrelation. For each outcome and symbolic level, we fit R = 8,192 candidate models. Fig 7A visualises the estimated relative importance, σ_x, for each variable. With the exception of macro-family for $\hat{K}^{best}$, all considered random effects have high relative importance for all three outcomes and across all three levels. Conversely, there is only weak evidence for the importance of most fixed effects, with the clear exception of speaker population size, which is maximally important–relative to the other variables–for both $\hat{h}^{best}$ and L, but not for $\hat{K}^{best}$.

To investigate if there is also evidence for a trade-off between entropy and length similar to the results presented in the previous section, Fig 7B plots the FMA-estimated effect, $\tilde{\beta}_x$, of speaker population size on each outcome per symbolic level. There is a positive and significant effect of population size on $\hat{h}^{best}$ across all three symbolic levels and a negative significant effect on L for both characters and BPE. For $\hat{K}^{best}$, there is no significant evidence of an effect of population size on any of the three symbolic levels. Table 5 lists the corresponding estimates for all investigated fixed effects, showing that the only consistent evidence of a noteworthy effect is for speaker population size on either entropy or length.

These results substantiate the evidence of a trade-off between entropy and length and indicate that languages with more speakers tend to have higher entropic values, i.e., are more complex, but also tend to produce shorter messages, i.e., are more efficient.

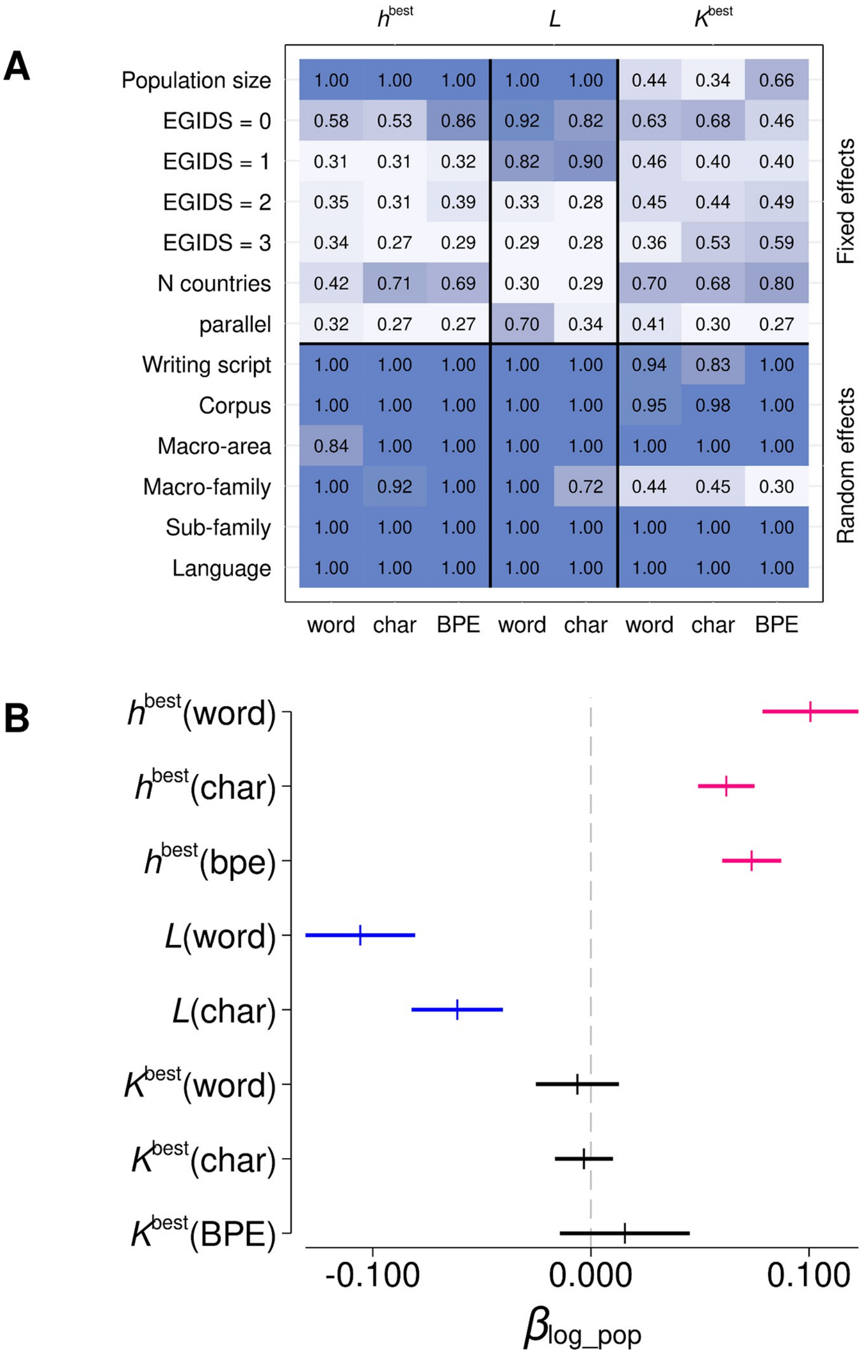
3.2.2. Random effects and slope models. As argued in the introduction (Sect. 1), an approach that focuses exclusively on random effects ignores variation within language families and geographical units [49–51,53,128]. We thus proceed by including random slopes, i.e., we allow the effect of population size to vary across different groups, representing deviations from the overall mean linear effect of speaker population size. The methodological question in this context is, which random effects should include random slopes?

We do not include a random slope for language since population size does not vary within languages. Due to geographic proximity and phylogenetic non-independence, it makes sense to include random slopes for macro-area, macro-family, and sub-family as geographic and phylogenetic structures can be critical for understanding language diversity and evolution, which justifies including these random slopes to account for shared inheritance and environmental factors [49].

There are, however, no a priori reasons for whether or not random slopes for writing script and corpus also make sense since both are not necessarily tied to population size or linguistic features in a way that would suggest significant variability in the effect of population size across different scripts. Including random slopes for both without clear justification could lead to overfitting, adding unnecessary complexity to the model and especially reducing the power to detect true effects: as noted by [129], a too-complex MLEM, by including excessive random slopes, may inflate Type I error rates and reduce the ability to identify significant predictors due to overfitting. This underscores the importance of balancing model complexity with the need to capture meaningful variability.

We thus opted for a two-stage estimation process. We first include random slopes for macro-area, macro-family, and sub-family only in our multi-model multilevel approach (R = 17,920). As a second step, we then additionally include random slopes for writing script and corpus (R = 35,200). Fig 8 visualises the results.

With respect to the relative importance of the fixed and the random effects, both Fig 8A and 8B largely point in the same direction as the random-effects-only approach (see Fig 7A). A noteworthy exception in both cases is that speaker population size is not only maximally important in predicting both $\hat{h}^{best}$ and L, but also very to maximally important in predicting



	h^{best}			L		K^{best}		
	word	char	BPE	word	char	word	char	BPE
Population size	1.00	1.00	1.00	1.00	1.00	0.44	0.34	0.66
EGIDS = 0	0.58	0.53	0.86	0.92	0.82	0.63	0.68	0.46
EGIDS = 1	0.31	0.31	0.32	0.82	0.90	0.46	0.40	0.40
EGIDS = 2	0.35	0.31	0.39	0.33	0.28	0.45	0.44	0.49
EGIDS = 3	0.34	0.27	0.29	0.29	0.28	0.36	0.53	0.59
N countries	0.42	0.71	0.69	0.30	0.29	0.70	0.68	0.80
parallel	0.32	0.27	0.27	0.70	0.34	0.41	0.30	0.27
Writing script	1.00	1.00	1.00	1.00	1.00	0.94	0.83	1.00
Corpus	1.00	1.00	1.00	1.00	1.00	0.95	0.98	1.00
Macro-area	0.84	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Macro-family	1.00	0.92	1.00	1.00	0.72	0.44	0.45	0.30
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

Fig 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, $\hat{\beta}_x$, of population size, here denoted as $\beta_{\text{log\_pop}}$, 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 95%-CI crosses zero (vertical dashed grey line), whereas blue and pink indicate significant negative and positive effects. Interactive LM-specific visualisations are available at https://www.owid.de/plus/tradeoffvis/.

https://doi.org/10.1371/journal.pcsy.0000032.g007

Table 5. FMA-estimated fixed effects on each outcome per symbolic level. Each cell lists $\tilde{\beta}_x$, here denoted as β and the corresponding p-value (cf. Sect. 2.6.2 for details). 1st column: fixed effect. 2nd– 4th column: results for $\hat{h}^{best}$. 5th + 6th column: results for L. 7th– 9th column: results for $\hat{K}^{best}$. Cell content is highlighted in bold if $p < .05$.

	$\hat{h}^{best}$			L		$\hat{K}^{best}$		
	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

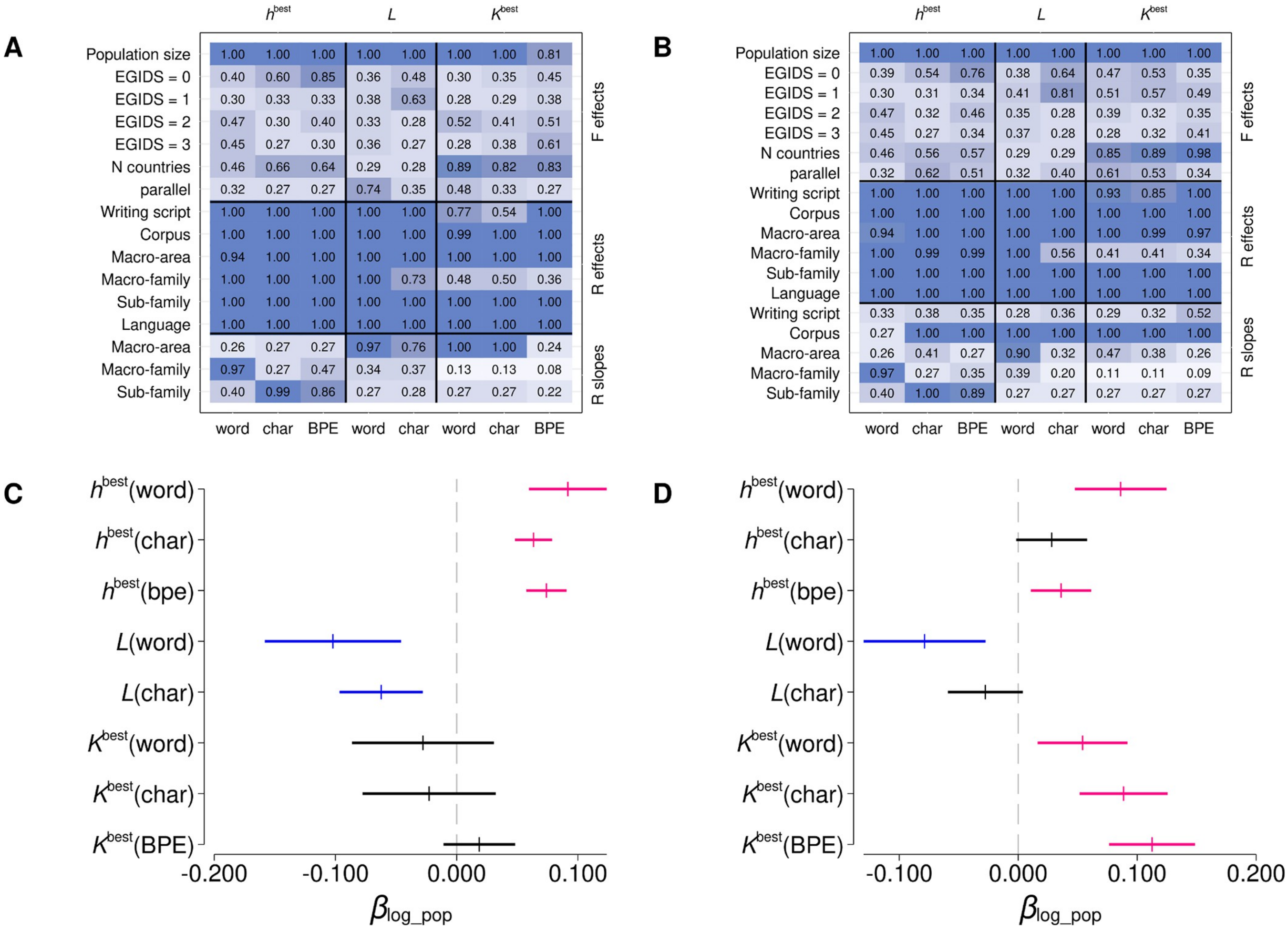
https://doi.org/10.1371/journal.pcsy.0000032.t005

$\hat{K}^{best}$. Regarding the relative importance of the variables for which we include random slopes in both scenarios, there is significant agreement for the two slopes included for phylogenetic non-independence (macro- and sub-family): neither seems to play an important role in predicting either L or $\hat{K}^{best}$.

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 $\hat{K}^{best}$ 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 $\hat{K}^{best}$ on both the word and the character level. The inclusion of writing script as a random slope does not seem to be very important. However, including corpus seems to make a difference for all seven parameter combinations, except for $\hat{h}^{best}$ on the word level.

Regarding the FMA-estimated effect ($\tilde{\beta}_x$) of speaker population size, Fig 8C shows that in the first scenario, the results for each outcome and symbolic level are qualitatively identical to the random-effects-only approach (see Fig 7B): (i) a significant positive effect of population size on $\hat{h}^{best}$, (ii) a significant negative effect on L, and (iii) no significant evidence for any effect on $\hat{K}^{best}$. This again suggests a trade-off between entropy and length.

In the second scenario, while there remains stable evidence for an entropy-length trade-off for words as symbols, at the character level, neither the positive effect on $\hat{h}^{best}$ nor the negative effect on L reaches significance at the 5% level ($p = 0.064$ for $\hat{h}^{best}$ and $p = 0.086$ for L). Unexpectedly, for $\hat{K}^{best}$, there is a significant positive effect of population size across all three levels. Future work could determine whether this indicates a true effect or if this result arises from an overly complex model structure that is unable to detect true effects accurately. To foster such endeavours, we provide a dataset on our OSF repository (https://osf.io/93csg/) that contains all information needed to replicate our findings and conduct follow-up investigations. This dataset includes estimates for both (i) $\hat{h}^{best}$ and K^{best} and (ii) LM-specific estimates, $\hat{h}^{LM}$ and $\hat{K}^{LM}$ for the seven investigated LMs across all three symbolic levels and two different outcome



	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	

	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	0.37	0.28	0.28	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	0.93	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	

Fig 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 $\beta_{\text{log\_pop}}$, 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 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/.

https://doi.org/10.1371/journal.pcsy.0000032.g008

transformations (standardised vs. logged, see Sect. 2.6.2 for details), totalling information for R = 3,520,000 individual models.

Further note that as mentioned above (see Sect. 2.6.2), we chose AIC$_j$ as the criterion to weigh models. AIC is computed by balancing goodness-of-fit with model complexity, i.e., $-2\hat{\mathcal{L}}_j + 2(k_j^f + k_j^r)$, where $\hat{\mathcal{L}}_j$ denotes the maximized log-likelihood of model j and k_j^f and k_j^r are the numbers of estimated fixed effects and random effects (including random slopes) parameters, respectively. However, AIC is not the only potential choice as a criterion. For example, we can choose a variant of the Bayesian Information Criterion (BIC) [130–132] that imposes a more substantial penalty on the inclusion of additional parameters than AIC, computed as $-2\hat{\mathcal{L}}_j + 2(\log(N - k_j^f)k_j^r)$. If we use this criterion to compute results for the second scenario, all obtained FMA-estimated effects of speaker population size, i.e., positive effects for

h and negative effects for L, reach significance at $p < 0.05$ on all three symbolic levels. We invite interested readers to further explore this using our interactive visualization tool available at https://www.owid.de/plus/tradeoffvis/, where we offer results for a total of four different information criteria.

4. Discussion

As written above, Baroni argues that LMs should be seen as distinct algorithmic linguistic theories rather than "blank slates," as they inherently encode structural biases that shape their linguistic capabilities. Each LM thus represents a "general theory defining a space of possible grammars." [48]. Put differently, an LM can be seen as a model of an idealised language learner [10,133,134]. Hence, we can think of an LM that is trained on language-specific data "as a *grammar*, that is, a computation system that, given an input, can predict whether the sequence is acceptable to an idealized speaker of the language" [48]. In this paper, we investigated very different types of LMs, belonging to different classes: statistical, machine learning, and deep learning models. Each class exemplifies unique learning capabilities and limitations:

PPM can be seen as an idealised learner focused on identifying and refining local word sequence patterns. By memorising partial matches with the last few symbols (e.g., 2 for PPM2, 6 for PPM6), PPM constructs a probabilistic grammar of language that adjusts predictions based on immediate context. This model is particularly effective at learning the structure of short-term dependencies and frequent n-grams within sentences based on very little input, making it adept at capturing local grammatical rules but limited in handling long-term dependencies.

PAQ functions as an idealised learner that integrates insights from multiple models to form a coherent representation of grammar. Each model within PAQ contributes specialised ‘knowledge’, and a gated linear network moderates these contributions to balance and refine various perspectives. This process allows PAQ to learn from multiple strings simultaneously, adapting to a wide range of linguistic patterns and refining predictions through consensus. However, the complexity of integrating multiple models may limit its ability to focus on highly specific patterns or rare linguistic structures.

LSTM networks represent an idealised learner adept at preserving and utilizing temporal sequences. By using memory cells and gates, LSTMs maintain continuity and context over longer sequences, enabling them to learn long-term dependencies and sequential information. This capability allows LSTMs to represent the progression of language over time, capturing narrative coherence and integrating both long-term and short-term adjustments. However, LSTMs may struggle with complex hierarchical structures due to their sequential processing nature and consequently require extensive training time.

Lastly, the Transformer-XL model [86], which the NNCP compressor used by us is based on [83], can be seen as an idealised learner that is proficient at mapping global context through its self-attention mechanism. This mechanism enables each symbol to dynamically relate to others, constructing a comprehensive web of interactions between symbols. Transformers are highly capable of learning complex dependencies and contextual relationships within entire sentences or documents. Their ability to process all preceding tokens simultaneously allows them to represent language in a broad and interconnected manner, making them particularly adept at generating coherent and contextually relevant text. However, this ability also comes at a cost, as Transformer models need to be trained on huge amounts of data to achieve this level of performance (see, e.g., Fig 2A–2C).

Our first main result (Sect. 3.1.1) indicates that the choice of the LM has very little impact on the obtained results. Given the far-reaching architectural differences between the investigated LMs, we think this is a surprising result. For instance, a PPM2 model, by design, lacks the memory to store long-term dependencies, yet Fig 2D–2F shows that the results are highly comparable across LMs. This trend holds even for our largest corpus, UNPC ($\tilde{L}_W$ = 341,723,872, see Sect. 2.1.3), which contains information for six languages. For example, the median value of ρ_{both} between $\hat{h}^{PPM2}$ and $\hat{h}^{TRFmed}$, the former estimate being based on an LM with a context window of exactly two symbols and the latter based on a transformer model with ~19.1 million parameters (see Table 2), across symbolic levels ($N_\rho = 9$) is $\tilde{\rho}_{\text{both}} = 0.90$. Similarly, take H_{Cr}, which is based on an even simpler LM, i.e., a 1-gram LM that does not consider any relationships between words but only their frequency of occurrence. Yet the median of ρ_{both} between H_{Cr} and $\hat{h}^{LM}$ for the seven considered LMs across levels and corpora ($N_\rho = 840$) is $\tilde{\rho}_{\text{both}} = 0.42$. This consistency across LMs is an important observation for several reasons. Firstly, larger LMs are notably more expensive to train, requiring substantially more computational resources as compared to smaller models like PAQ, as outlined in Table 2. Additionally, larger models need a lot of training data to achieve optimal performance (see Fig 2). This cost-effectiveness makes smaller LMs particularly attractive for cross-linguistic studies, especially when computational resources are scarce. Furthermore, as written above, the available electronic data for many languages, particularly those spoken by smaller populations, is very limited [36]. Our results indicate that training smaller LMs in such endeavours seems to be a viable option.

It is important to point out that our demonstration of consistency across LMs occurs under a specific scenario: we examined whether the information-theoretic complexity of languages relative to each other remains consistent regardless of the LM used. In other words, if language A is deemed more complex than language B when analysed with one particular LM, this relationship persists even when a completely different type of LM is employed. Future cross-linguistic research should explore whether this agreement across LMs extends to other types of analyses and questions, particularly those that are not strictly based on information-theoretic measures.

As our second main result (Sect. 3.1.2), we re-evaluated the hypothesis that all languages are equally complex. To address this, we developed a statistical approach that integrates machine learning with spatial filtering methods. This methodology, detailed in Sect. 2.6.1, was designed to control for language- and document-specific characteristics, as well as the phylogenetic and geographic relationships between languages. We used this approach to compare entropy estimates across corpora and showed that for different LMs, different types of symbols as information encoding units, and under control of potential sources of influence, a language with a high/low entropy rate in one corpus also tends to be more/less complex in another corpus. While entropy rate measures the difficulty in predicting text after the statistical structure of the input language has been learned [135], we can further explore the learning process itself by examining how difficult it is for LMs to learn these predictions. In our prior work [37], we showed that languages which are harder to predict often tend to be easier/faster to learn for PPM models. Building on this, recent findings [54] provide evidence of a relationship between learning difficulty and speaker population size, suggesting–contrary to expectations derived from previous research–that languages with larger speaker populations tend to be harder for LMs to learn. These results, along with the analyses presented here, offer information-theoretic evidence that challenges the equi-complexity hypothesis at the level of written language.

This result inevitably leads to the question: as higher complexity in language results in more demanding processing efforts, why should there be a trend towards increased complexity in certain languages? We provide a potential answer to this question as our third main result (Sect. 3.1.2): we showed that there is a trade-off between the distributions of estimated entropy rates and length across corpora and across LMs. Given that, from an information-theoretic perspective, message length quantifies efficiency [15,136] we argue that this result suggests that higher complexity is compensated by higher efficiency.

In discussions about the obtained trade-off between L and the average amount of information per symbol, i.e., h, a recurring objection was that this result seems trivial. As some colleagues pointed out, if the same message is encoded in two languages, each symbol in the language with the shorter message length must transmit more information, almost by definition. In a similar vein, in a recent publication a large-scale quantitative information-theoretic analysis of parallel corpus data in almost 1,000 languages was presented to show that there are apparently strong associations between the way languages encode information into words and patterns of communication, e.g. the configuration of semantic information [137]. This publication was criticised by [138] who demonstrated that the results presented by [137] are systematically biased by varying text lengths, which is a very well-known fact in quantitative linguistics as most, if not all, quantities in the context of word frequency distributions vary systematically with text length [139–141]. The authors of [137] responded that what they call "information density" and text length are "two sides of the same coin" and that the Gibbs-Shannon entropy and text length, conditional on the same content, measure the same underlying construct [142]: "because the information content of the parallel translations is the same across comparison languages, we can infer that the more words present [sic] within a document covering the same material, the less information is encoded in each." Another recent response to [137] made a similar argument: "if it takes a language more words to convey the 'same' message, then each word conveys less information." [143]. Both arguments are incorrect. First, let us note that a trade-off between h and L does not only occur for parallel corpora, but it is also observed (i) for comparable corpora (the median adjusted correlation, ρ_{both}, between $\hat{h}^{best}$ and L for the seven comparable corpora in our database (see Sect. 2.1 and S3 Text) amounts to $\tilde{\rho}_{\text{both}}$ = -0.47 for words, $\tilde{\rho}_{\text{both}}$ = -0.59 for characters and $\tilde{\rho}_{\text{both}}$ = -0.59 for BPE, $N_{\rho} = 49$), and (ii) across corpora–in other words: if we know the entropy distributions in one multilingual text collection (parallel or not), we can predict the length in another corpus. For example, the adjusted pairwise correlation ($N = 73$) between $\hat{h}^{best}$ on the BPE level for the UDHR (parallel) corpus and L on the character level for the LCC news (comparable) corpus is ρ_{both} = -0.79.

Secondly, Eq 1 clearly demonstrates that the Gibbs-Shannon unigram entropy is not simply equivalent to text length, but rather a diversity index [45] that measures the amount of "freedom in the combination of symbols" [144], as it is a function of the number of different symbols and how evenly those symbols are distributed. Thirdly, the main problem with such arguments is that the information-theoretic concept of information is fully agnostic about the content of messages [15]. As Shannon, the founding father of information theory, puts it [136]:

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages."

In fact, the use of "information" as a label (e.g., "amount of information" or "information content") has been a "continuous source of misunderstanding" since the very inception of information theory, as pointed out by Bar-Hillel [145], because "it is psychologically almost impossible not to make the shift from the one sense of information, [. . .], i.e., information = signal sequence, to the other sense, information = what is expressed by the signal sequence". In information theory, messages are only treated as signal sequences but not as "content-bearing entities" [145]. Bar-Hillel shows that only under very special circumstances, which do not typically apply to human language, can we infer the amount of information conveyed merely by the length of a message. There is no logical connection between the concept of semantic information (i.e., what is "expressed" by a transmitted signal sequence) and the rarity or improbability of the signal, which is measured by h. Put differently, information in the information-theoretic sense has nothing to do with meaning, but only measures the amount of uncertainty that is removed when receiving a message or the average amount of information learned about upcoming symbols when observing a symbol of the message.

To illustrate this, consider the sentence "Rain occurs most frequently in spring." Using LCC word frequency information (cf. Sect. 2.1.1), the predictive uncertainty or self-information, I, of the sentence, calculated based on a unigram LM (see Eq 1), is I_{word}^{1gram} = 12.531 bits per symbol (bps). Now compare this with the sentence "Liquid precipitation occurs most frequently in spring." If [142,143] were correct, we should expect that this sentence has lower per-symbol information content, as it contains the same message but uses 7 instead of 6 words. However, the self-information for this sentence is I_{word}^{1gram} = 13.646 bps and therefore higher. This is due to the fact that "rain" is much more frequent in our example data (f_{LCC} = 1,594) than "liquid" (f_{LCC} = 324) or "precipitation" (f_{LCC} = 96). Since a unigram LM does not consider any contextual information, it makes sense to additionally use a large LM to compute self-information. Using a GPT-2 model with ~1.5 billion parameters, the self-information of the first sentence amounts to I_{word}^{GPT-2} = 0.952 bps. Again, the self-information of the second sentence is higher with I_{word}^{GPT-2} = 0.965 bps.

As a further illustration, take the following two sentences: "I watched TV." and "I watched television." From the point of view of propositional content, both sentences contain the same message. Following [142,143], we thus should expect that the self-information is the same in both cases. However, both I_{word}^{1gram} and I_{word}^{GPT-2} are lower for the first sentence (I_{word}^{1gram} = 11.915, I_{word}^{GPT-2} = 3.068) compared to the second one (I_{word}^{1gram} = 12.397, I_{word}^{GPT-2} = 3.210) again because "television" is much less frequent (f_{LCC} = 1,568) than "TV" (f_{LCC} = 4,269).

Finally, consider the sentence pair "They will attack at dawn." versus "They will attack at 5am." From the point of view of propositional content, the "total amount of information" of the second sentence should be higher because it specifies a precise time. Based on a unigram LM, the total amount of information is I_{total}^{1gram} = 53.389 bits for the first sentence and I_{total}^{1gram} = 55.884 bits for the second one. However, based on the arguably much better GPT-2 LM, we obtain the opposite result with I_{total}^{GPT-2} = 7.000 bits for the first sentence and I_{total}^{GPT-2} = 6.620 bits for the second one.

These illustrations demonstrate that both H and h reflect the statistical structure of "linguistic sequences independently of the specific information that [is] being encoded" [146]. As such, it is important to point out that information theory cannot be used to measure something like the total amount of semantic information/propositional content of a sentence. Instead, information theory provides a framework for quantifying the efficiency of symbol sequences in transmitting data, focusing on the probabilistic structure and redundancy of the language rather than its semantic content.

To elaborate further, it should be noted that translation from one language into another is fundamentally different from what is meant by encoding in the information-theoretic sense. In early works on machine translation, such as Warren Weaver's famous memorandum [147], translation was sometimes understood to be similar to, e.g., cryptographic decoding. But encoding schemes are basically just biunique mappings of an information source to symbol sequences that permit exact recovery of the original symbols, whereas there is no information-theoretically well-defined sense in which one (of possibly very many) translation of a text into another human language conveys the "same content". The difference is similar to that between sign language and fingerspelling. A written English text can be translated into a sign language, such as American Sign Language (ASL). Since ASL is a full-blown natural language of its own (with no relation whatsoever to spoken English), there are usually many possible translations none of which allows unambiguous reconstruction of the source text from it. On the other hand, fingerspelling (which is a part of ASL) enables users to mechanically render e.g. English words through sequences of signs representing letters and can thus, in principle, be back-translated losslessly to the oral language original.

The biuniqueness of encoding schemes is the reason why text length and entropy rate are indeed trivially inversely related to each other when comparing different encodings of a source text, since different encodings of the same source can always be compressed to the exact same outcome and the length of that outcome is then used to estimate the entropy rate. For translations between human languages no such argument is available. Indeed, as we will demonstrate in what follows, shorter translations may in principle come with a lower entropy rate instead of a higher one.

Before further interpreting the complexity-efficiency trade-off, we will discuss the results, presented in the Supporting Information, from several additional analyses we conducted to test the reliability and validity of this trade-off. First, practicing what we preached above, we rule out the possibility that the association between h and L is simply the result of a well-known systematic text-length bias (S4 Text). Secondly, we demonstrate that there is clear evidence for an entropy-length trade-off only *between* languages, but not *within* languages (S5 Text). Thirdly, we show in S6 Text both theoretically and empirically that the trade-off is indeed not trivial, as one can define processes that increase both entropy and length at the same time. Fourthly, we control for cross-linguistic differences in the number of different words and characters that might stem from, amongst others, morphological typology, different writing systems, varying phonological constraints etc., by using the number of different word-form types and the number of different characters (both log-transformed averages over all translations per language in the *BibleNT* collection) as additional control variables when assessing the (dis)similarity of entropy and length distributions across corpora (see S7 Text). Fifthly, to address the issue of varying levels of orthographic depth (see Sect. 2.4), we incorporated data from ref. [148], who uses a transformer LM to estimate the percentage of correct predictions in phoneme-to-grapheme and grapheme-to-phoneme "translation" tasks across 16 languages. These estimates, which serve as measures of phonemic transparency, provide a quantitative assessment of the ease with which written text can be converted to or from its spoken form, i.e., how directly characters map to phonemes. In S8 Text, we used this information as an additional covariate to control for orthographic depth in our analyses. In both S7 Text and S8 Text our results remain robust: (i) languages with high/low entropy rates in one corpus also tend to be more/less complex in another corpus, (ii) an analogous pattern holds for cross-language length distributions, and (iii) there is a trade-off between the distributions of estimated entropy rates and text length across corpora. Sixthly, we show in S9 Text that the patterns observed in the other investigated multilingual text collections align with those from the PBC. As written above (see Sect. 2.4), we believe that this consistency bolsters confidence in the robustness and validity of our quantitative results, due to the substantial effort and typologically informed curation of the PBC (see Sect. 2.1.1). Seventhly, our study focuses exclusively on written language. To explore a potential connection to spoken language, we utilised data from the *VoxClamantis* corpus [149], which is based on audio recordings of the New Testament of the Bible. In S10 Text, 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.

Let us speculate that in large societies, institutionalised education potentially makes greater linguistic complexity possible by providing systematic and formalised language learning, which, in turn, can support the acquisition and use of more complex linguistic structures. In line with this, a recent large-scale study found a positive statistical correlation between grammatical complexity and speaker population size [53]. At the same time, the importance of written communication in larger societies might create a natural pressure towards shorter messages, as it saves costs for producing, storing, and transmitting written texts (e.g., book paper, storage space, bandwidth). This dual influence–educational systems enabling complexity and the practical need for efficiency in written communication–could help explain why languages in larger communities might evolve to balance these pressures, resulting in shorter but more complex messages. Testing this hypothesis is an important avenue for future research.

Supporting information

S1 Text. Language modelling details.
(PDF)

S2 Text. Scaling σ_x.
(PDF)

S3 Text. Numeric IDs.
(PDF)

S4 Text. Testing for a potential systematic length bias.
(PDF)

S5 Text. Associations between and within languages.
(PDF)

S6 Text. The entropy-length trade-off is not trivial.
(PDF)

S7 Text. Controlling for the number of different words and characters.
(PDF)

S8 Text. Controlling for orthographic transparency/depth.
(PDF)

S9 Text. Alignment with the PBC.
(PDF)

S10 Text. Testing the trade-off on the phonemic level.
(PDF)

Author Contributions

Conceptualization: Alexander Koplenig, Peter Meyer.

Data curation: Alexander Koplenig.

Formal analysis: Alexander Koplenig.

Investigation: Alexander Koplenig.

Methodology: Alexander Koplenig.

Resources: Alexander Koplenig, Jan Oliver Rüdiger.

Software: Alexander Koplenig, Sascha Wolfer.

Validation: Alexander Koplenig, Sascha Wolfer.

Visualization: Alexander Koplenig, Sascha Wolfer.

Writing – original draft: Alexander Koplenig.

Writing – review & editing: Alexander Koplenig, Sascha Wolfer, Peter Meyer.

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. https://doi.org/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]. https://doi.org/10.1038/s41586-024-07487-w PMID: 38718835
4. Gruver N, Finzi M, Qiu S, Wilson AG. Large Language Models Are Zero-Shot Time Series Forecasters. arXiv; 2023. https://doi.org/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. https://doi.org/10.1038/s41587-019-0224-x PMID: 31477924
6. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018; 559: 547–555. https://doi.org/10.1038/s41586-018-0337-2 PMID: 30046072
7. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 2015; 12: 931–934. https://doi.org/10.1038/nmeth.3547 PMID: 26301843

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1111/cogs.13256 PMID: 36840975

12. Grindrod J. Modelling Language. arXiv; 2024. https://doi.org/10.48550/ARXIV.2404.09579

13. Abramski K, Citraro S, Lombardi L, Rossetti G, Stella M. Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students. BDCC. 2023; 7: 124. https://doi.org/10.3390/bdcc7030124

14. Mahoney M. Data Compression Explained. Dell Inc.; 2013. Available: http://mattmahoney.net/dc/dce.html

15. Gibson E, Futrell R, Piandadosi ST, Dautriche I, Mahowald K, Bergen L, et al. How Efficiency Shapes Human Language. TRENDS in Cognitive Science. 2019; 23: 389–407. https://doi.org/10.1016/j.tics.2019.02.003 PMID: 31006626

16. Cysouw M, Wälchli B. Parallel texts: using translational equivalents in linguistic typology. Language Typology and Universals. 2007; 60: 95–99. https://doi.org/10.1524/stuf.2007.60.2.95

17. Bentz C, Ferrer-i-Cancho R. Zipf's law of abbreviation as a language universal. In: Bentz C, Jäger G, Yanovich I, editors. Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. Tübingen: University of Tübingen; 2016. https://doi.org/10.15496/publikation-10057

18. Bentz C. The Low-complexity-belt: Evidence For Large-scale Language Contact In Human Prehistory? In: Roberts SG, Cuskley C, McCrohon L, Barceló-Coblijn L, Fehér O, Verhoef T, editors. The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11). Online at http://evolang.org/neworleans/papers/93.html; 2016.

19. Koplenig A. Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. Royal Society Open Science. 2019; 6: 181274. https://doi.org/10.1098/rsos.181274 PMID: 30891265

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

21. Kauhanen H, Einhaus S, Walkden G. Language structure is influenced by the proportion of non-native speakers: A reply to Koplenig (2019). Journal of Language Evolution. 2023; lzad005. https://doi.org/10.1093/jole/lzad005

22. Koplenig A. Still No Evidence for an Effect of the Proportion of Non-Native Speakers on Natural Language Complexity. Entropy. 2024; 26: 993. https://doi.org/10.3390/e26110993 PMID: 39593937

23. Greenberg JH. A Quantitative Approach to the Morphological Typology of Language. International Journal of American Linguistics. 1960; 26: 178–194.

24. Bentz C, Ruzsics T, Koplenig A, Samardzic T. A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora. Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC). Osaka, Japan; 2016. pp. 142–153.

25. Wälchli B, Cysouw M. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. Linguistics. 2012; 50. https://doi.org/10.1515/ling-2012-0021

26. Östling R. Word Order Typology through Multilingual Word Alignment. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing, China: Association for Computational Linguistics; 2015. pp. 205–211. https://doi.org/10.3115/v1/P15-2034

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

28. Koplenig A, Meyer P, Wolfer S, Müller-Spitzer C. The statistical trade-off between word order and word structure–Large-scale evidence for the principle of least effort. Smith K, editor. PLOS ONE. 2017; 12: e0173614. https://doi.org/10.1371/journal.pone.0173614 PMID: 28282435

29. Pimentel T, Meister C, Salesky E, Teufel S, Blasi D, Cotterell R. A surprisal–duration trade-off across and within the world's languages. Proceedings of the 2021 Conference on Empirical Methods in

Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. pp. 949–962. https://doi.org/10.18653/v1/2021.emnlp-main.73

30. de Vries LJ. Some remarks on the use of Bible translations as parallel texts in linguistic research. Sprachtypologie und Universalienforschung. 2007; 60: 148–157. https://doi.org/10.1524/stuf.2007.60.2.148

31. Wälchli B. Advantages and disadvantages of using parallel texts in typological investigations. Language Typology and Universals. 2007; 60: 118–134. https://doi.org/10.1524/stuf.2007.60.2.118

32. Tiedemann J. Parallel Data, Tools and Interfaces in OPUS. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: ELRA; 2012. pp. 2214–2218. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

33. Mayer T, Cysouw M. Creating a Massively Parallel Bible Corpus. In: Calzolari N, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, et al., editors. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA); 2014.

34. Levshina N. Verbs of letting in Germanic and Romance languages: A quantitative investigation based on a parallel corpus of film subtitles. LiC. 2016; 16: 84–117. https://doi.org/10.1075/lic.16.1.04lev

35. Goldhahn D, Eckart T, Quasthoff U. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey; 2012. pp. 759–765. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf

36. Scannell KP. The Crúbadán Project: Corpus building for under-resourced languages. Proceedings of the 3rd Web as Corpus Workshop: Building and Exploring Web Corpora. 2007. pp. 5–15. Available: http://cs.slu.edu/~scannell/pub/wac3.pdf

37. Koplenig A, Wolfer S, Meyer P. A large quantitative analysis of written language challenges the idea that all languages are equally complex. Sci Rep. 2023; 13: 15351. https://doi.org/10.1038/s41598-023-42327-3 PMID: 37717109

38. Ehret K, Blumenthal-Dramé A, Bentz C, Berdicevskis A. Meaning and Measures: Interpreting and Evaluating Complexity Metrics. Front Commun. 2021; 6: 640510. https://doi.org/10.3389/fcomm.2021.640510

39. Sampson G. A linguistic axiom challenged. In: Sampson G, Gil D, Trudgill P, editors. Language complexity as an evolving variable. Oxford: Oxford University Press; 2009. pp. 1–18.

40. Ehret K, Berdicevskis A, Bentz C, Blumenthal-Dramé A. Measuring language complexity: challenges and opportunities. Linguistics Vanguard. 2023; 0. https://doi.org/10.1515/lingvan-2022-0133

41. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Hoboken, N.J: Wiley-Interscience; 2006.

42. Futrell R, Hahn M. Information Theory as a Bridge Between Language Function and Language Form. Front Commun. 2022; 7: 657725. https://doi.org/10.3389/fcomm.2022.657725

43. Ren G, Takahashi S, Tanaka-Ishii K. Entropy Rate Estimation for English via a Large Cognitive Experiment Using Mechanical Turk. Entropy. 2019; 21: 1201. https://doi.org/10.3390/e21121201

44. Meister C, Pimentel T, Wiher G, Cotterell R. Locally Typical Sampling. arXiv; 2022. https://doi.org/10.48550/ARXIV.2202.00666

45. Kolmogorov AN. Three approaches to the quantitative definition of information. International Journal of Computer Mathematics. 1968; 2: 157–168. https://doi.org/10.1080/00207166808803030

46. Kontoyiannis I. The Complexity and Entropy of Literary Styles. NSF Technical Report, Department of Statistics, Stanford University. 1996;97.

47. Cover TM. Kolmogorov Complexity, Data Compression, and Inference. In: Skwirzynski JK, editor. The Impact of Processing Techniques on Communications. Dordrecht: Springer Netherlands; 1985. pp. 23–33. https://doi.org/10.1007/978-94-009-5113-6_2

48. Baroni M. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. arXiv; 2021. https://doi.org/10.48550/ARXIV.2106.08694

49. Bromham L. Solving Galton's problem: practical solutions for analysing language diversity and evolution. PsyArXiv; 2022 May. https://doi.org/10.31234/osf.io/c8v9r

50. Bromham L, Yaxley KJ. Neighbours and relatives: accounting for spatial distribution when testing causal hypotheses in cultural evolution. Evolut Hum Sci. 2023; 5: e27. https://doi.org/10.1017/ehs.2023.23 PMID: 37829289

51. Guzmán Naranjo M, Becker L. Statistical bias control in typology. Linguistic Typology. 2022; 26: 605–670. https://doi.org/10.1515/lingty-2021-0002

52. Claessens S, Kyritsis T, Atkinson QD. Cross-national analyses require additional controls to account for the non-independence of nations. Nat Commun. 2023; 14: 5776. https://doi.org/10.1038/s41467-023-41486-1 PMID: 37723194

53. Shcherbakova O, Michaelis SM, Haynie HJ, Passmore S, Gast V, Gray RD, et al. Societies of strangers do not speak less complex languages. Sci Adv. 2023; 9: eadf7704. https://doi.org/10.1126/sciadv.adf7704 PMID: 37585533

54. Koplenig A, Wolfer S. Languages with more speakers tend to be harder to (machine-)learn. Sci Rep. 2023; 13: 18521. https://doi.org/10.1038/s41598-023-45373-z PMID: 37898699

55. Hall S, Moskovitz C, Pemberton M, Text Recycling Research Project. Understanding text recycling. A guide for researchers V.1. 2021. Available: https://textrecycling.org/files/2021/06/Understanding-Text-Recycling_A-Guide-for-Researchers-V.1.pdf

56. Gutierrez-Vasques X, Bentz C, Samardžić T. Languages Through the Looking Glass of BPE Compression. Computational Linguistics. 2023; 49: 943–1001. https://doi.org/10.1162/coli_a_00489

57. Simons GF, Fennig CD. Global Dataset Ethnologue: Languages of the World, Twentieth edition. SIL International; 2017. Available: http://www.ethnologue.com.

58. Bromham L, Dinnage R, Skirgård H, Ritchie A, Cardillo M, Meakins F, et al. Global predictors of language endangerment and the future of linguistic diversity. Nat Ecol Evol. 2022; 6: 163–173. https://doi.org/10.1038/s41559-021-01604-y PMID: 34916621

59. Hammarström H, Forkel R, Haspelmath M, Bank S. glottolog/glottolog: Glottolog database 4.8. Zenodo; 2023. https://doi.org/10.5281/ZENODO.8131084

60. Simons GF, Fennig CD, editors. Ethnologue: languages of Africa and Europe. Twentieth edition. Dallas, TX: SIL; 2017.

61. Simons, Gary F, Fennig CD. Ethnologue: Languages of the World. Dallas, Texas: SIL International; 2017.

62. Lewis MP, Simons GF. Assessing Endangerment: Expanding Fishman's GIDS. Revue Roumaine de Linguistique. 2010; 55: 103–120.

63. Baker M. Corpus Linguistics and Translation Studies—Implications and Applications. In: Baker M, Francis G, Tognini-Bonelli E, editors. Text and Technology. Amsterdam: John Benjamins Publishing Company; 1993. pp. 233–252. https://doi.org/10.1075/z.64.15bak

64. Cotterell R, Mielke SJ, Eisner J, Roark B. Are All Languages Equally Hard to Language-Model? Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018. pp. 536–541. https://doi.org/10.18653/v1/N18-2085

65. Jäger G. Global-scale phylogenetic linguistic inference from lexical resources. Scientific Data. 2018; 5: 180189. https://doi.org/10.1038/sdata.2018.189 PMID: 30299438

66. Wichmann S, Holman EW, Brown CH, Forkel R, Tresoldi T. CLDF dataset derived from Wichmann et al.'s "ASJP Database" v17 from 2016. Zenodo; 2016. https://doi.org/10.5281/ZENODO.3835942

67. Stewart WA. A Sociolinguistic Typology for Describing National Multilingualism. In: Fishman JA, editor. Readings in the Sociology of Language. De Gruyter Mouton; 1968. pp. 531–545. https://doi.org/10.1515/9783110805376.531

68. Bentz C, Dediu D, Verkerk A, Jäger G. The evolution of language families is shaped by the environment beyond neutral drift. Nature Human Behaviour. 2018; 2: 816–821. https://doi.org/10.1038/s41562-018-0457-6 PMID: 31558817

69. Jaeger TF, Graff P, Croft W, Pontillo D. Mixed effect models for genetic and areal dependencies in linguistic typology. Linguistic Typology. 2011; 15. https://doi.org/10.1515/lity.2011.021

70. Shannon CE. Prediction and Entropy of Printed English. Bell System Technical Journal. 1951; 30: 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

71. Chaitin GJ. On the intelligibility of the universe and the notions of simplicity, complexity and irreducibility. arXiv:math/0210035. 2002 [cited 25 Mar 2021]. Available: http://arxiv.org/abs/math/0210035

72. Delétang G, Ruoss A, Duquenne P-A, Catt E, Genewein T, Mattern C, et al. Language Modeling Is Compression. arXiv; 2023. https://doi.org/10.48550/ARXIV.2309.10668

73. Rissanen JJ. Generalized Kraft Inequality and Arithmetic Coding. IBM Journal of Research and Development. 1976; 20: 198–203. https://doi.org/10.1147/rd.203.0198

74. Jurafsky D, Martin JH. Speech and Language Processing. 3rd ed. 2021. Available: https://web.stanford.edu/~jurafsky/slp3/

75. Cleary J, Witten I. Data Compression Using Adaptive Coding and Partial String Matching. IEEE Transactions on Communications. 1984; 32: 396–402. https://doi.org/10.1109/TCOM.1984.1096090

76. Shkarin D. PPM: one step to practicality. Proceedings DCC 2002 Data Compression Conference. Snowbird, UT, USA: IEEE Comput. Soc; 2002. pp. 202–211. https://doi.org/10.1109/DCC.2002.999958

77. Pavlov I. 7-zip. 2023. Available: https://7-zip.org/

78. Knoll B, Freitas N de. A Machine Learning Perspective on Predictive Coding with PAQ8. 2012 Data Compression Conference. Snowbird, UT, USA: IEEE; 2012. pp. 377–386. https://doi.org/10.1109/DCC.2012.44

79. Veness J, Lattimore T, Budden D, Bhoopchand A, Mattern C, Grabska-Barwinska A, et al. Gated Linear Networks. 2019 [cited 27 Mar 2023]. https://doi.org/10.48550/ARXIV.1910.01526

80. Mahoney M. PAQ8. 2007. Available: http://mattmahoney.net/dc/paq8l.zip

81. Mahoney M. Adaptive weighing of context models for lossless data compression. Florida Tech.; 2005. Available: http://hdl.handle.net/11141/154

82. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997; 9: 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 PMID: 9377276

83. Bellard F. NNCP v3.1: Lossless Data Compression with Transformer. 2021. Available: https://bellard.org/nncp/

84. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. pp. 6000–6010.

85. Witten IH, Bell TC. The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory. 1991; 37: 1085–1094. https://doi.org/10.1109/18.87000

86. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. arXiv:190102860 [cs, stat]. 2019 [cited 22 Oct 2020]. Available: http://arxiv.org/abs/1901.02860

87. Bellard F. Lossless Data Compression with Neural Networks. 2019. Available: https://bellard.org/nncp/nncp.pdf

88. Bellard F. NNCP v2: Lossless Data Compression with Transformer. 2021. Available: https://bellard.org/nncp/nncp_v2.1.pdf

89. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323: 533–536. https://doi.org/10.1038/323533a0

90. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014 [cited 15 May 2023]. https://doi.org/10.48550/ARXIV.1412.6980

91. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. 2019. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

92. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv; 2019. https://doi.org/10.48550/ARXIV.1910.03771

93. Unicode Consortium. Unicode Text Segmentation. In: Unicode® Standard Annex #29 [Internet]. 2019 [cited 23 Jul 2019]. Available: http://www.unicode.org/reports/tr29/#Word_Boundaries

94. Koplenig A. Stata tip 129: Efficiently processing textual data with Stata's new Unicode features. Stata Journal. 2018; 18: 287–289.

95. Mielke SJ, Cotterell R, Gorman K, Roark B, Eisner J. What Kind of Language Is Hard to Language-Model? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. pp. 4975–4989. https://doi.org/10.18653/v1/P19-1491

96. Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics; 2016. pp. 1715–1725. https://doi.org/10.18653/v1/P16-1162

97. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd ed. 2024. Available: https://web.stanford.edu/~jurafsky/slp3/

98. Darmon ANM, Bazzi M, Howison SD, Porter MA. Pull out all the stops: Textual analysis via punctuation sequences. Eur J Appl Math. 2021; 32: 1069–1105. https://doi.org/10.1017/S0956792520000157

99. Moran S, Cysouw M. The Unicode Cookbook For Linguists: Managing Writing Systems Using Orthography Profiles. Berlin: Language Science Press; 2018. https://doi.org/10.5281/zenodo.1296780

100. Haspelmath M. The indeterminacy of word segmentation and the nature of morphology and syntax. Folia Linguistica. 2011;45. https://doi.org/10.1515/flin.2011.002

101. Jacobs J. Grammatik ohne Wörter? In: Engelberg S, Holler A, Proost K, editors. Sprachliches Wissen zwischen Lexikon und Grammatik. Berlin, Boston: DE GRUYTER; 2011. Available: http://www.degruyter.com/view/books/9783110262339/9783110262339.345/9783110262339.345.xml

102. Lyons J. Introduction to Theoretical Linguistics. 1st ed. Cambridge University Press; 1968. https://doi.org/10.1017/CBO9781139165570

103. Aaron PG, Joshi RM. Written Language Is as Natural as Spoken language: A Biolinguistic Perspective. Reading Psychology. 2006; 27: 263–311. https://doi.org/10.1080/02702710600846803

104. Berg K. Graphemic Analysis and the Spoken Language Bias. Front Psychol. 2016; 7. https://doi.org/10.3389/fpsyg.2016.00388 PMID: 27047416

105. Chafe W, Tannen D. The Relation Between Written and Spoken Language. Annu Rev Anthropol. 1987; 16: 383–407. https://doi.org/10.1146/annurev.an.16.100187.002123

106. Haspelmath M. Defining the word. WORD. 2023; 69: 283–297. https://doi.org/10.1080/00437956.2023.2237272

107. Geertzen J, Blevins J, Milin P. Informativeness of linguistic unit boundaries. Apollo—University of Cambridge Repository. 2016 [cited 23 Jul 2019]. https://doi.org/10.17863/cam.69

108. Bickel B, Nichols J. Inflectional morphology. 2nd ed. In: Shopen T, editor. Language Typology and Syntactic Description. 2nd ed. Cambridge University Press; 2007. pp. 169–240. https://doi.org/10.1017/CBO9780511618437.003

109. Evans N, Sasse H-J. Introduction: problems of polysynthesis. In: Evans N, Sasse H-J, editors. Problems of Polysynthesis. Berlin: AKADEMIE VERLAG; 2002. pp. 1–14. https://doi.org/10.1524/9783050080956.1

110. Schürmann T, Grassberger P. Entropy estimation of symbol sequences. Chaos: An Interdisciplinary Journal of Nonlinear Science. 1996; 6: 414. https://doi.org/10.1063/1.166191 PMID: 12780271

111. MacKay DJC. Information theory, inference, and learning algorithms. Cambridge, UK ; New York: Cambridge University Press; 2003.

112. Yaglom AM, Yaglom IM. Probability and information. Dordrecht, Holland ; Boston : Hingham, MA: D. Reidel ; Sold and distributed in the U.S.A. by Kluwer Boston; 1983.

113. Adami C. What is information? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2016; 374: 20150230. https://doi.org/10.1098/rsta.2015.0230 PMID: 26857663

114. Wolf L, Pimentel T, Fedorenko E, Cotterell R, Warstadt A, Wilcox E, et al. Quantifying the redundancy between prosody and text. arXiv; 2023. https://doi.org/10.48550/ARXIV.2311.17233

115. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996; 58: 267–288.

116. Belloni A, Chernozhukov V, Wei Y. Post-Selection Inference for Generalized Linear Models With Many Controls. Journal of Business & Economic Statistics. 2016; 34: 606–619. https://doi.org/10.1080/07350015.2016.1166116

117. Moran PAP. Notes on Continuous Stochastic Phenomena. Biometrika. 1950; 37: 17. https://doi.org/10.2307/2332142 PMID: 15420245

118. Kelejian HH, Prucha IR. On the asymptotic distribution of the Moran I test statistic with applications. Journal of Econometrics. 2001; 104: 219–257. https://doi.org/10.1016/S0304-4076(01)00064-1

119. Griffith DA. A Spatial Filtering Specification for the Autologistic Model. Environ Plan A. 2004; 36: 1791–1811. https://doi.org/10.1068/a36247

120. Tiefelsdorf M, Griffith DA. Semiparametric Filtering of Spatial Autocorrelation: The Eigenvector Approach. Environment and Planning A. 2007; 39: 1193–1221. https://doi.org/10.1068/a37378

121. Oberdabernig DA, Humer S, Crespo Cuaresma J. Democracy, Geography and Model Uncertainty. Scottish J Political Eco. 2018; 65: 154–185. https://doi.org/10.1111/sjpe.12140

122. Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language. 2008; 59: 390–412. https://doi.org/10.1016/j.jml.2007.12.005

123. Nettle D. Social scale and structural complexity in human languages. Philosophical Transactions of the Royal Society B: Biological Sciences. 2012; 367: 1829–1836. https://doi.org/10.1098/rstb.2011.0216 PMID: 22641821

124. Burnham KP, Anderson DR, editors. Model Selection and Multimodel Inference. New York, NY: Springer New York; 2004. https://doi.org/10.1007/b97636

125. Buckland ST, Burnham KP, Augustin NH. Model Selection: An Integral Part of Inference. Biometrics. 1997; 53: 603. https://doi.org/10.2307/2533961

126. Steel MFJ. Model Averaging and Its Use in Economics. Journal of Economic Literature. 2020; 58: 644–719. https://doi.org/10.1257/jel.20191385

127. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19: 716–723. https://doi.org/10.1109/TAC.1974.1100705

128. Claessens S, Atkinson Q. The non-independence of nations and why it matters. PsyArXiv; 2022 Apr. https://doi.org/10.31234/osf.io/m6bsn

129. Matuschek H, Kliegl R, Vasishth S, Baayen H, Bates D. Balancing Type I error and power in linear mixed models. Journal of Memory and Language. 2017; 94: 305–315. https://doi.org/10.1016/j.jml.2017.01.001

130. Schwarz G. Estimating the Dimension of a Model. Ann Statist. 1978; 6. https://doi.org/10.1214/aos/1176344136

131. Vonesh E, Chinchilli VM. Linear and Nonlinear Models for the Analysis of Repeated Measurements. 0 ed. CRC Press; 1996. https://doi.org/10.1201/9781482293272

132. Gurka MJ. Selecting the Best Linear Mixed Model Under REML. The American Statistician. 2006; 60: 19–26. https://doi.org/10.1198/000313006X90396

133. Yang Y, Piantadosi ST. One model for the learning of language. Proc Natl Acad Sci USA. 2022; 119: e2021865119. https://doi.org/10.1073/pnas.2021865119 PMID: 35074868

134. Wolff JG. Language acquisition, data compression and generalization. Language & Communication. 1982; 2: 57–89. https://doi.org/10.1016/0271-5309(82)90035-0

135. Takahira R, Tanaka-Ishii K, Dębowski Ł. Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. Entropy. 2016; 18: 364. https://doi.org/10.3390/e18100364

136. Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal. 1948; 27: 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

137. Aceves P, Evans JA. Human languages with greater information density have higher communication speed but lower conversation breadth. Nat Hum Behav. 2024 [cited 20 Feb 2024]. https://doi.org/10.1038/s41562-024-01815-w PMID: 38366103

138. Koplenig A. Corpus size strongly matters when analysing word frequency distributions. 2024. https://doi.org/10.31219/osf.io/p5nhd

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

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

141. Koplenig A, Wolfer S, Müller-Spitzer C. Studying Lexical Dynamics and Language Change via Generalized Entropies: The Problem of Sample Size. Entropy. 2019; 21. https://doi.org/10.3390/e21050464 PMID: 33267178

142. Aceves P, Evans J. Conditional Word Count and Huffman Code Size are Two Sides of the Same Coin: Response to Koplenig. 2024. https://doi.org/10.31219/osf.io/4b7mc

143. Lupyan G, Contreras Kallens P, Dale R. Information density as a predictor of communication dynamics. Trends in Cognitive Sciences. 2024; 28: 489–491. https://doi.org/10.1016/j.tics.2024.03.012 PMID: 38632006

144. Schmitt AO, Herzel H. Estimating the Entropy of DNA Sequences. Journal of Theoretical Biology. 1997; 188: 369–377. https://doi.org/10.1006/jtbi.1997.0493 PMID: 9344742

145. Bar-Hillel Y. An Examination of Information Theory. Philosophy of Science. 1955; 22: 86–105.

146. Montemurro MA, Zanette DH. Towards the quantification of the semantic information encoded in written language. Advances in Complex Systems. 2010; 13: 135–153. https://doi.org/10.1142/S0219525910002530

147. Translation Weaver W. In: Locke WN, Boothe AD, editors. Machine Translation of Languages. Cambridge, MA: MIT Press; 1949. pp. 15–23.

148. Marjou X. OTEANN: Estimating the Transparency of Orthographies with an Artificial Neural Network. In: Vylomova E, Salesky E, Mielke S, Lapesa G, Kumar R, Hammarström H, et al., editors. Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. Online: Association for Computational Linguistics; 2021. pp. 1–9. https://doi.org/10.18653/v1/2021.sigtyp-1.1

149. Salesky E, Chodroff E, Pimentel T, Wiesner M, Cotterell R, Black AW, et al. A Corpus for Large-Scale Phonetic Typology. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 4526–4546. https://doi.org/10.18653/v1/2020.acl-main.415

150. Wray A, Grace GW. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. Lingua. 2007; 117: 543–578. https://doi.org/10.1016/j.lingua.2005.05.005

151. Lupyan G, Dale R. Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity. TRENDS in Cognitive Science. 2016; 20: 649–660. https://doi.org/10.1016/j.tics.2016.07.005 PMID: 27499347

152. Lupyan G, Dale R. Language Structure Is Partly Determined by Social Structure. O'Rourke D, editor. PLoS ONE. 2010; 5: e8559. https://doi.org/10.1371/journal.pone.0008559 PMID: 20098492

153. Atkinson M, Kirby S, Smith K. Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages. Caldwell CA, editor. PLOS ONE. 2015; 10: e0129463. https://doi.org/10.1371/journal.pone.0129463 PMID: 26057624

154. Raviv L, Meyer A, Lev-Ari S. Larger communities create more systematic languages. Proceedings of the Royal Society B: Biological Sciences. 2019; 286: 20191262. https://doi.org/10.1098/rspb.2019.1262 PMID: 31311478

155. Frank S, Smith K. Natural Population Growth Can Cause Language Simplification. In: Ravignani A, Barbieri C, Martins M, Flaherty M, Jadoul Y, Lattenkamp E, et al., editors. The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13). 2020. https://doi.org/10.17617/2.3190925

156. Raviv L, Peckre LR, Boeckx C. What is simple is actually quite complex: A critical note on terminology in the domain of language and communication. Journal of Comparative Psychology. 2022 [cited 7 Dec 2022]. https://doi.org/10.1037/com0000328 PMID: 36222620