

Towards robust complexity indices in linguistic typology

A corpus-based assessment

Yoon Mi Oh and François Pellegrino

Ajou University | CNRS, University of Lyon

There is high hope that corpus-based approaches to language complexity will contribute to explaining linguistic diversity. Several complexity indices have consequently been proposed to compare different aspects among languages, especially in phonology and morphology. However, their robustness against changes in corpus size and content hasn't been systematically assessed, thus impeding comparability between studies. Here, we systematically test the robustness of four complexity indices estimated from raw texts and either routinely utilized in crosslinguistic studies (Type-Token Ratio and word-level Entropy) or more recently proposed (Word Information Density and Lexical Diversity). Our results on 47 languages strongly suggest that traditional indices are more prone to fluctuation than the newer ones. Additionally, we confirm with Word Information Density the existence of a cross-linguistic trade-off between word-internal and across-word distributions of information. Finally, we implement a proof of concept suggesting that modern deep-learning language models can improve the comparability across languages with non-parallel datasets.

Keywords: complexity metric robustness, complexity trade-off, linguistic typology, morphological complexity, non-parallel corpus

1. Introduction

Language complexity is a notion that has hovered in linguistics for more than a century, both as an implicit scale in a language description (with one feature being considered as more complex than another, such as passive vs. active syntactic constructions for instance) and as a general backdrop against which differences among languages could be dismissed or highlighted, depending on one's position on the equi-complexity hypothesis (see Joseph & Newmeyer 2012, for a

discussion). For most of the twentieth century, those discussions have arguably remained impressionistic but in the last two decades, considerable progress has been achieved in giving language complexity a more tangible substance. On the one hand, large databases based on thorough linguistic descriptions, such as LAPSyD (Maddieson et al. 2013) in phonology or the more comprehensive and multi-domain WALS (Dryer & Haspelmath 2013) and AUTOTYP (Bickel et al. 2022), offered gold mines that led to the emergence and specification of enlightening hypotheses in historical and typological linguistics and to the adoption of a broader worldwide perspective than the (mostly) euro-centric analysis grid that was previously prevalent. On the other hand, the increasing availability of large multilingual corpora allowed a quantitative estimation of several complexity indices that can be estimated from the raw text itself or from linguistically-informed representations that follow cross-linguistically consistent schemes such as Universal Dependencies (de Marneffe et al. 2021) for part-of-speech tagging or UniMorph (McCarthy et al. 2020) for inflectional morphology annotation (See Section 2 for a brief introduction).

Such corpora have been instrumental in confirming remarkable hypothesized relationships, such as the trade-off in information distribution between the within-word and the across-word dimensions (e.g. Ehret & Szmrecsanyi 2016; Koplenig et al. 2017, see also Bentz et al. 2022 for a recent meta-analysis) and the correlation between morphological complexity and sociolinguistic variables (Koplenig 2019; Miestamo et al. 2008; Sinnemäki & Di Garbo 2018). As a consequence, language complexity provides now a useful and intuitive means to tackle open questions in the fields of linguistic typology, both by providing a way to compare languages against a common complexity index and by emphasizing the existence of communicative and linguistic pressures within a language, as revealed by complexity trade-offs such as the aforementioned one.

Linguists are now equipped with an efficient toolbox of complexity metrics, although there exists no unique and authoritative definition of how language complexity should be assessed, so far and arguably ever, because of the highly multidimensional phenomena at play (Ehret et al. 2021). Still, language complexity is far from being fully understood and the recent studies shed light on some pieces of the puzzle, leaving others in the dark. For instance, large-scale studies that involve hundreds of languages often provide an amazing picture on global variations and common trends among languages at the expense of their interpretability in terms of language-specific phenomena (e.g. Koplenig et al. 2017; Pimentel et al. 2021). Conversely, finer-grained studies allowing thorough interpretation mostly focus on small sets of languages, often limited to European languages (but see also Gutierrez-Vasques & Mijangos 2020 and Vera & Palma 2020 for more diverse language samples), at the expense of statistical robustness and

breadth in diversity (Easterday et al. 2021 being a notable exception, with a broad and thorough evaluation of the relation between phonological and morphological complexities).

Another limitation of most of these approaches is that they use textual corpora selected for their comparability across languages at the expense of their naturalness. A parallel corpus, made of linguistic material translated from one (or a few) languages to others (such as the Parallel Bible Corpus (Mayer & Cysouw 2014) used in this study, see Section 3 for details) is thus the archetype analyzed in most studies on linguistic complexity since its language-specific versions are comparable by design. One can assume that differences induced by idiosyncratic semantic and stylistic characteristics are quasi-neutralized and conversely that regular differences among languages are underlined. In other words, a textual parallel corpus provides a material akin to but also distant from our everyday spontaneous linguistic experience, since it is polished by several authors and lacks any temporal dimension. As a consequence, one must be careful in their interpretation and the generalization of conclusions drawn from such corpora should remain cautious. A few attempts have recently been made to break this **parallel corpus barrier** to close the distance to more natural linguistic content. The most comprehensive contribution so far is arguably the work by Ehret & Szmrecsanyi (2016), who compared complexity hierarchies estimated from several configurations ranging from parallel to non-parallel corpora on a small sample of nine European languages. They observed a positive correlation between the resulting hierarchies, but also differences calling for further analyses on a larger and more diverse language sample. More recently, von Prince & Demberg (2018) explored the use of information-theoretic indices to characterize syntactic complexity on a non-parallel corpus of 25 languages (including a few non-European languages) and show a promising correlation with expert judgment.

Here, we propose to thoroughly investigate the potential consequences of some methodological choices in such quantitative typological studies of language complexity. Following Bentz et al. (2016); Ehret et al. (2021) and Koplenig et al. (2017), among others, we focus on complexity metrics estimated from raw text and our main objective is to assess their consistency and robustness against variations in the analyzed corpora. Additionally, we put our methodological results in perspective with the aforementioned trade-off between word-internal and across-word distributions of information and we consider how recent language models developed in Natural Language Processing (NLP) can help extending cross-language comparisons with non-parallel datasets.

Our study is based on a dataset consisting of 47 typologically diverse languages, which conveniently opens a window into linguistic diversity and allows us to keep track of each individual language's characteristics in a manageable way.

The main focus of the study is on the morphological domain, but syntactic aspects are also tangentially addressed when we evaluate several complexity metrics in light of expert judgements on complexity, as made available in the WALS and AUTOTYP initiatives.¹

The paper is organized as follows: Section 2 offers a short introduction to the measurement of linguistic complexity. Section 3 describes the corpus and our sub-sampling strategy. We introduce in the first part of Section 4 the grammar-based complexity indices that provide the expert background against which our corpus-based study is elaborated and interpreted. In the second part of Section 4, we describe the four indices of morphological complexity assessed in the paper, along with the methodology implemented to test its robustness. Results on morphological complexity are then provided. In Section 5, we study the interaction between the within-word and across-word information distribution, taking advantage of the Word Information Density index introduced in Section 4. In Section 6, a proof of concept is proposed and evaluated to alleviate the need for a parallel dataset in quantitative typological studies. The paper ends with a general discussion and a few perspectives.

2. Linguistic complexity across languages: A short overview

In the last 25 years, the notion of linguistic complexity has pervaded most language sciences, with various foci on its phylogeny and ontogeny (e.g. Givón 2009), its online cognitive processing (e.g. Gibson 1998), or its distribution across the world's languages (see below). Numerous books and collective volumes have tackled these issues from various perspectives and epistemological stances, offering comprehensive reviews (e.g. Dahl 2004; Hawkins 2004; Kortmann & Szmrecsanyi 2012; Kusters 2003; Miestamo et al. 2008; Mufwene et al. 2017; Trudgill 2011). In this short overview, we will focus on how linguistic complexity can be measured from a cross-language comparison perspective, a framework in which most developments pertain to the **phonological** and **morphological** domains.

Phonological complexity is often operationalized as the number of units (syllables, phonemes, features or gestures) or contrasts between units implemented in

1. The paper is provided with a Rmarkdown html document incorporating the analysis code in R, the main results detailed in the paper, and results from additional analyses. This document is referred to as Supplementary Information throughout the paper. It is freely available in the GitHub repository (<https://github.com/yoonmioh/RobustMorphComp>, last access 2 December 2022).

a language (e.g. Maddieson 2009; Shosted 2006). This descriptive perspective can furthermore be supplemented by considering how much information is encoded in these contrasts, either in a paradigmatic (e.g. Oh et al. 2015; Wedel et al. 2013) or syntagmatic perspective (e.g. Coupé et al. 2019; Pimentel et al. 2020; Pimentel et al. 2021). Phonological complexity has nevertheless attracted less interest than morphological complexity so far, and more specifically than inflectional complexity (for recent reviews, see Arkadiev & Gardani 2020, Baerman et al. 2015).

At least since Joseph Greenberg's seminal study (Greenberg 1960), the striking differences observed across languages in their verbal and nominal inflectional systems have nurtured a very prolific literature on how to account for a language's inflectional complexity. In an influential paper, Ackerman & Malouf convincingly argued that this complexity goes beyond the size of inflectional paradigms and the number of distinct inflected forms, and should also consider to what extent predicting the inflected form for a given lexeme in a given slot of the paradigm is difficult (Ackerman & Malouf 2013). Their approach paved the way for many computational approaches based on morphologically tagged corpora or lexicons. However, "The complexity of morphological inflection is only a small bit of the larger question of morphological typology" (Cotterell et al. 2019: 339), and deriving morphologically tagged corpora from the kind of raw texts we address in this paper is still an open issue, despite recent progress (Erdmann et al. 2019; Malouf 2017).

Fortunately, even a raw text can help open a window on morphological complexity by applying agnostic methods focused on wordform frequency distribution or on derived indices, such as the Type-Token Ratio (TTR). This index finds its origin in psycholinguistic research in the middle of the twentieth century, and is meant to be "a measure of vocabulary 'flexibility' or variability" (Johnson 1944: 1). Defined as the ratio of the number of distinct wordforms over the text length (total number of word tokens), it is influenced by a given language's propensity to shape words through morphological processes such as inflection, derivation, reduplication or compounding. Its conceptual and practical simplicity led to its dissemination and adoption as a typological index of morphological complexity (Kettunen 2014) either in its original formulation or more elaborated ones such as MATTR (Covington & McFall 2010) or MTLD (McCarthy & Jarvis 2010). The word entropy (denoted H) is a somewhat similar index of morphological complexity stemming from the information-theoretical framework. It is influenced by both the number of different wordforms in a text and the distribution of their relative frequency (see Section 4.2 for details). TTR, MATTR, MTLD, and H are known as distribution-based indices, following the terminology proposed by Bentz et al. (2016). These authors also introduced a translation-based index conceptually similar to the comparison of the relative lengths of the same

text translated in several languages. Being intrinsically comparative, this kind of approaches provide an interesting way to compare how different languages distribute an identical semantic content over its words.

Finally, other approaches have been elegantly derived from the concept of Kolmogorov complexity (Juola 1998). By comparing the mathematical compressibility of different distorted variants of the same text, these **distortion/compression** methods aim to quantify morphological and syntactic complexities in terms of within-word and inter-word regularities respectively (Ehret & Szmrecsanyi 2016; Juola 1998; Moscoso del Prado 2011).

All of these distribution-based, translation-based, and distortion/compression-based approaches are blindly applicable to raw text corpora. In Section 4, we implement TTR and H as they are the most widespread and standard complexity metrics used in crosslinguistic studies. They consequently provide the baseline for our robustness study. We also implement the MTLD index because it is supposed to overcome some of the known shortcomings of TTR (and especially its dependency to the text length) without requiring any additional parameter tuning, which is an advantage over MATTR. These distribution-based indices are then compared to Word Information Density (WID), a translation-based index. Finally, a distortion/compression method is applied in Section 5 to investigate the relative importance of the within-word and inter-word information and their potential trade-off.

3. Corpus description and subsampling strategy

We used a subset of the Parallel Bible Corpus (Mayer & Cysouw 2014) which was selected and preprocessed by the organizers of the Interactive Workshop on Measuring Language Complexity (IWMLC2019), with additional preprocessing by the authors for Burmese,² Egyptian Arabic, and Persian.³ The dataset⁴ contains 1,150 verses which are fully parallelized with maximum overlaps across 47 typologically and geographically diverse languages (Figure 1a). No linguistic annotation is included. However, since it is fully parallelized, we assume that the semantic content S of each verse v in a language L is equivalent for all languages (S^L). By equivalent, we mean that each verse translation is likely to convey the

2. Segmentation into words was performed by the authors with the Myan-word-breaker library (downloaded at <https://github.com/stevenay/myan-word-breaker>, last access 2 December 2022).

3. Punctuation marks were removed by the authors for Egyptian Arabic and Persian.

4. http://www.christianbentz.de/MLC2019_data.html, last access 2 December 2022.

same meaning on average, despite fluctuations (but see Christodouloupoulos & Steedman, 2015 for a discussion on the translation process).

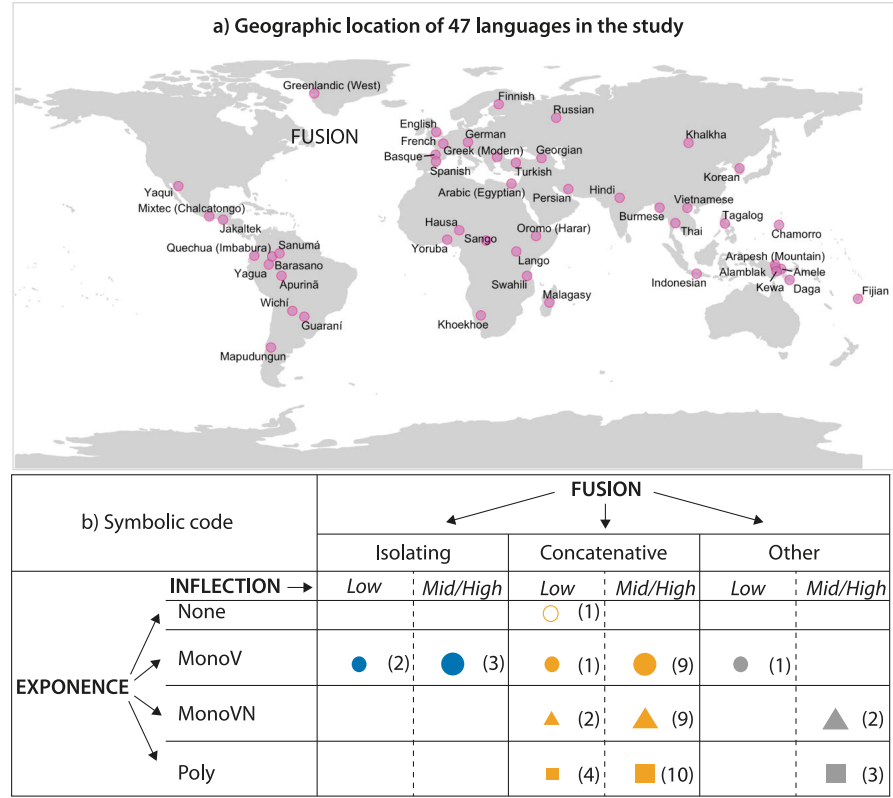


Figure 1. (a) Geographical distribution. (b) Distribution of the languages among WALS classical typological features and symbolic codes. Marker color and shape respectively encodes the **fusion** strategy and the **exponence** category. Marker size further indicates whether **verbal inflection** is limited (small size for Low values) or more extended (large size for Mid and High values). In each cell, the number of languages is displayed when different from zero. See text for details and Supplementary Information (§3) for the list of languages belonging to each category

We have selected this corpus because it offers a sound balance between its coverage in terms of linguistic diversity (both typological and areal, see Figure 1) and its size, in coherence with the linguistically-informed and quantitative approach adopted here. Our rationale is that this dataset size offers a testbed that is large enough to reveal or confirm cross-language trends while allowing to individually tag all languages in each figure. While it is also easily extendable to other lan-

guages available in the Parallel Bible Corpus, we also decided to conform to the selection made for the IWMLC2019 workshop to guarantee the coherence with other studies initiated in this framework (cf. Bentz et al. 2022, Çöltekin & Rama 2022; Gutierrez-Vasques & Mijangos 2020). As explained in the introduction, such a parallel corpus is far from being ideal for comparing spontaneous language complexities, but it is well-adapted to our purpose of assessing the robustness of complexity indices by the subsampling approach we propose in this paper.

The first aim of our study is indeed to assess whether several corpus-based metrics (described in Section 4) are robust complexity indices. Throughout the study, we thus implement an iterated subsampling strategy by comparing, for each metric, the values estimated from the whole corpus to repeated values estimated by splitting the corpus into several smaller subsets. This way, we can both evaluate each index sensitivity to subset length and, for a given length, its sensitivity to differences in the texts themselves. More specifically, we proceed with an increasing number of subsets extracted from the Parallel Bible Corpus. For a given subset number and size, each subset contains the same number of verses across languages. The indices' robustness is thus assessed by comparing between six different configurations with respect to the total number of subsets: whole (1,150 verses), 5 (230 verses in each dataset), 10 (115 verses), 20 (57 verses), 40 (28 verses), and 60 subsets (19 verses). With such a procedure, a complexity index is considered robust if it is characterized by a stability of the language-specific values across the six configurations, along with a limited standard deviation within each configuration. Additionally, one also expects that a robust index would keep the language ranking constant across the configurations, even if the index values are prone to a limited fluctuation.

Two metrics of morphological complexity are acquired and computed from a top-down typological approach by means of two linguistic databases: Grammar-based Morphological Complexity derived from WALS (GMC_W) and Grammar-based Morphological Complexity derived from AUTOTYP (GMC_A), and a bottom-up corpus-based approach is additionally applied to the Parallel Bible Corpus to compute four indices: Word Information Density (WID), Type-Token Ratio (TTR), Measure of Textual Lexical Diversity (MTLD), word-level Entropy (H).

In addition, we adopt a symbolic code to visually represent the languages throughout the paper against a traditional morphological typology backdrop (Figure 1b). For each language marker, its color, shape, and size respectively encodes **Fusion** strategy (derived from WALS Chapter 20), **Exponence** (derived from WALS Chapter 21B), and the amount of **Verbal inflection** (WALS Chapter 22).

Finally, information on phonological complexity is also provided though only tangential to our study (see Supplementary Information § 5.2.1).

4. Morphological complexity

4.1 Grammar-based morphological complexity indices

4.1.1 Methods

The measure of Grammar-based Morphological Complexity derived from WALS (GMC_W) is adapted from the methodology proposed by Lupyan & Dale (2010) where 29 linguistic features relevant to the inflectional morphology are chosen from WALS. The score of GMC_W is calculated by distinguishing between lexical and inflectional coding strategies and summing assigned values (−1 for lexical and 0 for morphological strategies) to the linguistic features which are accounted for by continuous or categorical variables. (See Supplementary Information § 4.1.1 for details.) Following the method in Oh (2015), the current study does not convert continuous variables into dichotomous variables, which differs from Lupyan & Dale (2010). Instead, continuous variables, such as the number of case categories (WALS feature 49A) and the number of grammatical categories expressed by the inflectional synthesis of the verb (WALS feature 22A), are normalized between 0 and −1 to better represent the degree of morphological complexity. The score is obtained by dividing the overall sum by the total number of available linguistic features in each language.⁵

Following Kettunen (2014) and Sinnemäki & Di Garbo (2018) (see also Nichols & Bentz 2019), another Grammar-based Morphological Complexity metric is derived from AUTOTYP (GMC_A). More specifically, we integrate the degree of inflectional synthesis of verbs. One of the linguistic features aforementioned and derived from WALS (Inflectional synthesis of the verb, 22A) was originally based on this dataset of AUTOTYP (Bickel & Nichols 2013) and classified languages into seven categories according to the number of grammatical categories which can be expressed by a maximally inflected verb form. Since such classification results in a loss of information as described by Sinnemäki & Di Garbo (2018), we decided to use the actual maximum number of inflectional categories per word. For Malagasy and Wichí, two languages with missing information, we used a score obtained from their variant language present in AUTOTYP with a distinct Glottocode (Malagasy (mala1537) and Mataco (wich1263), respec-

5. On average, there are 27 linguistic features available (out of 29) per language, with a minimum number of 23 features for Mountain Arapesh and all 29 features available for 16 languages.

tively). To summarize, the dataset is studied in the light of both a comprehensive morphological complexity index (GMC_W) and a more specific one (GMC_A), in order to provide two levels of granularity in expert-based indices.

4.1.2 Crosslinguistic overview

On a WALS-based MC scale (GMC_W index, Figure 2), languages distribute from -0.92 (Vietnamese) to -0.34 (Turkish), arguably revealing three groups according to their dominant strategy: languages with MC values higher than $-\frac{1}{2}$ favor inflectional coding strategies (hence MORPH strategy). Conversely below $-\frac{2}{3}$, we see that languages favoring lexical coding strategies (hence LEXICAL) span over a larger range of values. A third group consists of languages for which a more balanced strategy is observed (hence BALANCED). Please note that this tri-
chotomy is indicative and does not pretend to identify clear-cut boundaries along the GMC_W scale but rather a hint on each language's strategic orientation. Additionally, one can expect that adding more languages to the dataset would both fill the gaps within and extend beyond the current distribution.

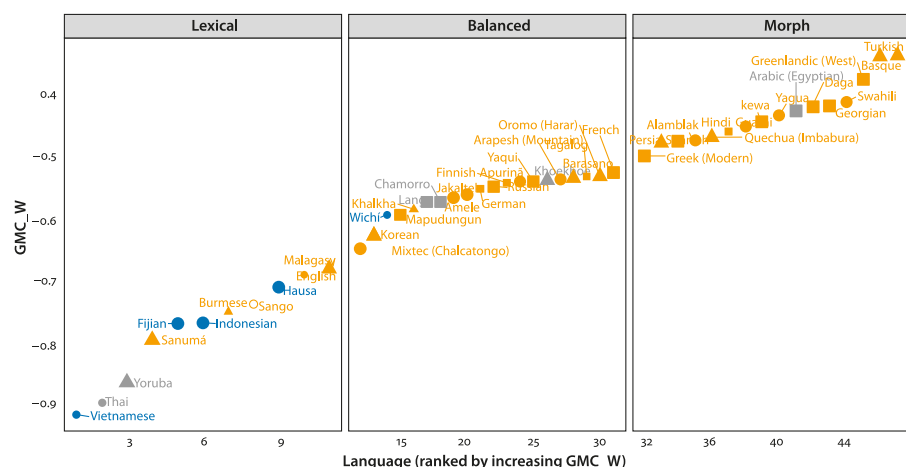


Figure 2. Grammar-based Morphological Complexity based on WALS (GMC_W). On the x-axis, languages are ordered by increasing GMC_W values from left to right. GMC_W is by definition normalized between -1 and 0 . Marker convention is shown in Figure 1b (also in Supplementary Information § 8.1)

Unsurprisingly, the GMC_W index is visually coherent with the WALS feature encoded in each language's color: Isolating languages (in blue) lead to low complexity values (but note the exception of Wichí) while Concatenative languages (in orange) tend to show higher values. In contrast, the connections between the complexity index and the exponence (encoded by the marker shape) or the

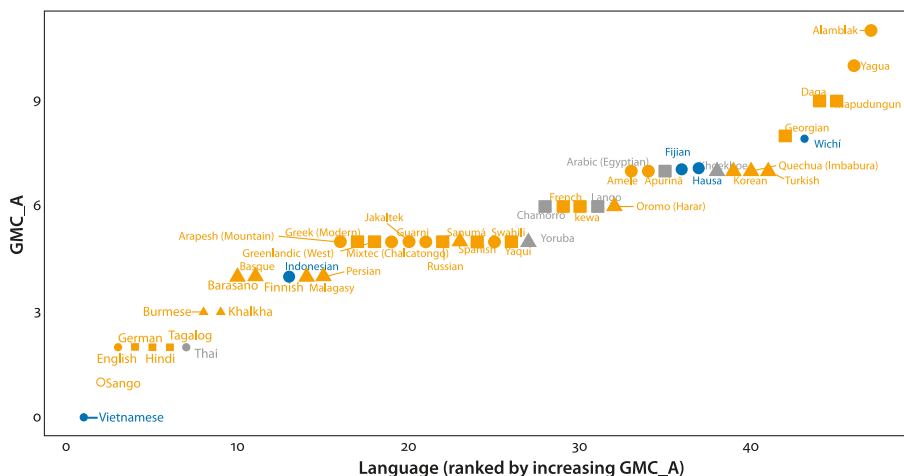


Figure 3. Grammar-based verbal inflectional complexity based on AUTOTYP (GMC_A). On the x-axis, languages are ordered by increasing GMC_A values from left to right. Marker convention is shown in Figure 1b (also in Supplementary Information § 8.1)

amount of verbal inflection (encoded by the marker size) are less straightforward and require a more detailed analysis.

On an AUTOTYP-based MC scale (GMC_A, Figure 3), languages are distributed from 0 (Vietnamese) to 11 (Alamblak). At first glance, it appears that the grammatical information captured by the index is quite different from the WALs-based analysis, with for instance Yoruba ($GMC_W = -0.87$) and Greenlandic (West) ($GMC_W = -0.37$) sharing the same GMC_A value (5) with ten other languages, despite their differences in terms of traditional morphological typology.

This first impression is confirmed by a Bayesian correlation analysis.⁶ The Bayes factor ($BF = 13.19$) indicates strong evidence toward a relatively low positive correlation found between the two indices ($r_{\text{median}} = 0.37$, 95% CI [0.12, 0.58]) as illustrated by the scatter plot in Figure 4. While a few languages oriented towards

6. All correlations reported in this paper are estimated in a Bayesian framework with the BayesFactor and bayestestR packages in R. We used the default prior options of the following two functions (correlationBF and describe_posterior) for the correlation analysis (cf. see Supplementary Information § 4.3.2 for the code). Each correlation is reported as the median Bayesian posterior estimate r_{median} , along with the 95% credible intervals for each correlation coefficient under a two-sided alternative hypothesis. The BayesFactor (BF) in support of the alternative hypothesis (viz. the existence of a correlation) is also reported. $BF > 10$ indicates a strong support in favor of the existence of a correlation. $3 < BF < 10$ indicates a moderate support while BF values between one and three are considered weak.

a LEXICAL coding strategy share low values on both axes (Vietnamese, Thai, Sango, and English), most languages are spread over a large surface, underlying the highly multidimensional nature of morphological complexity. Isolating languages (in blue) also show an interesting pattern: they occupy a quite large portion of the total variation range observed on both axes, and their distribution suggests the existence of a higher correlation than for the other languages (not statistically tested here because of the very limited subsample of isolating languages). This distribution more largely underlines that the form taken by morphological complexity across the world's languages is quite variable because of the numerous dimensions involved, their interactions, and their potential relative weights in complexity metrics. Moreover, even the language classification in broad categories can be debated and three of the languages considered here as isolating (following WALS) would be considered morphologically complex in other accounts (Fijian: Aranovich 2013; Hausa: Newman 2003; Wichí: Nercesian 2014). We leave for future work a thorough study of these aspects and simply emphasize that GMC_A and GMC_W give only partial visions of the processes of word formation and usage at play in each language. Interested readers can further find a breakdown of the complexity distribution by more specific descriptors (Morphological Strategy, Fusion, Verb Inflection, Syllable Structure, and Tonal system) in the Supplementary Information (§ 4.3.1).

4.2 Towards robust indices of morphological complexity

4.2.1 Methods

The following four metrics estimate morphological complexity by means of the Parallel Bible Corpus. From an information-theoretic angle, the average amount of information conveyed per unit of language (such as word in this study) WI can be defined in Equation (1) as the semantic content S of verse v in language L (S_L^v) divided by the number of its words (N_L^v).

$$(1) \quad WI_L^v = \frac{S_L^v}{N_L^v}$$

By adopting the method in (Oh 2015; Pellegrino et al. 2011), Word Information Density (WID) is calculated by a pairwise comparison between the number of words in our reference language and a target language L . English is used here as a reference language, because it offers a convenient background shared all over the world. Since we use a parallel corpus, the semantic content S of each verse is assumed to be equivalent across all languages ($S_L^v = S_{ENG}^v$). Therefore, WID is estimated by a pairwise ratio between the number of words of verse v in English

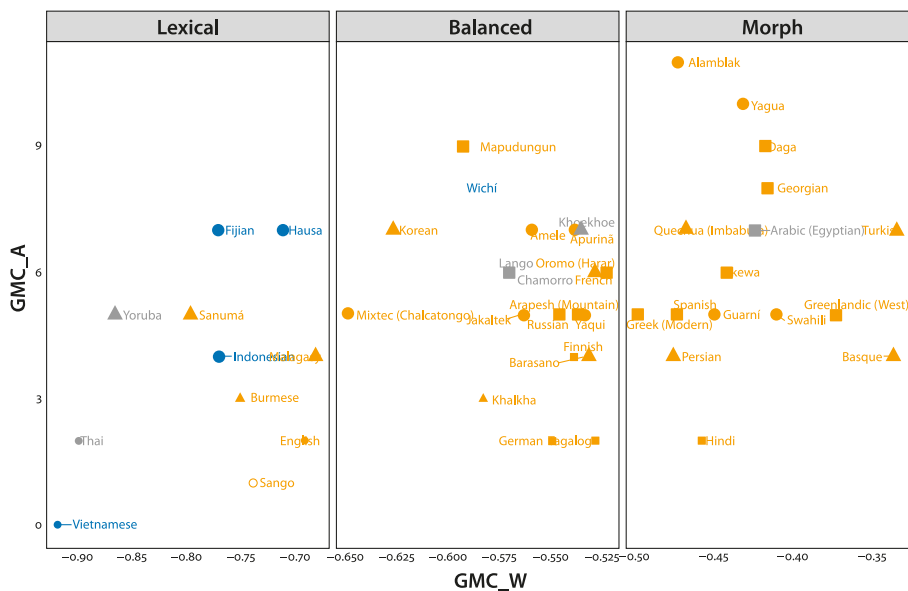


Figure 4. Language distribution in a two-dimensional space defined by Grammar-based Morphological Complexities based on WALS (GMC_W, x-axis) and AUTOTYP (GMC_A, y-axis). Marker convention is shown in Figure 1b (also in Supplementary Information § 8.1)

(N_{ENG}^v) and in a target language (N_L^v). At a global level, for each language L , WID_L is obtained by averaging over the V verses in the dataset:

$$(2) \quad WID_L = \frac{1}{V} \sum_{v=1}^V \frac{WI_L^v}{WI_{ENG}^v} = \frac{1}{V} \sum_{v=1}^V \frac{S_L^v}{N_L^v} \times \frac{N_{ENG}^v}{S_{ENG}^v} = \frac{1}{V} \sum_{v=1}^V \frac{N_{ENG}^v}{N_L^v}$$

In contrast, the following three corpus-based metrics are directly calculated at the level of the entire text, without averaging nor normalizing with reference to English.

Type-Token Ratio (TTR) measures the degree of lexical diversity and morphological productivity by means of vocabulary size (number of unique word types) and text length (total number of word tokens). One should note that it is known to be influenced by the text length (Covington & McFall 2010). Following other similar studies using parallel corpus (Bentz et al. 2016; Gutierrez-Vasques & Mijangos 2020, inter alia), we do not try to match text lengths across languages because they are mainly influenced by each language's morphology and wordhood, the semantic content being kept equivalent. TTR is calculated for language

L as the ratio of the number of unique word types T_L over the word token count N_L in the dataset:

$$(3) \quad TTR_L = \frac{T_L}{N_L}$$

A higher score of TTR denotes there are more word types generated by a language, which approximates its morphological productivity and complexity. However, as the text length gets longer, the likelihood of encountering a new word type in the text decreases.

Another measure of lexical diversity, Measure of Textual Lexical Diversity (MTLD) has been proposed by McCarthy & Jarvis (2010) as a solution to the length-sensitivity problem of TTR. MTLD is defined as the average number of words required within a text to reach the same TTR value of 0.72 (the threshold point of stabilization established in McCarthy & Jarvis 2010). MTLD is calculated by using a code released by John Frens.⁷ A higher score of MTLD means the text requires more words to reach the point of stabilization and therefore, it can be considered more complex in terms of morphological productivity. The advantage of MTLD is that it was designed to be less prone than TTR to variations in text length even on small dataset ranging from 100 to 2k words, thanks to its intrinsic averaging nature (McCarthy & Jarvis 2010). We thus hypothesize that our study will confirm this enhanced robustness.

Word-level (unigram) Entropy (H) is defined as the average amount of information (or unpredictability) of words and it can be estimated by Equation (4), where Language L consists of a finite set of T_L unique word types $\{W_1, \dots, W_{T_L}\}$ and P_{w_i} is the probability of i th word type, estimated from the corpus.

$$(4) \quad H_L = - \sum_{i=1}^{T_L} P_{w_i} \times \log_2(P_{w_i})$$

It is another measure of morphological complexity which is also sensitive to the corpus size, as the accuracy of estimating the distribution of word probabilities depends on the corpus size. On the one hand, in Oh (2015), syllable unigram entropy has been shown to be more robust as the corpus size grows, with a range of convergence threshold between 50k and 70k word tokens for 4 languages. On the other hand, in Bentz & Alikaniotis (2016), convergence points for the word unigram block entropy of 21 languages have been established with text sizes between 20k and 60k word tokens, with an average of 38k. Since the average

7. Downloaded from https://github.com/jennafrens/lexical_diversity/, last access 2 December 2022.

number of tokens per language in our parallel corpus is 25.4k, with a maximum of 57.3k for Sanumá and a minimum of 10.8k for Greenlandic (West), our study sheds light on the extent to which those limited lengths affect the reliability of H .

4.2.2 Results: Type-Token ratio and entropy

TTR is easy to compute and its interpretation is rather intuitive, but it is known to be prone to fluctuation caused by variations in the examined corpus. Our sensitivity methodology based on subsampling the data set in a variable number of subsets illustrates this trend. For each language, increasing the number of subsets (and coincidentally reducing their size) increases both the overall range and the variation in TTR among the subsets for a given language. This is the combined effect of estimating TTR on smaller samples and of the inherent differences existing among the samples themselves, even though they are all extracted from the same source text. Figure 5 illustrates this effect according to the subset configuration (Whole, 5, 10, 20, 40, and 60 subsets).

A more important consequence is that the relative ranking of the languages is highly dependent on the corpus itself, as illustrated at the end of the present section in Figure 9 (leftmost panel). Each dot represents the rank (from 1 for the highest TTR to 47 for the lowest, y-axis) obtained by one language in one subset configuration on average (Whole dataset, 5, 10, 20, 40, and 60 datasets, x-axis), and for each language, an edge connects its ranks across configurations if they are not identical. Conversely, perfectly consistent rankings are displayed as gray dots and labels on the figure, without any edge. The desirable property of a consistent index, immune to fluctuation, would be to maintain rank estimations similar (leading to almost horizontal segments for small variations or no segment at all for identical ranks), which is clearly not the case with TTR, where non-marginal changes are visible for the majority of the sample.

Applying the same methodology to the word-level Entropy H yields similar results and observations. When the Whole dataset is considered, H values spread over a quite large range and seem to be quite informative on each language, but the smaller subsets are considered, the more overlap is present (Figure 6). In turn, this instability results in abundant reorganizations of the language ranking throughout different sampling configurations, as shown in Figure 9 (second panel).

4.2.3 Results: Measure of textual lexical diversity and word information density

The Measure of Textual Lexical Diversity proposed by McCarthy & Jarvis (2010) is conceived to be less prone to fluctuations induced by changes in corpus lengths. It is nevertheless somehow sensitive to variations in the corpus content itself, as

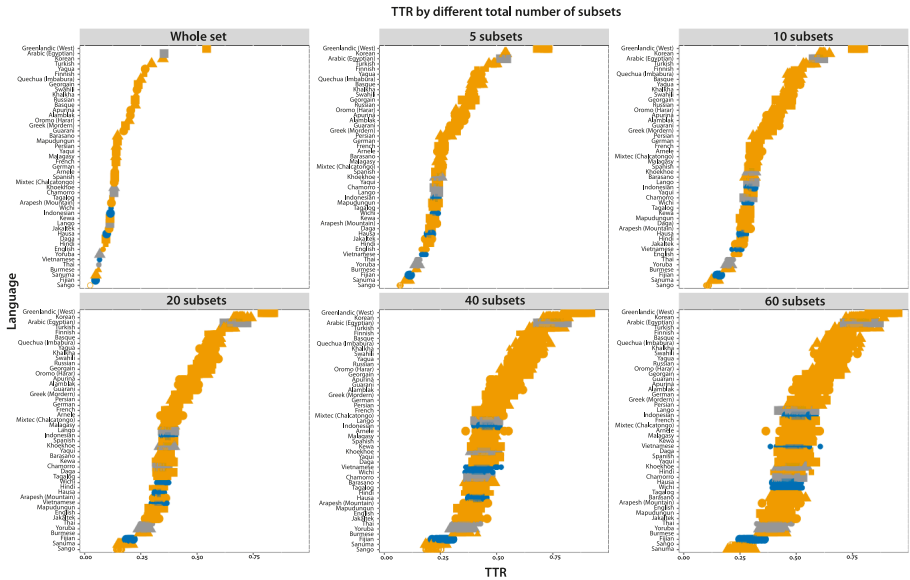


Figure 5. Languages ranked by Type-Token Ratio (TTR, x-axis). Each panel corresponds to a different corpus sampling configuration, from one unique sample (Whole set, top left panel) to 60 samples (bottom right panel). In each panel, languages are ranked by average TTR over the subsets, potentially leading to differences in ranking across the panels

illustrated when comparing the MTLD range and variation in estimated values resulting from different subset configurations (see Figure 7). However, these variations are mostly limited to few languages, principally Greenlandic (West), and Arabic (Egyptian) and yet, it is interesting to note that the range of across-language difference in MTLD is also quite limited and on par with the within-language fluctuation for a large proportion of the sample. Despite those variations, MTLD is quite consistent in the relative rank allocated to each language, as shown by the small changes in rank visible between the Whole vs. 60-subset configurations, consisting mostly of swaps between adjacent languages on the MTLD scale (see Figure 9, third panel).

TTR, H , and MTLD are based on the mathematical properties of the word-form distribution in a corpus. On the contrary, WID approximates the difference between the morphological strategy of a given language and the strategy at work in the English corpus: a language for which each English word is consistently translated into two words would get a WID of $\frac{1}{2}$ and its words would be considered less dense than their English counterpart in terms of information encoding. Is such an index prone to or on the contrary resistant to variations in corpora?

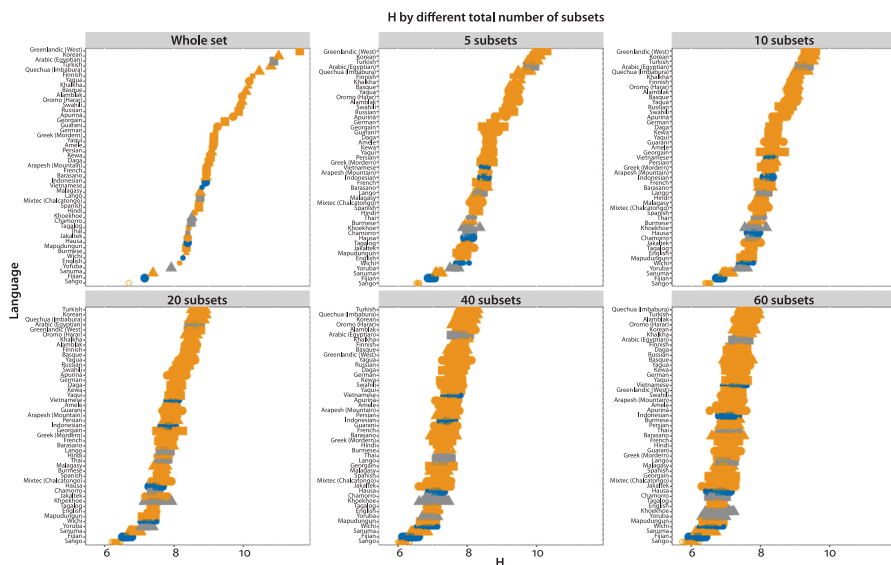


Figure 6. Languages ranked by word-level Entropy (H , x-axis). The Figure convention is the same as in Figure 5

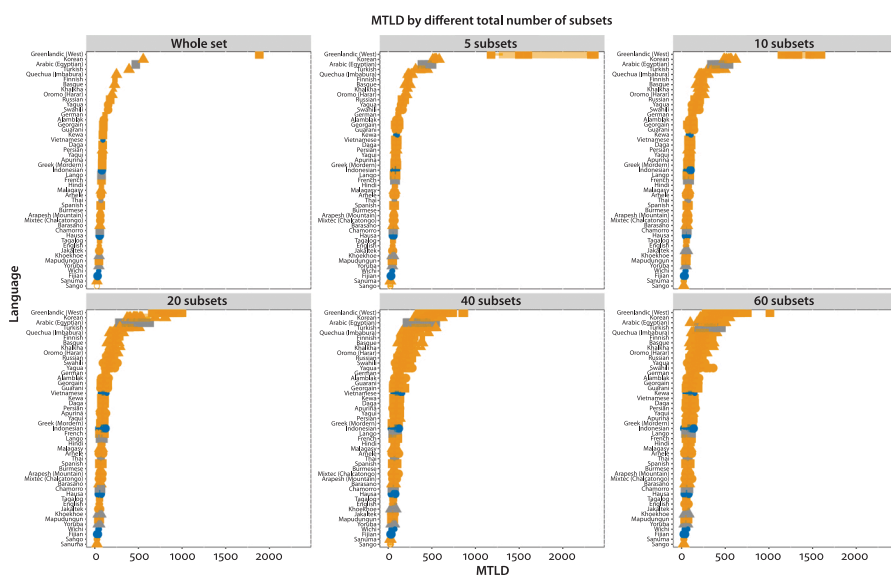


Figure 7. Languages ranked by Measure of Textual Lexical Diversity (MTLD, x-axis). The Figure convention is the same as in Figure 5

As shown in Figure 9 (rightmost panel), WID is quite consistent, with very few changes in language ranks induced by changes in the corpus configuration. Moreover, languages are more evenly distributed on a range quite stable across configurations except for Greenlandic (West), which is a potential clue of an interesting **discriminative power** to differentiate languages, even if a large number of small subsets are considered (see Figure 8).

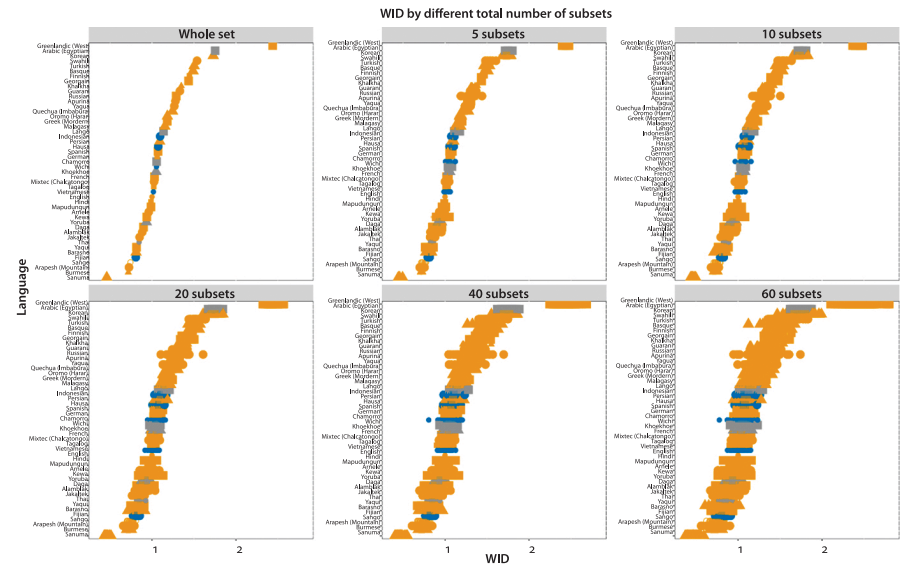


Figure 8. Languages ranked by Word Information Density (WID, x-axis). The Figure convention is the same as in Figure 5

4.3 Comparing corpus-based and grammar-based indices

In the previous sections, we showed that standard indices such as TTR and H are quite sensitive to changes in the material they are estimated on, in contrast with MTLT and WID, which seem more resilient to the sensitivity test we implemented. However, there is no doubt that TTR and H can be correctly estimated on large corpora and as such provide useful information on the range of variation in morphological complexity among a language sample, as shown by Kettunen (2014) among others. In a sample of more than 500 languages, Bentz et al. (2016) furthermore found a positive correlation of $r_p=0.402$ between a morphological complexity index based on WALS⁸ and TTR (and similarly a correlation of

8. The WALS-based index in Bentz et al. (2016) is different from the one we implemented in the present study since they took a variable number of morphological features per language,

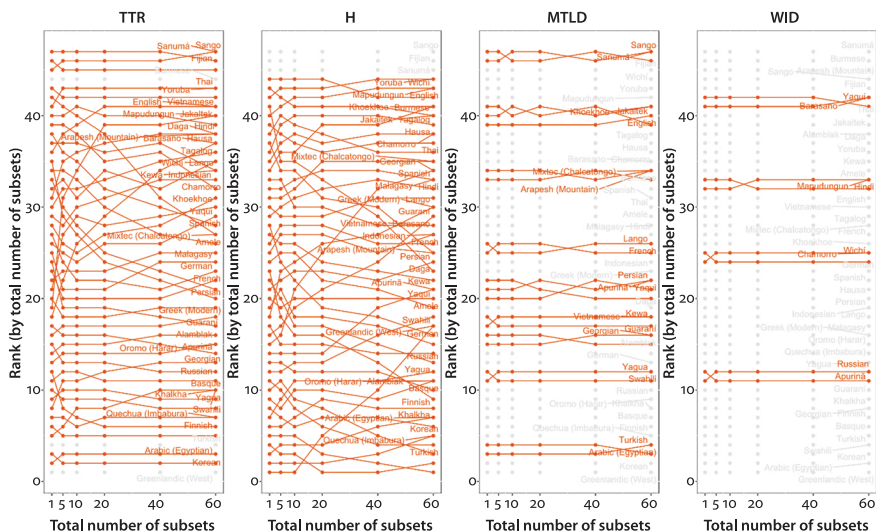


Figure 9. Impact of the sampling configuration on the four complexity indices (TTR: Type-Token Ratio; H : Word-level Entropy; MTLD: Measure of Textual Lexical Diversity; WID: Word Information Density). In each panel, the y-axis shows the language ranks according to the sampling configuration (whole set, 5, 10, 20, 40, and 60 subsets, on the x-axis). Languages are displayed in gray when rank is preserved throughout all configurations and in orange when changes occur, with orange edges underlying the changes

$r_\rho = 0.362$ with H), demonstrating that these corpus-based indices can be informative about a language's morphological strategy. This potential is confirmed in our sample when for each language, the whole dataset is used without sub-sampling to evaluate the indices (see Supplementary Information §4.3.1 for a global overview of the correlations observed across the indices over the Whole set sampling configuration and §4.3.2 for Bayesian correlation analyses). A median Bayesian posterior estimate $r_{\text{median}} = 0.55$, 95% CI [0.29, 0.72], BF = 3514, (resp. $r_{\text{median}} = 0.58$, 95% CI [0.32, 0.73], BF = 7720) – higher than the one reported in Bentz et al. (2016) – is present between GMC_W and TTR (resp. H). The existence of a moderate positive correlation is also strongly supported between WID and GMC_W ($r_{\text{median}} = 0.46$, 95% CI [0.24, 0.65], BF = 137.79) while the Bayesian

because of the very large language sample they investigated. As explained in Section 4.1, the number of morphological features taken into account in our study is less variable, ranging from 23 features for Mountain Arapesh to 29 features for 16 languages. For this reason, GMC_W is probably more consistent and comparable throughout our language sample, at the expense of a narrower coverage of linguistic diversity.

analysis brings only a weak support to the existence of a correlation between MTLD and GMC_W ($r_{\text{median}} = 0.29$, 95% CI [0.01, 0.52], $\text{BF} = 3.04$). Interestingly, the Bayesian analysis does not support the existence of correlations between GMC_A and neither WID ($r_{\text{median}} = 0.13$, 95% CI [-0.14, 0.38], $\text{BF} = 0.496$) nor MTLD ($r_{\text{median}} = 0.05$, 95% CI [-0.22, 0.31], $\text{BF} = 0.347$), underlying the multidimensional kinds of grammatical information subsumed in these corpus-based indices. Moreover, the four corpus-based indices are substantially positively correlated (see Supplementary Information § 4.3.1), and they consequently partially encode redundant information. In the rest of the paper, we focus on WID, to further test whether this innovative index brings new opportunities and understanding on language complexity.

5. Beyond word complexity

Morphological complexity is only one facet of the overall complexity of a linguistic system. As mentioned in the introduction, several studies have revealed a compensatory relationship between the complexity (and thus informativeness) within words (morphology or word constituency) and across words (syntactical or sentence constituency). In this section, we investigate how WID interacts with an inter-word index of complexity and information (Inter-Word Information, or IWI, see below) in order to test whether this seemingly robust index confirms previous results on compensation.

5.1 Methods

Inter-Word Information (IWI) estimates the amount of information across words by measuring the average compression ratio for each language L , i.e., the change in the size of compressed text files C in Language L before (CA_L) and after (CP_L) distorting word order (by a random permutation). In order to increase robustness, this procedure is iterated 10 times per language and the average value $\text{Mean}C_L$ is reported. This estimation method of Kolmogorov complexity for linguistic description was introduced by Juola (1998) and adopted for cross-linguistically comparing (pseudo-)syntactic complexity (see also Kettunen et al. 2006). It is based on the ability of the compression algorithm to achieve higher compression rates for texts exhibiting regularities in their word order, compared to texts where the transitions between adjacent words are difficult to predict.

$$(5) \quad \text{Mean}C_L = 1 - \frac{1}{10} \sum_{i=1}^{10} \frac{CA_L^i}{CP_L^i}$$

In essence, one expects that compressibility will be almost unchanged between the original and the randomized versions in a putative language with a totally free word-order. Such a language should result in a ratio between CA_L and CP_L close to one, and consequently to a $MeanC_L$ close to zero. Conversely, compressibility should be dramatically altered in a language with a strict word order, leading to a $MeanC_L$ value less than one (the compressed randomized version being on average larger than the compressed original text).

English being used as a pivot language in our study, IWI is computed by a pairwise comparison between the average compression ratio in a target language ($MeanC_L$) and in English ($MeanC_{ENG}$) in order to get a normalized value.

$$(6) \quad IWI_L = \frac{MeanC_L}{MeanC_{ENG}}$$

Given the abundant literature suggesting the existence of a trade-off in how a language weighs the within-word and across-word dimensions in complexity, we expect a negative correlation between WID and IWI, probably quite strong because WID is expected to be quite robust to idiosyncratic fluctuations across languages.

WID and IWI respectively quantify whether a given language relies more or less than English on the within-word and across-word dimensions. Since both indices are dimensionless and normalized with regard to English, by definition $WID_{ENG} = IWI_{ENG} = 1$ and we can take advantage of this normalization to compose an index that accounts for within-word and across-word dimensions together, maintaining English as a reference. We thus define Language EXplicitness (LEX) as the product of WID with IWI.

$$(7) \quad LEX_L = WID_L \times IWI_L$$

A language with both WID and IWI larger than one automatically yields a LEX value larger than one. Conceptually, it means that such a language **explicitly** encodes more information in its text than English. Assuming a negative correlation between WID and IWI analog to the one found in previous studies, we expect that the LEX variation range will be rather limited, but still present because of crosslinguistic variation in the components that are obligatorily expressed (see Bisang 2014, 2015 for discussions of the notion of overt vs. covert complexity) and in average speech rates which can be seen as the other aspect of information transmission (Coupé et al. 2019; Pellegrino et al. 2011). In the absence of any specific prediction on the position of English in the potential range of variation of LEX, we expect that languages will distribute on a limited range below and above one.

5.2 Results

Koplenig et al. (2017) have nicely shown at a large scale (more than 1,000 languages) that there exists a statistical trade-off in how information is split between across-word and within-word regularities (Spearman correlation of at least $r_\rho = -0.71$ depending on the dataset considered). Projecting our language sample in a two-dimension space defined by Inter-Word Information (IWI) and WID offers a similar balance, with a correlation of $r_{\text{median}} = -0.79$, 95% CI $[-0.88, -0.67]$, $\text{BF} = 6.24\text{e}+09$. Since both dimensions are normalized with regard to English (at coordinates [1,1] on Figure 10), one can formulate a few additional comments.

First, there exists a large range of variation on both axes: Greenlandic (West) conveys less than half of English-equivalent information on the across-word axis and more than twice on the within-word axis, while at the other end on the plot, Sanumá exhibits the opposite proportions. Greenlandic (West) is well-known to be a polysynthetic language, according to a traditional typological account, and its extreme position on this figure is thus not surprising. Sanumá on the other end, is a Yanomaman language classified as concatenative in WALS but “while not strictly an isolating language, Sanumá appears to be less polysynthetic than many Amazonian languages” (Derbyshire & Payne 1990: 246) and our analysis shows that it heavily relies on the word order to convey information. Three other languages show a remarkably high bias towards IWI: Sango is a tonal creole affiliated to the Ubangi languages (Thornell 1997) with little affixation. Being used as a lingua franca, Sango may have undergone changes under the kind of sociolinguistic pressures for learnability improvement discussed in Lupyan & Dale 2010; Sinnemäki & Di Garbo 2018; Thomason & Kaufman 1992; Trudgill 2001, and Wray & Grace 2007, among others. Burmese is well-known for its pervasive particles and the very productive reduplication at work, leading to a salient amount of regular collocations that are broken apart by the randomization process. Fijian is an Austronesian isolating language with a quite strict ordering of the constituents in the predicate and noun phrase structure (Dixon 1988). More generally, strictly isolating languages and languages with a non-concatenative morphology tend to be in the low-right quadrant of the figure.

Secondly, no language is characterized by low values on both dimensions (the rectangle defined by values below English on both axes is empty). Similarly, no language jointly exhibits high values on both axes. Still, there is room for cross-linguistic differences as illustrated by the dispersion around a theoretical hyperbolic regression line of equal explicit information amount (not shown on the figure).

Thirdly, the European languages present in the sample cover a quite limited area (roughly approximated by a rectangle with a diagonal defined by the Finnish-English segment), underlying the absolute need for addressing complexity issues from a broader perspective before drawing any conclusion on linguistic diversity and universal tendencies.

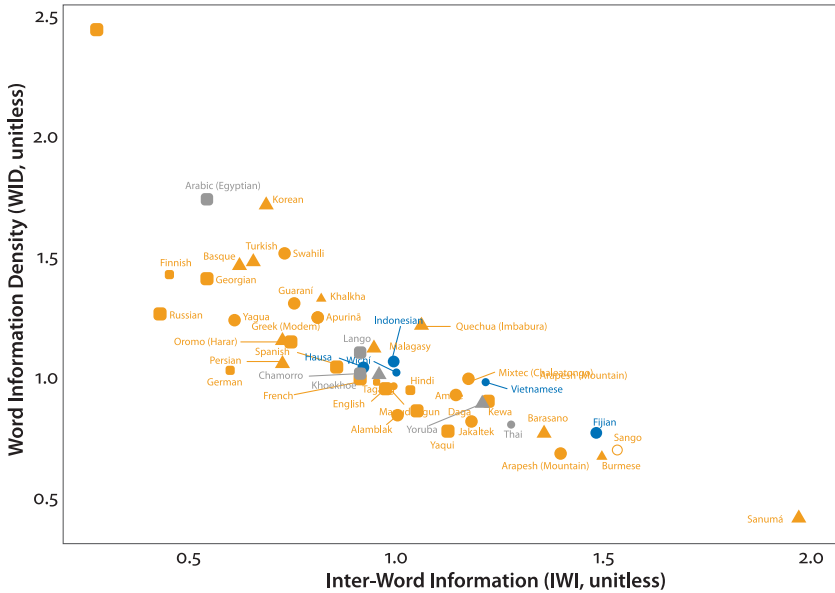


Figure 10. Distribution of the languages according to their information encoding strategies along the Inter-Word Information and Word Information Density (IWI and WID, respectively) dimensions. Marker convention is shown in Figure 1b (also in Supplementary Information § 8.1)

When we turn to the explicit information encoded (LEX), we see that English lies in the middle of the distribution (ranked 23/47) which spreads from 0.57 for Russian to 1.32 for Quechua (Imbabura), with a rather narrow peak (mean 0.99, median 1, sd 0.16). Highest values are reached by languages with a limited imbalance between the within-word and inter-word dimensions (Quechua (Imbabura): WID=1.24, IWI=1.06; Vietnamese: WID=1.02, IWI=1.22; Mixtec (Chalcatongo): WID=1.02, IWI=1.18). On the contrary, the two languages with **extreme** strategies (Greenlandic (West) toward word-internal information encoding (WID=2.44) and Sanumá toward Inter-Word Information encoding (IWI=1.97)) exhibit modest LEX values (0.7 and 0.91, respectively).

In addition, LEX is more strongly positively correlated to IWI ($r_{\text{median}} = 0.52$, 95% CI [0.29, 0.69], $\text{BF} = 818.41$) than negatively correlated to WID ($r_{\text{median}} = -0.24$, 95% CI [-0.48, 0.04], $\text{BF} = 1.55$), suggesting that languages encoding more explicit information rely more heavily on syntax (as approximated by IWI) than on morphology to do so. This is somewhat confirmed by the negative correlation also observed between LEX and GMC_W ($r_{\text{median}} = -0.36$, 95% CI [-0.57, -0.10], $\text{BF} = 10.58$) that illustrates that **highly explicit** languages tend to adopt more lexical than morphological strategies.

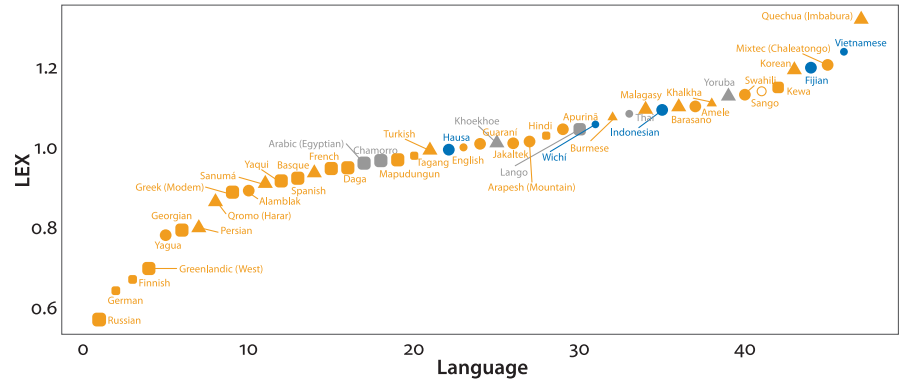


Figure 11. Languages ranked by increasing Language EXPLICITNESS (LEX). Marker convention is shown in Figure 1b (also in Supplementary Information § 8.1)

6. Breaking the parallel corpus barrier: A proof of concept

6.1 Experimental framework

Most studies on language complexity, including the one presented in the previous section, are based on **parallel** corpora with the assumption that they convey a similar informational content and can consequently reveal the genuine cross-language range of variation by neutralizing (or at least limiting) differences that would be induced by semantic dissimilarities in **non-parallel** corpora. A few studies have nevertheless ventured into non-parallel territories (e.g. Kettunen 2014), but comparing languages based on indices such as TTR remains a scaffolding whose robustness is challenged by the results shown in the Morphological section of this paper.

In this section, we investigate whether this parallelism constraint can be partially relaxed while preserving robust and informative indices, using our results on Word Information Density as a testbed. By definition a non-parallel dataset

induces the existence of crosslinguistic differences in the amount of semantic information encoded. The challenge in extending the WID approach to a non-parallel dataset is thus to neutralize those differences and, in this section, we propose a proof of concept based on the utilization of a Transformer language model to quantify this amount of semantic information (see below).

Figure 12 displays the implemented procedure. Three configurations are presented in the left (Full Parallel or FP), central (Pairwise Parallel or PP), and right (Non-Parallel or NP) panels, respectively. In each panel, the datasets are schematized by a box outlined in green, English (the reference language) is denoted in a shaded box and the other languages in boxes outlined in gray. The modifications in WID calculation induced by each configuration are shown in the formulas at the bottom of the panels.

In the **Full Parallel** configuration (left panel), the dataset is unique and shared by all languages, including English. This configuration is the one presented in the previous sections: a common semantic information S is shared by all languages. As explained in Equation (2), WID then simplifies to the ratio between the number of words in English and in any other language, and estimating S itself is unnecessary. In the following, this WID index resulting from this configuration is referred to as WID_FP in order to distinguish it from the other configurations introduced hereafter. The equation shown at the bottom of the panel summarizes WID_FP calculation and is identical to Equation (2).

In the **Pairwise Parallel** configuration (central panel), a **different** subset k is analyzed for each language k and matched to its translation in English. The semantic information is no longer the same **across** the languages, but for **each** given language k , it is shared with its English counterpart on subset k . WID_PP can be conveniently computed without explicit estimation of S_k (the semantic information in subset k) since it reduces as previously to the ratio of the number of words in subset k in English and in language k . One can nevertheless expect that differences across datasets will result in fluctuations in word information density estimations. The equation shown at the bottom of the panel summarizes WID_PP calculation and its simplification.

In the **Non-Parallel** configuration (right panel), each language, including English, is analyzed on a different subset. By definition, we now have different semantic information S_k for each language, including English. Computing WID_NP thus requires evaluating S_{ENG} for English and S_k for each other language k , since those semantic components do not simplify anymore when the ratio between language k and English is computed, as illustrated in Equation (8) where N_k and N_{ENG} refer to the number of words in different subsets assigned to languages k and English respectively:

$$(8) \quad WID_NP_k = \frac{S_k}{N_k} \times \frac{N_{ENG}}{S_{ENG}}$$

We thus need a way to estimate S_k . In the next subsection, we will show how semantic information can conceptually be connected to the information-theoretical notion of surprisal and how a recent Transformer language model can provide estimations of surprisal that satisfy our requirements.

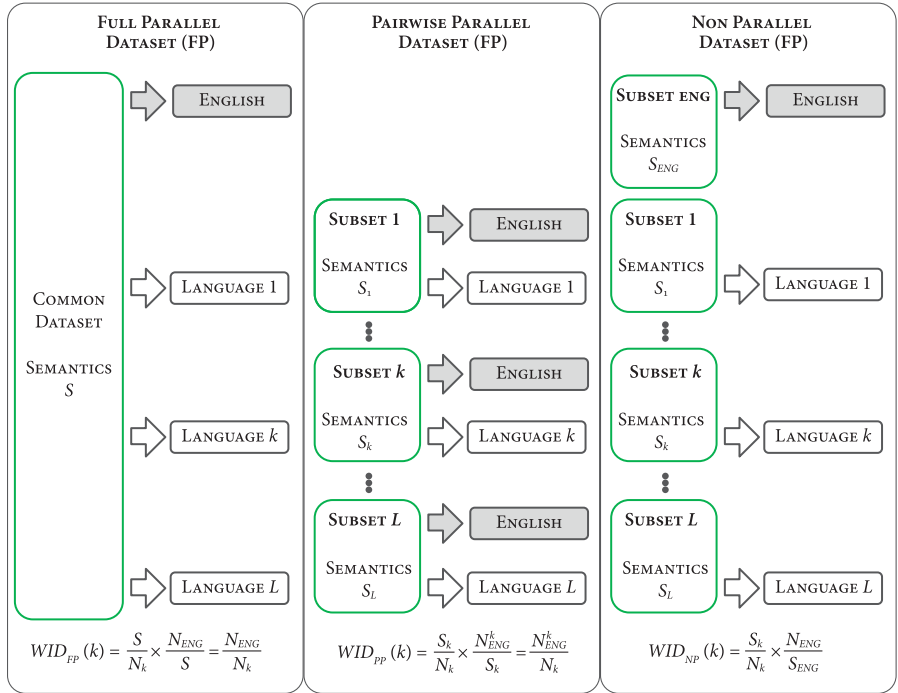


Figure 12. Schematic representation of the three cross-linguistic estimations of Word Information Density (WID) implemented. Left panel: Full Parallel configuration with a common dataset for all languages. Central panel: Pairwise Parallel configuration, with a shared subset for each language and its English translation but different subsets across languages. Right panel: Non-Parallel configuration, with a different subset for each language, including English

6.2 Evaluating information content

Several recent studies have convincingly showed that word-level and sentence-level surprisals estimated with state-of-the-art Transformer language models are intimately related to cognitive language processing, revealed by behavioral and electrophysiological cues in experimental comprehension tasks (Merkx & Frank

2021; Schrimpf et al. 2021; Wilcox et al. 2020). Word-level surprisal is inversely linked to the predictability of a word given its context and as such, it quantifies the information borne by this word and its cognitive processing cost (see Frank 2013, among many others).

A thorough analysis of these recent studies is beyond the scope of this paper but they provide a solid ground for considering that the surprisal estimated by such models is a good proxy of the amount of semantic information in a sentence or verse. However, despite the impressive results achieved by these language models on a few languages, their deployment in a multilingual context is still in its infancy (for a discussion, see Gerz et al. 2018; Hollenstein et al. 2021; Mielke et al. 2019, *inter alia*). This situation is rapidly improving but at the moment, the absence of a robust multilingual language model able to correctly handle the specific 47 languages considered here leads us to consider for each subset k that the semantic information S_k is correctly approximated by the surprisal estimated on its English version.⁹ There is undoubtedly a speculative dimension here, because using a text surprisal estimated in one language as a predictor of the surprisal of a translated version in another language is not routinely done, meaning that our approach is exploratory. This workaround is only temporary but we consider that it still enables us to assess this approach as a proof of concept, leaving improvements for further studies based on the ongoing research – especially on the multilingual prediction of human reading times (associated to cognitive processing and surprisal, see Hollenstein et al. 2022 for a recent evaluation).

As in the previous sections, we implemented a manifold repeated random sampling in order to get an estimation of the procedure robustness. WID_FP is estimated on 20 random subsets drawn from the whole corpus, as we did in one configuration of the previous sections. In parallel, the whole corpus is randomly split into $N=47$ subsets, corresponding to the 47 languages (including English). A twenty-fold permutation is then applied, in which each subset is randomly assigned to a language and WID_PP and WID_NP are calculated following Equations (2) and (8) in each permuted configuration.

9. More specifically, we estimated surprisal at the verse level with the *lm-scorer* package downloaded from <https://github.com/simonepri/lm-scorer>, last access 2 December 2022, using the GPT-2 model (Radford et al. 2019). More powerful language models have been released since then, but performance optimization is not the goal of this exploratory study.

6.3 Comparing information density estimations from parallel and non-parallel corpora

Figure 13 displays the similarities and differences induced by the experimental setting in terms of word information density ranking. In each panel, the left column indicates the original ranks assessed using the whole parallel corpus (these ranks are thus identical to the leftmost ones reported in Figure 9 in the WID panel) while the right column shows the average rank over the twenty-fold sampling for WID_FP, WID_PP, and WID_NP in left, central, and right panel respectively. By construction, the left panel shows results identical to the WID panel shown in Figure 9, with only three rank swaps between WID on the whole dataset and the 20-fold subset sampling. As expected, relaxing the parallelism constraints introduces larger differences: ranking is modified for 22 languages and 18 languages in the WID_PP and WID_NP configurations, respectively. Most alterations nevertheless remain between adjacent languages, except for German and Khoekhoe which nevertheless stay in the same central area of the distribution. In other words, WID estimations remain remarkably consistent despite introducing fluctuations due to the use of different subsets across languages.

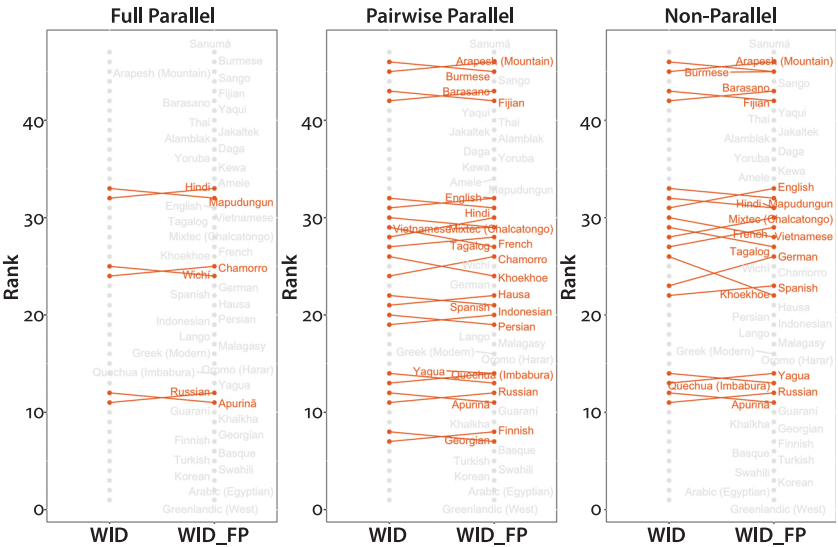


Figure 13. Comparison of the information density computed on the whole corpus (WID, left column in each panel) and the information densities implemented following Figure 12 (right column in each panel)

This result is corroborated by the scatterplot of the 20 subsets per language in the two-dimension space defined by WID_PP and WID_NP (Figure 14, background panel). The consistent ranking of the languages translates in a very high global correlation which is higher than one would get if semantic information was not taken into consideration, as revealed by comparing the actual correlation (red vertical line on the superimposed panel in Figure 14) to a distribution of simulated subsets with a 1000-fold shuffling of the semantic content associated with each subset (histogram in yellow on the same panel). This high positive correlation is partially due to differences among languages in terms of basic properties such as the average number of words per verse for instance and can thus hide a more complex reality within each language's dataset (a phenomenon known as Simpson's paradox, Blyth 1972). A visual inspection nevertheless reveals that a within-language positive correlation between WID_PP and WID_NP is present in 40/46 languages while a majority of their randomized counterparts logically exhibit a negative or no correlation at all (see Supplementary Information § 6.3.3 and § 6.3.4). In other words, for most languages, taking the semantic information approximated by a language model into account brings substantial information on the WID that would be calculated from the parallel version, both ruling out a pure "Simpsonian" correlation and supporting the initial assumption that surprisals calculated on English versions are informative on the semantic information present in their translated counterparts.

7. General discussion

In this paper, we first presented our main objective consisting in evaluating to which extent language comparisons based on raw complexity indices are reliable and robust against corpus content variation. Once this methodological goal set, we briefly presented several indices that have been used as yardsticks for measuring complexity in a typological perspective. It led us to select several indices computable from raw texts and that globally encompass morphological complexity rather than just inflectional morphology (Section 2).

In Section 4, we started by projecting the languages from our corpus (introduced in Section 3) into a bidimensional space defined by two complexity indices derived from WALs and AUTOTYP (GMC_W and GMC_A respectively). The moderate positive correlation present between these grammar-based indices calls for caution in their interpretation, since they only provide partial views on the overall morphological complexity. They nevertheless offer a rich and insightful background for systematically assessing the sensitivity of four indices of morphological complexity (TTR, *H*, MTL_D, and WID) against variation in the dataset

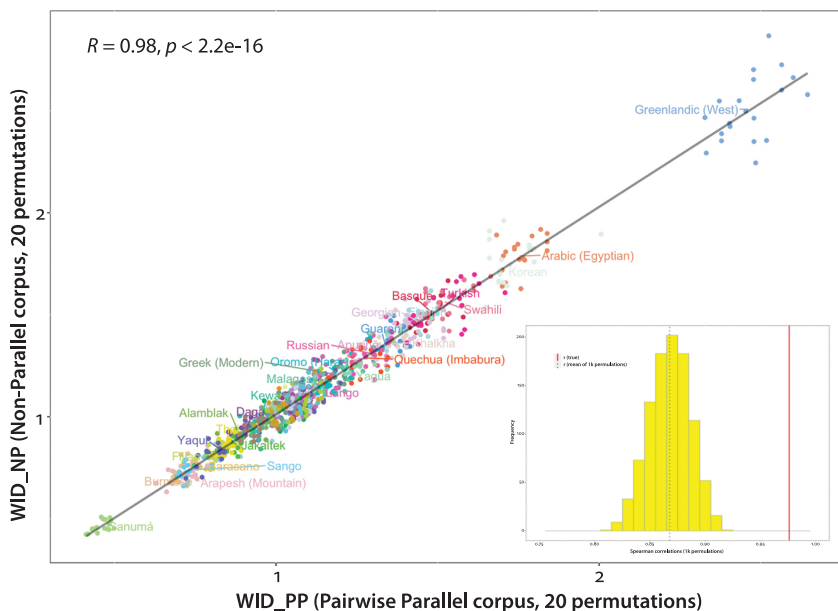


Figure 14. Large panel: Comparison between information densities estimated from Non-Parallel and Pairwise-Parallel configurations (20 subsets per language). Small panel: Comparison of the observed global correlation on the large panel (red vertical line) and the distribution of correlations obtained from 1000 randomized permutations (histogram in yellow)

on which they are estimated. Results show that while the indices positively correlate, partially encoding similar information, MTLD and WID are much more consistent in language complexity ranking than TTR or H across different dataset sizes and contents (results summarized in Figure 9). As a consequence, we insist on the fact that the interpretation of crosslinguistic differences in TTR and H should be cautious, and instead, we would suggest adopting MTLD or WID as more robust comparison metrics. Since this **stress test** was only performed on variable subsampled subsets of the Parallel Bible Corpus, we consider that replicating this observation on more diverse corpora will be a welcome improvement in the future. MTLD (introduced in McCarthy & Jarvis 2010) is not conceptually different from TTR and H since it is derived from the type and token frequencies estimated in the dataset.¹⁰ In contrast, WID (introduced in Pellegrino et al. 2011)

10. MTLD is also conceptually similar to another variation of TTR designed to deal with disparities across corpora, namely MATTR (Covington & McFall 2010), briefly mentioned in Section 2. However, MATTR is based on a sliding analysis window whose length should be manually set, with a recommended minimum length of a few hundred words, which can be

relies on a normalization with regard to a reference language (set to English in our study), and thus falls into the category of translation-based indices. It is robust, even on small subsets, and natively interpretable in a crosslinguistic framework, being comparative by design.

In Section 5, we confirmed that WID is meaningful as a morphological complexity index by showing that it is engaged in a multilingual trade-off with the pseudo-syntactic complexity (IWI) assessed by a randomized compression method (Figure 10). This result replicates previous findings (e.g. Koplenig et al. 2017, but see also Gutierrez-Vasques & Mijangos 2020 for a different perspective) and extends them: because WID and IWI are both unitless and normalized with regard to English, we can define Language EXplicitness (LEX) as their product, giving an insightful perspective on how different languages behave regarding their explicit vs. implicit information encoding. Another important aspect in our opinion is that these results are obtained on a corpus of 47 languages, which is by no means impressive in terms of coverage of the world's languages but still offers a diversity sufficient to give a gist of the shape and range of the complexity distribution one can expect. We thus advocate that, beside large-scale studies – especially useful for assessing statistical significance and defining the mathematical relationship binding several variables – there is still room for smaller scale studies as long as enough geographical and typological diversity is taken into consideration. In such a study for instance, it is quite easy to see how a specific language behaves and to take advantage of linguistic expertise, either encoded in WALS and AUTOTYP frameworks for instance, or directly by visually inspecting the language distributions.

In Section 6, we explored a quite different direction by bridging typological motivations and deep learning language models. We proposed a proof of concept which shows that, under certain circumstances, one can depart from strictly parallel corpora and still assess word information densities in a robust enough way that permits cross-linguistic comparisons, thanks to recent progress in natural language processing. The implemented approach nevertheless suffers from several limitations.

First, by subsampling our corpus, we introduced differences in the texts used for each language, but the general style and topic remain similar (excerpts of the King James Version of the bible). It means that the intrinsic coherence existing across the verses in terms of narration and vocabulary is partially preserved, leading to a high comparability of the semantic content across the datasets. There is no

problematic for the very small subsets we used here. In addition, Wu et al. (2019) showed that MTLT is more sensitive than MATTR in assessing meaningful differences in the context of language acquisition.

doubt that selecting more diverse language-specific datasets would increase variation in the informational content. Future studies will show whether the approach proposed here is viable in this context, but several encouraging observations can be drawn: (i) WID is based on densities defined as the ratio of a text informational content divided by its length. Adopting a diversified dataset will influence both the numerator and the denominator, and one can expect that language-specific regularities will preserve some invariance in WID; (ii) in the present study, each subset is very small (24 verses per language, because we needed distinct subsets for the 47 languages). Larger and more diverse corpora (both across-languages and within-language) will likely yield more robust and less text-specific estimations. (iii) Semantic embeddings (Mikolov et al. 2013) are methods developed to represent texts in a vector space sensitive to syntactic and semantic relationships. Initially developed in a monolingual framework, they can now provide a multilingual embedding space shared across an increasing number of languages (Artetxe & Schwenk 2019), paving the way to more accurate estimation of components of WID mostly influenced by the language itself or more idiosyncratic aspects (topic, style, register).

Secondly, this approach requires an estimation of the informational content of texts for each of the languages considered. Using the English version of the texts as a workaround is a trick in the sense that it considers this problem solved, by projecting all languages in the English language semantic and informational space. While not perfect, we argue that this kind of approach, somehow akin to the Wizard of Oz paradigm used in human-computer interfaces research, is a necessary step while the development of multilingual language models is still in progress. Even if reasonably optimistic on the future improvement of multilingual language models such as XLM (Conneau et al. 2019), available for 100 languages, we still consider their use for typological purposes problematic as long as they do not guarantee that their language-specific representations are directly comparable. In a sense, Gerz and her colleagues (Gerz et al. 2018) rightly advocate for taking typological considerations into account for developing “next-level language-agnostic [Language Model] Architecture”, coining an oxymoron that is still to be solved (see also Lake & Murphy 2021 and Rust et al. 2020 for a broader perspective).

Such language models have nevertheless undergone tremendous improvements since the introduction of the Transformer architecture (Vaswani et al. 2017, see also Wolf et al. 2020 for an introduction) and they can already help linguists to break the parallel corpus barrier. Additionally, adopting a complexity standpoint provides a promising avenue to mediate the dialogue between the linguistic typology and the NLP communities in a multilingual perspective. Indeed, several recent papers have investigated the intertwined relationship between the perfor-

mances reached by modern language model architectures and languages' typological features (see Gerz et al. 2018; Ponti et al. 2019, *inter alia*) and advocated for this much necessary cross-fertilization (see Choenni & Shutova 2020 and Gutierrez-Vasques et al. 2021 for fruitful examples).

8. Conclusions

In this study, we demonstrate a proof of concept suggesting that recent language models can improve the comparability across language corpora, both parallel and, as we believe, non-parallel ones. We consider that the methodologies introduced in this paper are an important and necessary step for the extension of quantitative complexity studies towards more natural and more diverse corpora than the ones traditionally used, often limited to religious or administrative texts. Obviously, even considering more diversified textual sources is not enough to get a comprehensive grasp on language diversity. As mentioned in the introduction (and advocated in Haig et al. 2021), there is an urgent need to study multilingual speech corpora recorded in natural interactions or narration if we want to put the notion of language complexity in its right place: as an indicator of a complex adaptive system at work where several dynamic mechanisms “*mesh* in the current moment” (MacWhinney 2005: 191, *italics in the original*). In this respect, ongoing initiatives such as the Multi-CAST (Haig & Schnell 2022), DoReCo (Paschen et al. 2020), and SCOPIC (Barth & Evans 2017) projects are important to offer a rich and diversified material for studying natural language and the reciprocal influences at play between grammar and real-time temporal constraints of speech, such as the ones recently demonstrated (e.g. Cohen Priva 2017; Coupé et al. 2019; Meister et al. 2021; Pimentel et al. 2021).

Funding




Yoon Mi Oh was funded by Ajou University (S-2019-G0001-00088).





This article was made Open Access under a CC BY 4.0 license through payment of an APC by or on behalf of the authors.

Acknowledgements

We thank the anonymous reviewers and the editor for helpful suggestions and comments. We also thank Christian Bentz and Ximena Gutierrez-Vasquez for providing us with the data used in their study.

References



-  Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89(3). 429–464.
-  Aranovich, Raúl. 2013. Transitivity and polysynthesis in Fijian. *Language* 89(3). 465–500.
-  Arkadiev, Peter & Francesco Gardani (eds.). 2020. *Introduction: The complexities of morphology*. Oxford: Oxford University Press.
-  Artetxe, Mikel & Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* 7. 597–610.
-  Baerman, Matthew, Dunstan Brown & Greville G. Corbett (eds.). 2015. *Understanding and measuring morphological complexity*. Oxford: Oxford University Press.
- Barth, Danielle & Nicolas Evans (eds.). 2017. *The Social Cognition Parallax Corpus (SCOPIC)* (Language documentation and conservation special publication no. 12). Honolulu: University of Hawai'i Press.
-  Bentz, Christian & Dimitrios Alikaniotis. 2016. The word entropy of natural languages. *arXiv preprint arXiv:1606.06996*. Available at: (last access 2 December 2022).
- Bentz, Christian, Tatyana Ruzsics, Alexander Koplenig & Tanja Samardžić. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Thomas François & Philippe Blache (eds.), *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 142–153. Osaka, Japan, December, 2016. As a part of COLING 2016. 26th International Conference on Computational Linguistics.
-  Bentz, Christian, Ximena Gutierrez-Vasques, Olga Sozinova & Tanja Samardžić. 2022. Complexity trade-Offs and equi-complexity in natural languages: A meta-analysis. *Linguistics Vanguard*.
- Bickel, Balthasar & Johanna Nichols. 2013. Chapter 22: Inflectional synthesis of the verb, In Matthew S. Dryer & Martin Haspelmath (eds.). *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at: <http://wals.info/chapter/22> (last access 2 December 2022).
- Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga & John B. Lowe. 2022. *The AUTOTYP database (v1.0.0) [Data set]*. Zenodo. Available at: <https://zenodo.org/record/5931509#Y24TeuxBxb8> (last access 2 December 2022).
-  Bisang, Walter. 2014. Overt and hidden complexity—Two types of complexity and their implications. *Poznan Studies in Contemporary Linguistics* 50(2). 127–143.
-  Bisang, Walter. 2015. Hidden complexity—the neglected side of complexity and its implications. *Linguistics Vanguard* 1(1). 177–187.
-  Blyth, Colin R. 1972. On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* 67(338). 364–366.
-  Choenni, Rochelle & Ekaterina Shutova. 2020. What does it mean to be language-agnostic? Probing multilingual sentence encoders for typological properties. *arXiv e-prints arXiv:2009.12862*. Available at: (last access 2 December 2022).

-  Christodouloupoulos, Christos & Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation* 49(2). 375–395.
-  Cohen Priva, Uriel. 2017. Not so fast: Fast speech correlates with lower lexical and structural information. *Cognition* 160. 27–34.
-  Çöltekin, Çağrı & Taraka Rama. 2022. What do complexity measures measure? Correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard*.
-  Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*. Available at: (last access 2 December 2022).
-  Cotterell, Ryan, Christo Kirov, Mans Hulden & Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics* 7. 327–342.
-  Coupé, Christophe, Oh Yoon Mi, Dan Dediu & François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances* 5(9). eaaw2594.
-  Covington, Michael A. & Joe D. McFall. 2010. Cutting the Gordian knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics* 17(2). 94–100.
-  Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.
-  de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2). 255–308.
- Derbyshire, Desmond C. & Doris L. Payne. 1990. Noun classification systems of Amazonian languages. In Doris L. Payne (ed.), *Amazonian linguistics: Studies in lowland South American languages*, 243–272. Austin: University of Texas Press.
- Dixon, Robert M. W. 1988. *A grammar of Boumaa Fijian*. Chicago: University of Chicago Press.
- Dryer, Matthew S. & Martin Haspelmath. 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
-  Easterday, Shelece, Matthew Stave, Marc Allasonnière-Tang & Frank Seifart. 2021. Syllable complexity and morphological synthesis: a well-motivated positive complexity correlation across subdomains. *Frontiers in Psychology* 12. 583. Available at: (last access 5 December 2022).
-  Ehret, Katharina & Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Raffaella Baechler & Guido Seiler (eds.), *Complexity, isolation, and variation*, 71–94. Berlin: De Gruyter Mouton.
-  Ehret, Katharina, Alice Blumenthal-Dramé, Christian Bentz & Aleksandrs Berdicevskis. 2021. Meaning and measures: Interpreting and evaluating complexity metrics. *Frontiers in Communication* 6. 640510. Available at: (last access 5 December 2022).
-  Erdmann, Alexander, Salam Khalifa, Mai Oudah, Nizar Habash & Houda Bouamor. 2019. A little linguistics goes a long way: Unsupervised segmentation with limited language specific guidance. In Garrett Nicolai & Ryan Cotterell (eds.), *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 113–124. Florence: Association for Computational Linguistics.

-  Frank, Stefan L. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science* 5(3). 475–494.
-  Gerz, Daniela, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart & Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In Ellen Riloff, David Chiang, Julia Hockenmaier & Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 316–327, Brussels: Association for Computational Linguistics.
-  Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68(1). 1–76.
-  Givón, Talmy. 2009. *The genesis of syntactic complexity: Diachrony, ontogeny, neuro-cognition, evolution*. Amsterdam: John Benjamins.
-  Greenberg, Joseph H. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics* 26(3). 178–194.
-  Gutierrez-Vasques, Ximena & Victor Mijangos. 2020. Productivity and predictability for measuring morphological complexity. *Entropy* 22(1). 48.
-  Gutierrez-Vasques, Ximena, Christian Bentz, Olga Sozinova & Tanja Samardzic. 2021. From characters to words: The turning point of BPE merges. In Paola Merlo, Jorg Tiedemann & Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3454–3468. Association for Computational Linguistics.
- Haig, Geoffrey & Stefan Schnell (eds.). 2022. *Multi-CAST: Multilingual corpus of annotated spoken texts*. Version 2108. Available at: <https://multicast.aspra.uni-bamberg.de> (last access 2 December 2022).
- Haig, Geoffrey, Stefan Schnell & Frank Seifart (eds.). 2021. *Doing corpus-based typology with spoken language corpora: State of the art*. Honolulu: University of Hawai'i Press.
-  Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
-  Hollenstein, Nora, Federico Pirovano, Ce Zhang, Lena Jäger & Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cottrell, Tanmoy Chakraborty & Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 106–123. Association for Computational Linguistics.
-  Hollenstein, Nora, Emmanuele Chersoni, Cassandra Jacobs, Yohei Oseki, Laurent Prévot & Enrico Santus. 2022. CMCL 2022 Shared Task on Multilingual and Crosslingual Prediction of Human Reading Behavior. In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot & Enrico Santus (eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 121–129. Dublin: Association for Computational Linguistics.
-  Johnson, Wendell. 1944. Studies in language behavior: I. A program of research. *Psychological Monographs* 56(2). 1–15.
-  Joseph, John E. & Frederick J. Newmeyer. 2012. 'All Languages Are Equally Complex': The rise and fall of a consensus. *Historiographia Linguistica* 39(2–3). 341–368.

-  Juola, Patrick. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* 5(3). 206–213.
-  Kettunen, Kimmo, Markus Sadeniemi, Tiina Lindh-Knuuttila & Timo Honkela. 2006. Analysis of EU languages through text compression. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo & Tapio Pahikkala (eds.), *International Conference on Natural Language Processing* (in Finland), 99–109. Berlin: Springer.
-  Kettunen, Kimmo. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics* 21(3). 223–245.
-  Koplenig, Alexander, Peter Meyer, Sascha Wolfer & Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLoS ONE* 12(3). e0173614.
-  Koplenig, Alexander. 2019. Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society Open Science* 6(2). 181274. Available at: (last access 5 December 2022).
-  Kortmann, Bernd & Benedikt Szmrecsanyi. 2012. *Linguistic complexity: Second language acquisition, indigenization, contact*. Berlin: De Gruyter Mouton.
- Kusters, Wouter. 2003. *Linguistic complexity: The influence of social change on verbal inflection*. Utrecht: Netherlands Graduate School of Linguistics.
-  Lake, Brenden M. & Gregory L. Murphy. 2021. Word meaning in minds and machines. *arXiv preprint ArXiv:2008.01766*. Available at: (last access 2 December 2022).
-  Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS ONE* 5(1). e8559.
-  MacWhinney, Brian. 2005. The emergence of linguistic form in time. *Connection Sciences* 17(3–4). 191–211.
-  Maddieson, Ian. 2009. Calculating phonological complexity. In François Pellegrino, Egidio Marsico, Ioana Chitoran & Christophe Coupé (eds.), *Approaches to Phonological Complexity*, 83–110. Berlin: De Gruyter Mouton.
-  Maddieson, Ian, Sébastien Flavie, Egidio Marsico, Christophe Coupé & François Pellegrino. 2013. LAPSyd: Lyon-Albuquerque phonological systems database. In Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Lori Lamel, François Pellegrino & Pascal Perrier (eds.), *Proceedings of the 14th Interspeech Conference, Lyon, France, 3022–3026* Lyon: International Speech Communication Association (ISCA).
-  Malouf, Robert. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology* 27. 431–458.
- Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel bible corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 3158–3163. Reykjavik: European Language Resources Association (ELRA).

- McCarthy, Arya D. Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden & David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, 3922–3931. Marseille: European Language Resources Association (ELRA).
-  McCarthy, Philip M. & Scott Jarvis. 2010. MTLT, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42. 381–392.
-  Meister, Clara, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell & Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia & Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 963–980. Punta Cana: Association for Computational Linguistics.
-  Merckx, Danny & Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot & Enrico Santus (eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 12–22. Association for Computational Linguistics.
-  Mielke, Sabrina J., Ryan Cotterell, Kyle Gorman, Brian Roark & Jason Eisner. 2019. What kind of language is hard to language-model? In Anna Korhonen, David Traum & Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4975–4989. Florence: Association for Computational Linguistics.
-  Miestamo, Matti, Kaius Sinnemäki & Fred Karlsson (eds.). 2008. *Language complexity: Typology, contact, change* (Studies in Language Companion Series 94). Amsterdam: John Benjamins.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26. 3111–3119.
- Moscato del Prado, Fermin. 2011. The mirage of morphological complexity. In Laura Carlson, Christoph Hoelscher & Thomas F. Shipley (eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 3524–3529. Austin: Cognitive Science Society.
-  Mufwene, Salikoko S., Christophe Coupé & François Pellegrino. 2017. *Complexity in language: Developmental and evolutionary perspectives*. Cambridge: Cambridge University Press.
-  Nercesian, Verónica. 2014. Wordhood and the interplay of linguistic levels in synthetic languages. An empirical study on Wichi (Mataguayan, Gran Chaco). *Morphology* 24. 177–198.
- Newman, Paul. 2003. Hausa and the Chadic languages. In Bernard Comrie (ed.) *The major languages of South Asia, the Middle East and Africa*, 177–192. London: Routledge.


- Nichols, Johanna & Christian Bentz. 2019. Morphological complexity of languages reflects the settlement history of the Americas. In Katerina Harvati, Gerhard Jäger & Hugo Reyes-Centeno (eds.), *New perspectives on the peopling of the Americas*, 13–26. Tübingen: Kerns Verlag.
- Oh, Yoon Mi. 2015. Linguistic complexity and information: Quantitative approaches. Lyon: University of Lyon Ph.D. dissertation.
-  Oh, Yoon Mi, Christophe Coupé, Egidio Marsico & François Pellegrino. 2015. Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of Phonetics* 53. 153–176.
- Paschen, Ludger, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave & Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2657–2666. Marseille: European Language Resources Association.
-  Pellegrino, François, Christophe Coupé & Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language* 87(3). 539–558.
-  Pimentel, Tiago, Brian Roark & Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics* 8. 1–18.
-  Pimentel, Tiago, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi & Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world’s languages. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia & Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 949–962. Punta Cana: Association for Computational Linguistics.
-  Ponti, Edoardo Maria, Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova & Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics* 45(3). 559–601.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8). 9.
-  Rust, Phillip, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder & Iryna Gurevych. 2020. How good is your tokenizer? On the monolingual performance of multilingual language models. *arXiv preprint arXiv:2012.15613*. Available at: (last access 2 December 2022).
-  Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum & Evelina Fedorenko. 2021. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *Proceedings of the National Academy of Sciences* 118(45). e2105646118.
-  Shosted, Ryan K. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10(1). 1–40.
-  Sinnemäki, Kaius & Di Garbo, Francesca. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology* 9. 1141.

- Thomason, Sarah Grey & Terrence Kaufman. 1992. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.
- Thornell, Christina. 1997. *The Sango language and its lexicon (Sêndâ-yângâ tî sângö)*. Vol. 32. Lund: Lund University.
- Trudgill, Peter. 2001. Contact and simplification: Historical baggage and directionality in linguistic change. *Linguistic Typology* 5(2). 371–374.
- Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30. 6000–6010.
-  Vera, Javier & Wenceslao Palma. 2020. Laplacian spectrum approach to linguistic complexity: A case study on indigenous languages of the Americas. *Europhysics Letters* 129(5). 58003.
- von Prince, Kilu & Vera Demberg. 2018. POS tag perplexity as a measure of syntactic complexity. In Alekandrs Berdicevskis & Christian Bentz (eds.), *Proceedings of the First Shared Task on Measuring Language Complexity*, 20–25. Uppsala: Uppsala University, Department of Linguistics and Philology.
-  Wedel, Andrew, Abby Kaplan & Scott Jackson. 2013. High functional load inhibits phonological contrast loss: A corpus study. *Cognition* 128(2). 179–186.
- Wilcox, Ethan Gotlieb, Jon Gauthier, Jennifer Hu, Peng Qian & Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In Stephanie Denison, Michael Mack, Yang Xu & Blair C. Armstrong (eds.), *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 1707–1713. Cognitive Science Society.
-  Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Qun Liu & David Schlangen (eds.) *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, 38–45. Association for Computational Linguistics.
-  Wray, Alison & George W. Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117(3). 543–578.
-  Wu, Shang-Yu, Rei-Jane Huang & I-Fang Tsai. 2019. The applicability of D, MTLD, and MATTR in Mandarin-speaking children. *Journal of Communication Disorders* 77. 71–79.

Address for correspondence

Yoon Mi Oh
Department of French Language and Literature
Ajou University
16499 Suwon Gyeonggi-do
South Korea
yoonmih@ajou.ac.kr
 <https://orcid.org/0000-0003-1164-6141>

Co-author information

François Pellegrino
Laboratoire Dynamique du Langage
UMR5596, CNRS
University of Lyon
Lyon, France
francois.pellegrino@univ-lyon2.fr
 <https://orcid.org/0000-0002-6456-1953>

Publication history

Date received: 30 May 2022
Date accepted: 10 November 2022
Published online: 20 December 2022