A Measure for Transparent Comparison of Linguistic Diversity in Multilingual NLP Data Sets

Anonymous ACL submission

Abstract

Typologically diverse benchmarks are increasingly created to track the progress achieved in multilingual NLP. Linguistic diversity of these data sets is typically measured as the number of languages or language families included in the sample, but such measures do not consider structural properties of the included languages. In this paper, we propose assessing linguistic diversity of a data set against a reference language sample as a means of maximising linguistic diversity in the long run. We represent languages as sets of features and apply a version of the Jaccard index (J_{mm}) suitable for comparing sets of measures. In addition to the features extracted from typological data bases, we propose an automatic text-based measure, which can be used as a means of overcoming the well-known problem of data sparsity in manually collected features. Our diversity score is interpretable in terms of linguistic features and can identify the types of languages that are not represented in a data set. Using our method, we analyse a range of popular multilingual data sets (UD, Bible100, mBERT, XTREME, XGLUE, XNLI, XCOPA, TyDiQA, XQuAD). In addition to ranking these data sets, we find, for example, that (poly)synthetic languages are missing in almost all of them.

1 Introduction

Data sets for training and testing NLP models are increasingly multilingual and aimed at broad linguistic coverage. These data sets are often claimed to represent a typologically diverse sample, including low-resource and endangered languages.

Linguistic diversity is typically described as the number of languages included in the data set, yet less often as the number of language families to which these languages belong. Both counts indicate a level of linguistic diversity: the more languages and families, the more diversity. But how do we know that included languages are indeed different? How can we define a desired or optimal diversity to set as a goal when composing multilingual data sets? These questions need to be addressed if our goal is to know how NLP technology generalises across diverse languages, without testing it on each single language (even if we had the necessary data for all languages).

The aim of this paper is to initiate and facilitate comparisons between multilingual NLP data sets with respect to a linguistic diversity reference. For this, we propose a measure of linguistic diversity and a method of comparison that identifies what kinds of linguistic features are missing. As an initial reference, we rely on a predefined sample of languages — the 100-language-sample (100L) selected by the Word Atlas of Language Structures (WALS; Comrie et al. (2013)) to represent geographic and phylogenetic diversity. As a comparison method, we formulate a version of the Jaccard index suitable for comparing measures. This measure allows us to quantify the distance between the observed and the reference diversity in terms of linguistic features, showing not only how diverse language samples are but also what kinds of linguistic phenomena are not represented in a given sample. To facilitate automatic extraction of linguistic features needed for assessing linguistic diversity, we complement the information from linguistic data bases with relevant text statistics.

Our proposals are intended to help researchers make informed choices when designing a multilingual data set. Representing a wider spectrum of linguistic diversity is not only a way to improve the cross-linguistic generalisation of NLP technology, but also a way to deal with biases against low-resource languages, which are harder to represent and thus more likely to be left behind (Joshi et al., 2020).

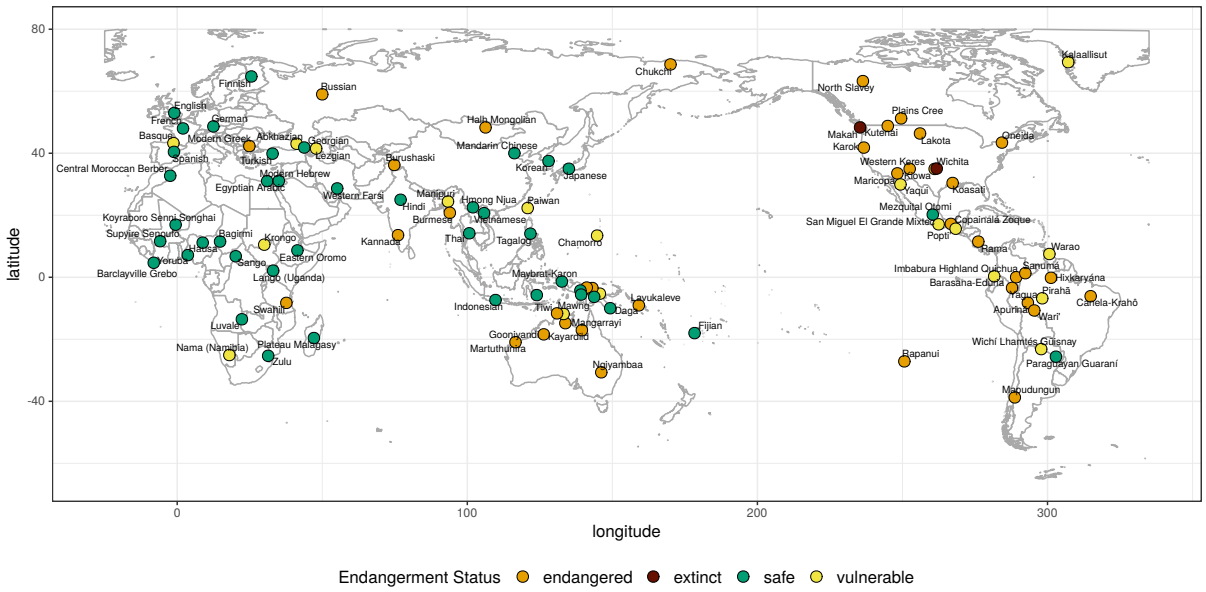

Figure 1: Geographic distribution of the languages included in the WALS 100L sample and their endangerment status.

2 Background and Related Work

Evaluating the linguistic diversity of data sets relies on comparable descriptions of languages. For instance, the (approximate) number of speakers is an attribute whose value can be found and compared for all registered languages. This attribute, however, does not describe the structure of languages. An example of a structural attribute would be the presence or the absence of adjectives in a language. To establish the value of this attribute for any language, we need a universal definition of what an adjective is. It turns out that such universal definitions are hard to formulate in a principled way (Haspelmath, 2007), which makes it hard to define objective measures of how similar or dissimilar any two languages are.

The most widely accepted method for comparing languages relies on genealogical classification: given a phylogenetic tree, we consider languages located in the same region of the tree to be similar. This method currently prevails in NLP (cf. the work discussed in Section 6). Typically, we regard languages that belong to the same *family* to be similar. To know which language belongs to which family, we turn to popular authorities such as WALS (Dryer and Haspelmath, 2013) or Glottolog (Hammarström et al., 2018). However, language families can be too broad for a meaningful comparison as they include typologically very different languages. For instance, English and Armenian belong to the same family, Indo-European, but are vastly different in terms of their phoneme inventories, morphology, and word order.

Another possibility to compare languages, starting to be used in NLP only recently, is to rely on grammatical features available in the WALS data base, which is a comprehensive source of information about linguistic structures despite being sparsely populated with features that are often known for only a few languages.[1] Ponti et al. (2020) propose a diversity score using the features from URIEL (Littell et al., 2017) (which is derived from WALS and other typological data bases). The score is called *typology index* and it is calculated as the entropy of feature values (averaged per data set).[2] In other NLP work, grammatical features (usually termed *typological*) are used for other purposes, such as predicting the features (Ponti et al., 2019) rather than using them for language sampling in creating multilingual data sets. Moran (2016) use WALS and AUTOTYP features (Stoll and Bickel, 2013) to compose a sample of 10 maximally diverse languages for a corpus-based study of language acquisition.

Finally, languages can be described using features derived from various text statistics, which is still not used as a method of sampling. Type-

[1] An alternative typological data base is AUTOTYP (Bickel et al., 2017), with a different design but similar coverage.

[2] They propose two more scores, *family* and *geography*, which do not make use of grammatical features.

token ratio (TTR) or unigram entropy of a text have been shown to correlate with grammar-based morphological complexity measures (Kettunen, 2014; Bentz et al., 2016). Many other methods have been proposed for assessing linguistic complexity using text statistics (see, for instance, Berdicevskis et al. (2018)). All of these measures can, in principle, be used for describing and comparing languages. Although such comparisons might seem counter-intuitive and hard to interpret in terms of genealogical classification, it is worth exploring them as complementary descriptions of languages, more directly relevant to text processing, which is the most common goal in NLP.

Transfer learning created a new need for nuanced languages comparison for NLP. While models can now be transferred across languages with zero-shot or few-shot learning (Pires et al., 2019), the success of the transfer depends on the similarity between languages. Lin et al. (2019) propose a range of measures that can be used in order to choose the best transfer language, which they divide into data-dependent (data size, token overlap, TTR) and data independent (various distance measures extracted from the URIEL data base). Lauscher et al. (2020) study how well different similarity scores predict the success of the transfer and they find that language family is, in fact, the one that is least helpful in all the tasks considered (with mBERT and XLM-R). Various criteria for assessing language similarity remain an open research area in NLP (Turc et al., 2021; Pelloni et al., 2022; Samardžić et al., 2022; de Vries et al., 2022). Our proposal for assessing linguist diversity is relevant to these efforts too, as its key component is language comparison at the level of features extracted from both typological data bases and text samples.

More generally, our work is intended to contribute to several wide-scope initiatives for improving the quality of data management in multilingual NLP (Bender and Friedman, 2018; Kreutzer et al., 2021; Lhoest et al., 2021) by focusing specifically on diversity assessments and data-independent scores for language comparison.

3 Comparing Data Sets with Jaccard Similarity

Our goal is to estimate the linguistic diversity of a data set with respect to some reference. Our score is thus a comparison between two data sets. More precisely, we compare scaled distributions

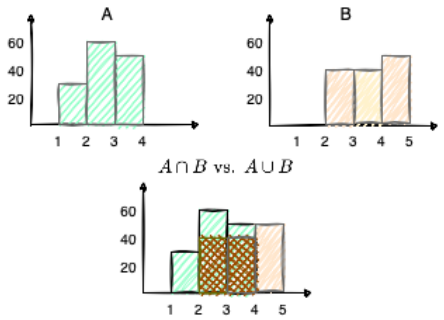

Figure 2: A toy example of comparing sets of measures with the minmax version of the Jaccard index.

of the values of a numerical attribute as shown in Figure 2. The upper part of the figure shows (constructed) examples of two data sets (A and B), which we compare assuming that A is the data set whose diversity we want to assess and B is the reference. The values of the numerical attribute (one measurement per language) are on the x-axis and the numbers of languages are on the y-axis. Each bar in the figures represents the number of languages in the given data set with the numerical value in the given range (bin). For instance, the first bar in the upper left plot shows that the first sample (A) has 30 languages, with the values of their numerical attributes falling between 1 and 2. The other sample (B) has no languages in this bin.

The width of the bins is arbitrary, but it does impact the score. Narrower bins capture more differences between two distributions than wider bins. By setting the width of the bins, we thus control the resolution at which we want to compare two data sets. In our example, the width is the distance between integers, but one can define different thresholds (as long as all of the bins are of the same width).

Since the data sets that we compare contain different numbers of languages, the values on the y-axis (counts of languages) are normalised in order to neutralise the effect of the size of the samples and focus rather on the diversity. We multiply all counts in the smaller set with the scalar c:

$$c = \frac{max(|A|, |B|)}{min(|A|, |B|)} \quad (1)$$

In this way, we increase the counts in the smaller set proportionally to obtain the same number of data points in both distributions and comparable numbers in each bin.[3]

[3]Another way to normalise the counts would be to divide them by the size of the set, but we chose the first option in

Once we have represented our two sets in this way, we compare them using a generalised version of Jaccard similarity. This score shows how much the two distributions overlap. Intuitively, it is the ratio between the intersection and the union of the two distributions (shown in the bottom part of Figure 2).

The original Jaccard index (Jaccard, 1912) compares two sets, but its generalised versions are suitable for comparing sets of measurements. Thus, we use the *minmax* version of the score (J_{mm}), initially proposed by Tanimoto (1958) for comparing vectors of binary values and then generalised to weight vectors by Grefenstette (1994). In our version, we compare two data sets as two vectors of weights: each bin is one dimension in the vectors and the number of languages in that bin is its weight.

Formally, we first map all the languages in all data sets to real numbers $m : \mathbb{L} \mapsto \mathbb{R}$, so that $\{Y = m(x) : x \in X\} = \{(x_i, y_i)\}$, where x is a language in a data set, y is its corresponding measurement ($y \in \mathbb{R}$) and the range of the index i $1 \ldots |X|$ is the set of languages included in a data set. We then group the measurements into bins by applying a given threshold: $\{Z = t(y) : y \in Y\} = \{(y_i, z_j)\}$, where z is the bin to which the measurement is assigned, the range of i is $1 \ldots |X|$ and the range of j is $1 \ldots |Z|$.

With this formalisation, we define the Jaccard minmax similarity of two data sets, $J_{mm}(A, B)$, as a similarity between two vectors of weights:

$$J_{mm}(\mathbf{a}, \mathbf{b}) = \frac{\sum_{j=1}^{|Z|} min(a_j, b_j)}{\sum_{j=1}^{|Z|} max(a_j, b_j)} \qquad (2)$$

The sum in the numerator represents the intersection and the sum in the denominator the union of the two sets of measurements. The weights a and b represent the number of measurements in the bin j.

The values of J_{mm} fall in the range $[0, 1]$, with higher values indicating more similarity between A and B, and, indirectly, better coverage of linguistic diversity in A.

What is especially interesting about using J_{mm} as a diversity score is its transparency in terms of individual measurements: we can visualise and interpret where exactly a data set departs from the reference.

order to preserve the notion of *number of languages*, which is helpful for the subsequent explanations.

4 Language Features

We now turn to the question of how to define and take measures (the values on the x-axis in Figure 2) that can be used for calculating Jaccard minmax similarity between sets of languages. We use two kinds of descriptions.

4.1 Grammar Features

Typological data bases are currently the principal source of information about the properties of languages, but NLP researchers are faced with many obstacles when using this information. The popular software package `lang2vec` associated with the URIEL data base (Littell et al., 2017) alleviates some of the obstacles. First, the package solves the problem of incompatible feature values across different sources by mapping the data from several original data bases to binary features. Second, the problem of sparsity of feature values is solved by imputing the missing values: instead of a missing feature value in a language, the package returns the observed value for the same feature in the closest language. In this way, features become available for all queried languages, which is necessary for estimating language diversity, but a large proportion (roughly 40%) of the returned features are imputed.

While `lang2vec` facilitates retrieving typological features, its use for describing languages is limited due to remaining obstacles that are hard to solve. First, it does not contain any morphological features, which are especially relevant to NLP due to the known difficulties with that morphologically rich languages (Tsarfaty et al., 2013). The second unsolved problem is the fact that typological features are hard to add for languages for which they are not already available. Adding any new feature or value requires human expertise in many different languages.

4.2 Text Features

As a solution to both of these problems, we use text statistics as language features. In this study, we focus on the *mean word length* as an approximation of aggregated morphological features, but other text-based features might be envisioned in future work. The intuition behind our proposal is that word length indicates morphological types: longer words can be expected in languages with rich morphology (large morphological paradigms, productive derivation), while shorter words are found in

languages with less morphology.[4] As empirical evidence of the expected relationship between the word length and morphology, we perform a correlation test between the mean word length and morphological complexity calculated over morphological features in WALS (Section 5).

Text features are especially interesting in the context of NLP because they can be calculated automatically and applied to any language in which there are any texts to process. An important advantage of word length over other text statistics in this regard is that it manifests itself in very small samples of text and remains stable across different sizes. A sample of contiguous text of only 500 tokens gives us already a very good estimation of the overall mean word length. This can be seen in Figure 4 in the Appendix A, which shows the values of the mean word length on random samples of the length 500, 2000 and 10000 tokens in 87 languages. A correlation score (also in the Appendix A) shows that languages are almost identically ranked with all the sample sizes.

4.3 Maximising Linguistic Diversity

The editors of the WALS data base have selected two samples of languages (100 and 200 sample) as a means of guidance in the collective effort to create linguistic descriptions on a wide scale. These samples maximise genealogical (language family) and areal (geographic) diversity. Completing their descriptions is expected to minimise a potential bias regarding the relative frequency of different types of linguistic features included in the data base (Comrie et al., 2013). Figure 1 shows the geographic distribution of the languages in the 100 sample and their endangerment status according to UNESCO.

Recently, text samples have been collected for most of the 100 languages in the TeDDi data set (Moran et al., 2022).[5] These text data are sampled from online resources, e.g., Project Gutenberg,[6] Open Subtitles (Lison and Tiedemann, 2016), The Parallel Bible Corpus (Mayer and Cysouw, 2014), the Universal Declaration of Human Rights,[7], but also from grammars and other language documentation sources. For languages not present in online resources, the texts were manually transcribed.

We take these two resources as the current reference that maximises linguistic diversity in terms of grammar features (WALS) and text features (TeDDi). We compare NLP data sets with these references, but our method can be applied to compare any given pair of data sets including potentially better references in the future.

5 Data and Methods

We calculate the Jaccard minmax diversity score (J_{mm}) for a number of popular multilingual data sets in comparison to the TeDDi sample.[8] Without attempting to provide an exhaustive evaluation, we review data sets that satisfy the following criteria: multilingual (containing ten or more languages), relatively widely used and recently released or updated. The list is given in Table 1 and discussed in more detail in Section 6. For reference, we compare our J_{mm} score to the typological index (TI) previously proposed as a linguistic diversity measure by Ponti et al. (2020) (see Section 2).

Descriptions of the data sets often do not include all the information that was needed for our comparison. In particular, the number of language families is often not stated. To add this information, we extracted language names from the data files, converted these names into ISO 639-3 codes manually, and then retrieved the corresponding families from the Glottolog data base (top level family). The numbers in the second and the third column marked with an asterisk are added or modified by us. The numbers without an asterisk are reported in the respective publications. Note that the conversion to ISO 639-3 codes led to some changes in the number of languages, compared to those cited in the data descriptions. For instance, the mBERT training data has only 97 distinct languages, not 104 as mentioned in the original description.

5.1 Methods for Text Features

We define words to be sequences of Unicode characters, delimited by spaces or other language-specific word delimiters, as defined by common multilingual tokenisers. We tokenise all the collected samples into word-level tokens using the Python library Polyglot (Al-Rfou, 2015).[9] If a resulting token does not contain any alphanumeric

[4] We give a more specific definition of the notion of a word as part of the methods in Section 5.

[5] https://github.com/MorphDiv/TeDDi_sample/tree/master

[6] https://www.gutenberg.org/

[7] http://unicode.org/udhr/

[8] In the final version, the link to the shared code for reproducing the calculations will be provided here.

[9] https://polyglot.readthedocs.io

characters, we discard it as punctuation. All the remaining tokens are further segmented into characters using the Python library `segments` (Moran and Cysouw, 2018). [10] We split words into sequences of characters and take their length as word length.[11] We apply this same definition to all scripts, but we discuss below potential adjustments in the case of (partially) logographic scripts.

Since the mean word length can be calculated on small samples, we take a single random sample for each language in a data set that we consider. To do this, we select a random position in the data set and extract contiguous text of the length up to 10K tokens starting from the random position. In case a data set does not contain such long texts (or sequences of paragraphs), we take smaller samples. The smallest samples are 200-300 tokens long.

As a result, we obtain a set of real numbers, each number representing a language in a data set. To turn these numbers into discrete features, we group them into bins of equal size. We set the bin width to 1.[12]

Mean word length vs. WALS morphology features Following Bentz et al. (2016), we calculate a complexity score (C_{WALS}) for each language using the set of 26 features that are relevant to describing morphology. This score is obtained by: 1) transforming the range of values each feature can take so that bigger values reflect the increasing use of morphology; 2) normalizing and averaging the resulting feature values per language. See Appendix B for more details. C_{WALS} ranges from 0 to 1, where values closer to one indicate that the language encodes more morphosyntactic distinctions, making its morphology richer. We observe a strong correlation ($\rho = 0.69$) between the mean word length and morphological complexity for 29 diverse languages (the subset of Teddi languages for which the 26 WALS features are known). The high correlation means that the variables quantify very similar phenomena and can be used interchangeably.

Adjustments for logographic scripts Words in languages with logographic scripts tend to be shorter due to the fact that a single symbol corresponds to several alphabetic symbols (Sproat and Gutkin, 2021). For instance, in Mandarin Chinese, types such as 的 *de* (possessive particle), 了 *le* (aspect particle), 是 *shì* (copular verb 'is'), 我們 *wǒmen* (pronoun 'us') are assigned lengths (1, 1, 1, 2) respectively when measured UTF-8 characters in the original script. When transliterated into Pinyin, the corresponding lengths are (2, 2, 3, 5). Hence, compared to Pinyin, the lengths are somewhat underestimated. It might seem more appropriate to convert the logographic scripts into their romanised counterparts to achieve cross-linguistic comparability. We opt for leaving such scripts without conversion, because we consider this phenomenon part of the diversity that we want to capture. Additional motivation for our choice is the fact that NLP systems have to deal with text as it is regardless of the mapping between written characters and sounds. To show that this decision does not impact our main findings, we report in Appendix C diversity scores with adjusted mean word length.

5.2 Linguistic Diversity Scores

With the grammar features extracted from URIEL, we calculate syntactic diversity according to both TI and J_{mm}.

Syntax Typological Index (TI_{syn}) Following the formulation by Ponti et al. (2020), we calculate the typological index for each data set. In this context, a language is characterized by 103 syntactic features with binary values[13]. For each feature, Shannon entropy is estimated using the distribution of feature values in a data set. The feature-specific entropy values are averaged over the full set of features to obtain a TI score ranging from 0 to 1. The TI values closer to 1 indicate a more diverse data set.

Syntax Jaccard (J_mm_syn) We apply Jaccard similarity for comparing each data set against the TeDDi sample. Here the measures are the counts of the observed values of the same 103 syntactic feature available in `lang2vec`. This means that the items on the x-axis in Figure 2 are the 103 values, while the y-axis represents the number of times each feature value was observed in a data set.

With text features (mean word length) extracted from TeDDI and the scored NLP data sets, we calculate morphological diversity according to both TI and J_{mm}.

[10] https://github.com/cldf/segments

[11] We use the units defined by the Unicode Standard as "user-perceived characters" (NFC).

[12] In addition to this, we also tried smaller bin sizes. We do not report the latter results, but the main trends did not change.

[13] We use the syntax_knn features available in lang2vec, which includes predicted values for those languages whose features are not available

Name and main references	N(L)	N(F)	TI_{syn}	J_{mm_syn}	TI_{morph}	J_{mm_morph}
Universal Dependencies (UD)	106*	20*	0.567	0.736	0.349	0.650
Bible 100	103*	30*	0.649	0.811	0.311	0.534
mBERT	97*	15*	0.559	0.710	0.323	0.603
XTREME	40	14	0.612	0.775	0.311	0.457
XGLUE	19	7*	0.517	0.674	0.307	0.504
XNLI	15	7*	0.557	0.711	0.339	0.598
XCOPA	11	11	0.586	0.737	0.361	0.608
TyDiQA	11	10	0.626	0.751	0.343	0.525
XQuAD	12*	6*	0.523	0.680	0.341	0.588
TeDDi	89	51	**0.706**	-	**0.369**	-

Table 1: Diversity of multilingual NLP data sets. N(L): the number of languages in the data set. N(F): the number of families to which the languages belong. TI: typology index Ponti et al. (2020). J_{mm}: Jaccard minmax similarity (this paper).

Morphology Typological index (TI_{morph}) We adapt the measure proposed by Ponti et al. (2020) to the text-based features (mean word length). Each bin of the mean word length values is a feature and the number of languages that fall in a given bin are the counts of feature values. In other words, the mean word length becomes a vector of binary values, 1 for the languages that are in the bin and 0 for all the other languages in the sample. The rest of the calculation is the same as in TI_{syn}.

Morphology Jaccard (J_mm_morph) Similarly to J_mm_syn, we calculate the Jaccard score by comparing the distributions of the mean word length: TeDDi vs. a given NLP data set.

6 Findings

Table 1 lists all the reviewed data sets with all the measures of linguistic diversity. The colour scale of the cells represents the relative ranking of data sets according to each measure separately. TeDDI data set obtains the highest diversity scores at both levels (syntax and morphology) using the TI measure. This confirms the role of these resources as the current reference regarding linguistic diversity.

TI and J_{mm} are consistent The rankings of data sets according to the J_{mm} score are very similar to those obtained with the TI score when the syntactic features are used. The agreement between the two measures is somewhat lower in the case of morphological features, but still rather high. The consistency between the two measures is not a trivial outcome given the entirely different approaches behind them. We can thus take this agreement as a validation of both measures. The main advantage of J_{mm} compared to TI is its transparency regarding the kinds of languages that are missing in a data set.

Diversity rankings of NLP data sets The highest rankings appear split between the two structural levels. Bible 100 (Christodouloupoulos and Steedman, 2015) and XTREME (Hu et al., 2020) are the two most syntactically diverse data sets, while their morphological diversity is moderate to low. The Bible data set contains mostly non Indo-European languages, while the collection criteria for the XTREME data set was to maximise diversity. On the other hand, Universal Dependencies (UD, Nivre et al. (2020), which are often seen as especially biased towards European languages, show the best morphological, but a moderate syntactic diversity. XCOPA (Ponti et al., 2020) and TyDiQA (Clark et al., 2020) are data sets containing relatively few languages, but designed to maximise linguistic diversity. They are both highly ranked on 3/4 measures (two syntactic and one morphological). Contrary to this, the linguistic diversity ranking of one of the most popular benchmarks that contain manual labels for several downstream tasks, XGLUE (Liang et al., 2020; Wang et al., 2019) is consistently low. XQuAD (Artetxe et al., 2020; Rajpurkar et al., 2016) fairs a little better, but it is still one of the least diverse data sets. The XNLI data set (Conneau et al., 2018; Bowman et al., 2015; Williams et al., 2018), which is compiled with the goal of spanning language families and which includes some low resource languages, remains of moderate linguistic diversity according to all measures. It is curious to see that the number of languages or even languages families included in a data set does not ensure a high linguistic diversity.

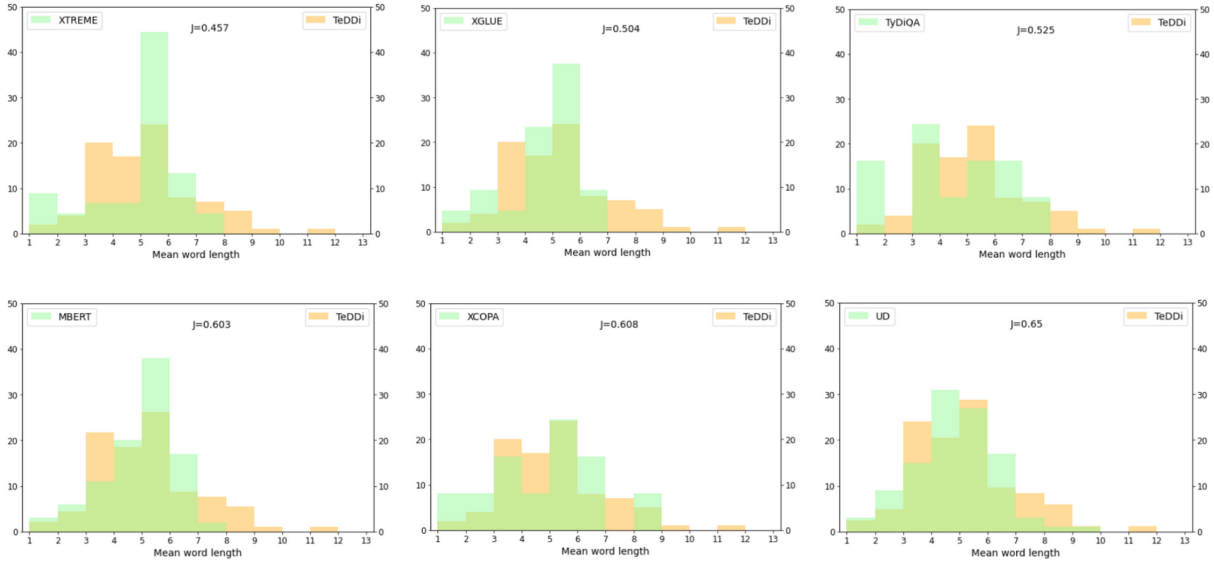

Figure 3: Union and intersection between the distributions of the mean word length in TeDDi and NLP data sets.

For example, the mBERT[14] data set contains 97 languages in 15 language families, but it turns out to be less diverse than smaller data sets such as XCOPA (on TI_{syn}, J_{mm_syn} and J_{mm_morph}) and TyDiQA (on TI_{syn}, J_{mm_syn} and TI_{morph}). The strategy of including the top 100 languages according to the size of their Wikipedia content (plus Thai and Mongolian), does not result in high diversity.

Underrepresented language types Figure 3 is a visualisation of the J_{mm_morph} score[15] for some of the data sets showing the overlap and differences with the reference (TeDDi). The recurrent difference is whether a data set includes languages with long words or not (mean length > 8). Those that contain at least some languages with long words (UD, XCOPA) score much better on J_{mm_morph} than those that remain completely on the short-middle side (EXTREME, XGLUE, TyDiQA, mBERT). The second important factor that leads to lower scores is a strong peak of the distribution indicating a bias towards one of the length bins (EXTREME, XGLUE, mBERT). The third factor is a different ("wrong") shape of the distribution (TyDiQA). The data set that diverges the most is EXTREME, exhibiting all three factors of disagreement. Overall, it seems that the right-hand side of the mean word length scale remains rather scarcely represented in all data sets, including the TeDDi sample itself. In future data collection, more effort should be put in representing languages with long words, especially because most of them are likely to be low-resource languages.

7 Conclusion

We have shown that the linguistic diversity of NLP data sets can be consistently assessed by two independent measures, TI (proposed in previous work) and J_{mm} (proposed in this paper). Both of these measures show that a high number of languages and language families included in a data set is not sufficient to ensure linguistic diversity.

To make the assessment of linguistic diversity automatic and rather simple, we show that text-based features such as the mean word length can be used as linguistic descriptors. These features can be easily calculated on very small text samples (of length of 500 tokens), overcoming the obstacles posed by the need to extract linguistic features from typological databases.

An advantage of the J_{mm} score over TI and other previous indicators of linguistic diversity is its capacity to show what kinds of languages are missing in a given data set in comparison to a reference. Assessing popular NLP data sets with this measure revealed that the most underrepresented languages are those with rich morphology. This kind of direct and transparent comparison can improve multilingual NLP coverage in the long run.

[14] https://github.com/google-research/bert/blob/master/multilingual.md

[15] We show the morphological diversity for convenience since visualising 103 syntactic features would required additional adaptations.

Limitations

Both measures of morphological diversity that we propose rely on text features (the mean word length). Although we show that the mean word length is strongly correlated with an independent measure of morphological complexity (WALS_C). it remains an aggregated measure (one feature per language). For an even more transparent measure of linguistic diversity, it would be desirable to obtain more nuanced morphological features.

Another limitation of our study is that we do not propose syntactic features that could be extracted from text. We focused here on the current gap in the available linguistic features (the lack of morphological features in `lang2vec`), but devising text-based syntactic features would deserve more attention in future work.

References

Rami Al-Rfou. 2015. *Polyglot: A massive multilingual natural language processing pipeline*. Ph.D. thesis, State University of New York at Stony Brook.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardzic. 2016. A comparison between morphological complexity measures: typological data vs. language corpora. In *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*, pages 142–153.

Aleksandrs Berdicevskis, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyan, Taraka Rama, et al. 2018. Using universal dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 8–17.

Balthasar Bickel, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and John B Lowe. 2017. The autotyp typological databases. version 0.1.2.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Bernard Comrie, Matthew S. Dryer, David Gil, and Martin Haspelmath. 2013. Introduction. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2018. *Glottolog 3.3*. Leipzig.

Martin Haspelmath. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, 11(1):119–132.

Martin Haspelmath. 2017. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica*, 51(s1000):31–80.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

P. Jaccard. 1912. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11:37–50.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pretraining, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings from LREC 2016*, pages 923–929. European Language Resources Association.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3158–3163.

Steven Moran. 2016. The ACQDIV database: Min(d)ing the ambient language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4423–4429, Portorož, Slovenia. European Language Resources Association (ELRA).

Steven Moran, Christian Bentz, Ximena Gutierrez-Vasques, Olga Pelloni, and Tanja Samardzic. 2022. TeDDi sample: Text data diversity sample for language comparison and multilingual NLP. In *Proceedings of the Thirteenth Language Resources*

and Evaluation Conference, pages 1150–1158, Marseille, France. European Language Resources Association.

Steven Moran and Michael Cysouw. 2018. *The Unicode cookbook for linguists*. Number 10 in Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Olga Pelloni, Anastassia Shaitarova, and Tanja Samardzic. 2022. Subword evenness (SuE) as a predictor of cross-lingual transfer to low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7428–7445, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Tanja Samardžić, Ximena Gutierrez-Vasques, Rob van der Goot, Max Müller-Eberstein, Olga Pelloni, and Barbara Plank. 2022. On language spaces, scales and cross-lingual transfer of UD parsers. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 266–281, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Richard Sproat and Alexander Gutkin. 2021. The Taxonomy of Writing Systems: How to Measure How Logographic a System Is. *Computational Linguistics*, 47(3):477–528.

Sabine Stoll and Balthasar Bickel. 2013. Capturing diversity in language acquisition research. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson, and Alan Timberlake, editors, *Language typology and historical contingency: studies in honor of Johanna Nichols*, pages 195–260. Benjamins, Amsterdam. [pre-print available at `http://www.psycholinguistics.uzh.ch/stoll/publications/stollbickel.sampling2012rev.pdf`].

T. T Tanimoto. 1958. *An elementary mathematical theory of classification and prediction*. International Business Machines Corporation.

Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, 39(1):15–22.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Alison Wray. 2015. Why are we so sure we know what a word is? In John R. Taylor, editor, *In: Taylor, John R. ed. The Oxford Handbook of the Word, Oxford Handbooks, Oxford: Oxford University Press, pp. 725-750.*, pages 725–750. Oxford University Press.

A Mean Word Length Correlation between Different Sample Size

To make sure that the stability across different sample sizes suggested by Figure 4 is not a mere consequence of a relatively small range of variation, we perform correlation tests between different samples and in comparison to other measures (TTR and unigram entropy (H)). Table 2 shows that the ranks of languages change considerably less across different sample sizes when considering the mean word length than in the other two measures.

Samples	MWL	H	TTR
500 tokens vs. max.	0.99	0.85	0.84
2K tokens vs. max	0.99	0.95	0.94

Table 2: Spearman rank correlation showing how much rankings of languages change with text measures taken on random samples of different size.

B Word length and morphological complexity

ISO396-3	MWL	C_{WALS}
abk	7.17	0.62
apu	7.67	0.60
arz	4.44	0.49
bsn	6.02	0.69
ckt	8.45	0.50
deu	4.87	0.55
ell	4.72	0.53
eng	4.18	0.42
eus	5.70	0.64
fin	6.23	0.66
fra	4.41	0.45
hae	5.91	0.53
hau	4.08	0.38
heb	3.94	0.54
ind	5.42	0.40
kan	5.22	0.65
kat	4.78	0.50
khk	5.66	0.53
kut	4.60	0.37
lvk	4.77	0.67
qvi	8.18	0.71
rus	4.79	0.52
spa	4.37	0.45
swh	5.72	0.71
tur	6.07	0.76
vie	3.20	0.21
yaq	5.31	0.57
yor	3.52	0.25
Spearmann correlation	$\rho = 0.69$	

Table 3: Mean Word length (MWL) and morphological complexity measure (C_{WALS})

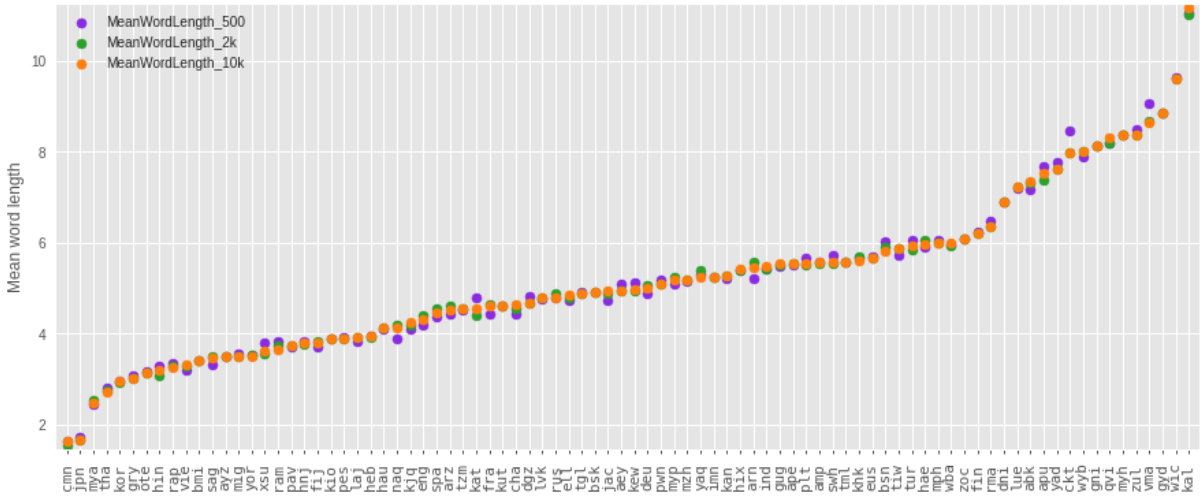

Figure 4: Mean word length measures at different text sizes in TeDDi. The languages on the x-axis are sorted according to the increasing value calculated on the biggest sample (10K). The values in the two smaller samples (2K and 500) depart very little from the main trend.

Chapter	Name	Categories	Transformation	Final Values
22A	Inflectional Synthesis	7 (ordinal)	none	1-7
26A	Prefixing vs. Suffixing in Inflectional Morphology	6 (non-ordinal)	binarization	0-1
27A	Reduplication	3 (non-ordinal)	binarization	0-1
28A	Case Syncretism	4 (ordinal)	reorder	1-4
29A	Syncretism in Verbal Person/Number marking	3 (ordinal)	none	1-3
30A	Number of Genders	5 (ordinal)	none	1-5
33A	Coding of Nominal Plurality	9 (partially ordinal)	binarization	0-1
34A	Occurrence of Nominal Plurality	6 (ordinal)	none	1-6
49A	Number of Cases	9 (ordinal)	remove	1-8
51A	Position of Case Affixes	9 (non-ordinal)	binarization	0-1
57A	Position of Pronominal Possessive Affixes	4 (non-ordinal)	binarization	0-1
59A	Possessive Classification	4 (ordinal)	none	1-4
65A	Perfective/Imperfective Aspect	binary	none	0-1
66A	The Past Tense	4 (ordinal)	reorder	1-4
67A	The Future Tense	binary	none	0-1
69A	Position of Tense/Aspect Affixes	5 (non-ordinal)	binarization	0-1
70A	The Morphological Imperative	5 (partially ordinal)	recategorization	1-4
73A	The Optative	binary	none	0-1
74A	Situational Possibility	3 (non-ordinal)	binarization	0-1
75A	Epistemic Possibility	3 (non-ordinal)	binarization	0-1
78A	Coding of Evidentiality	6 (non-ordinal)	binarization	0-1
94A	Subordination	5 (non-ordinal)	binarization	0-1
101A	Expression of Pronominal Subjects	6 (non-ordinal)	binarization	0-1
102A	Verbal Person Marking	5 (partially ordinal)	recategorization	1-3
111A	Nonperiphrastic Causative Constructions	4 (non-ordinal)	binarization	0-1
112A	Negative Morphemes	6 (non-ordinal)	binarization	0-1

Table 4: Subset of WALS features that we use for characterizing the morphological complexity of languages. The column "Final Values" gives the range of values each feature can take after transformations were performed to the original values (Bentz et al., 2016)

C Word Length Adjustments for Logographic Scripts

In the case of logographic scripts, we scale the observed word length proportionally to the difference between Chinese original script and Pinyin so that the scaled length is comparable to alphabetic scripts. We estimate the scalar s as a function of the character-level type-token ratio on a sample of text:

$$s_l = r \cdot ch\_ttr_{l\_500} \quad (3)$$

Where r is a constant representing the mean word length ratio between Pinyin and the original

Name and main references	N(L)	N(F)	TI_{syn}	J_{mm_syn}	TI_{morph}	J_{mm_morph}
Universal Dependencies (UD)	106*	20*	0.567	0.736	0.337	0.665
Bible 100	103*	30*	0.649	0.811	0.302	0.617
mBERT	97*	15*	0.559	0.710	0.316	0.617
XTREME	40	14	0.612	0.775	0.311	0.471
XGLUE	19	7*	0.517	0.674	0.297	0.580
XNLI	15	7*	0.557	0.711	0.321	0.704
XCOPA	11	11	0.586	0.737	0.336	0.634
TyDiQA	11	10	0.626	0.751	0.343	0.552
XQuAD	12*	6*	0.523	0.680	0.318	0.634
TeDDi	89	51	**0.706**	-	**0.361**	-

Table 5: Diversity of multilingual NLP data sets with adjustments for logographic scripts. Compared to the main results in Table 1, all TI_{morph} scores are slightly decreased and J_{mm_morph} slightly increased. The rankings of the t are mostly preserved, with the exception of XNLI, whose J_{mm_morph} ranking improves.

Chinese script, l is the language in question and ch_ttr_{500} is the character-level type-token ratio on a text sample of the length of 500 word-level tokens in a given language. The scalar should be applied only if its value is greater than 1.

Table 5 shows revised diversity scores after the mean word length adjustments for logographic scripts. In our study, only three languages Chinese, Japanese and Korean required adjustments.