Linguistic Diversity Scores for NLP Data Sets

Anonymous ACL submission

Abstract

Quantifying linguistic diversity in multilingual data sets is important for improving cross-linguistic coverage of NLP models. However, current linguistic diversity scores rely mostly on measures such as the number of languages in the sample, which are not very informative about the structural properties of languages. In this paper, we propose a score derived from the distribution of text statistics (mean word length) as a linguistic attribute suitable for cross-linguistic comparison. We compare NLP data sets (UD, Bible100. mBERT, XTREME, XGLUE, XNLI, XCOPA, TyDiQA, XQuAD) to a new data set designed specifically for the purpose of being typologically representative (WALS-SC). To do so, we apply a version of the Jaccard index (J_{mm}) suitable for comparing sets of measures. This diversity score can identify the types of languages that need to be included in multilingual data sets in order to reach broad linguistic coverage. We find, for example, that (poly)synthetic languages are missing in almost all data sets.

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 much diversity do we need? How can we define a desired or optimal diversity to set as a goal when composing multilingual data sets?

These questions are typically not addressed in multilingual NLP studies. However, they are important in assessing whether our methods generalise well across diverse languages, without the need to test them on each single language (even if we had the necessary data for all languages).

The aim of this paper is to start a discussion on how to define optimal diversity, and how to quantify the degree to which multilingual NLP data sets capture it. For this, we need a simple scalable method to describe and compare languages, ideally a numerical attribute that can be easily assigned to any language. To be able to describe low-resource languages, the value of the attribute should not depend on the data size. We also need a quantifiable definition of the desired diversity of the language sample, and a method to compare the actual diversity with the desired one.

We propose to use text statistics as a quantitative attribute for describing languages. As a representation of overall linguistic diversity, we propose to use a predefined sample of languages designed by linguists for that purpose — the 100-language-sample (100L) selected by the Word Atlas of Language Structures (WALS; Comrie et al. (2013)) to represent geographic and phylogenetic diversity. Since text data are needed for our language attributes and they are not easily accessible for all the languages in the 100L sample, we compile a new corpus which aims to cover the 100L sample — the WALS-sample corpus (WALS-SC). This new data set allows us to compare popular NLP data sets against an independent benchmark. As a comparison method, we propose to use a version of the Jaccard index suitable for comparing measures.

Thus, our study contributes a novel technique to estimate the linguistic diversity of a data set, which NLP researchers can easily apply and use as a complement to existing techniques. This helps researchers to make informed choices when designing a multilingual data set. Representing a wider spectrum of linguistic diversity is 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).

2 Related Work

Evaluating the linguistic diversity of data sets relies on comparable descriptions of languages. Therefore, we need to determine which languages are similar and which ones are dissimilar. Describing and comparing languages has a very rich tradition in linguistics, but the resulting descriptions tend to be rather language-specific, which makes cross-linguistic comparison a difficult task (Haspelmath, 2007).

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 7). 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 extracted from WALS.[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 databases). The score is called *typology* and it is calculated as the entropy of feature values (averaged per language).[2] Moran (2016) compose a sample of 10 maximally diverse languages selected from language clusters made with WALS and AUTOTYP features (Stoll and Bickel, 2013). Other work in NLP uses grammatical features (usually termed *typological*) for other purposes such as improving model performance or predicting the features (Ponti et al., 2019), not for sampling.

[1] An alternative typological database is AUTOTYP (Bickel et al., 2017).

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

Finally, languages can be described using features derived from various text statistics. These values could be the type-token ratio (TTR) or unigram entropy of a text, which have been shown to correlate with other 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 safe to regard 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 bast transfer language, which they divide into data-dependent (data size, token overlap, TTR) and data independent (various distance measures extracted from the URIEL database). 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). Turc et al. (2021) show that German is a better transfer language than English for some languages. Our proposal for assessing linguist diversity is relevant to these efforts too, as its key component is language comparison.

More generally, our work is intended to contribute to several wide-scope initiatives for improving the quality of data management in 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 WALS-SC as an Initial Capture of Overall Linguistic Diversity

The WALS-sample corpus (WALS-SC) is an ongoing collection effort for texts written in languages which are part of the WALS one hundred language sample. The WALS editors selected this language sample as guidelines for contributors of chapters.

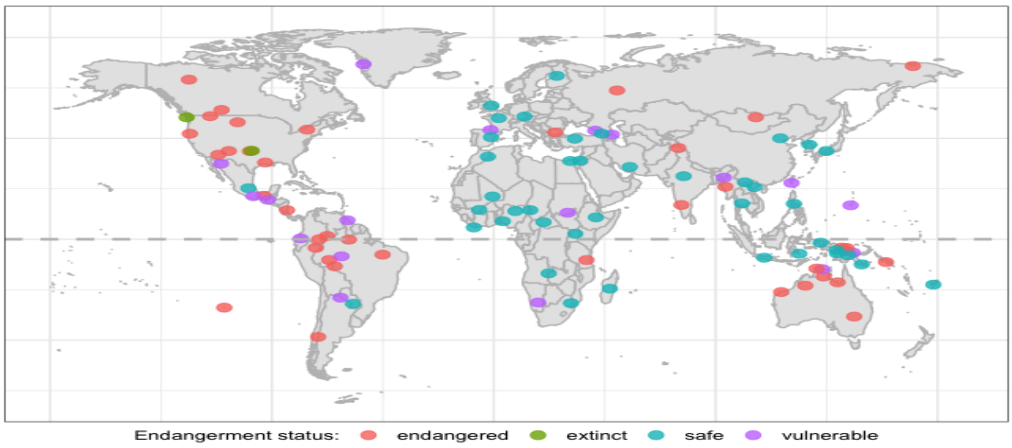

Figure 1: Geographic distribution of the languages included in the WALS 100L sample and their endangerment status.

They were asked to cover at least these one hundred languages in their collection efforts. The idea is to maximize genealogical (language family) and areal (geographic) diversity, and hence to minimize bias regarding the relative frequency of different types of linguistic features (Comrie et al., 2013). Figure 1 shows the languages and their endangerment status according to UNESCO.

The WALS-SC collection of text samples[3] aims at capturing cross-linguistic diversity in terms of languages and their modalities and genres by covering the WALS 100L sample. It is comprised of existing text resources, e.g., Project Gutenberg,[4] Open Subtitles (Lison and Tiedemann, 2016), The Parallel Bible Corpus (Mayer and Cysouw, 2014), the Universal Declaration of Human Rights,[5] and extended with manually collected translations, transcriptions, and grammatical annotations from sources of language documentation and description. Texts of various modes (spoken, written) and genres (conversation, technical, (non-)fiction) are included.

Due to the fact that the WALS-SC includes both high- and and low-resource languages, we have implemented a text sampling procedure to counterbalance the large divergence in text sizes. When we encounter a text with less than 50k word tokens, we include the entire document. For languages for which large corpora are already available, we randomly sample chunks of contiguous text of the length 50k word tokens. This procedure allows us to build corpora of comparable sizes for cross-linguistic comparison.

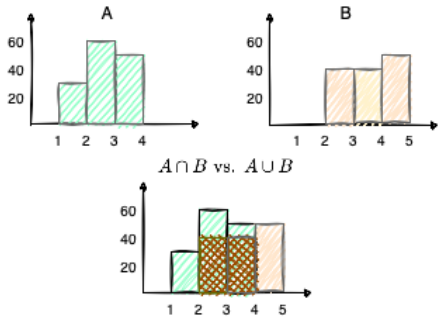

Figure 2: A toy example of comparing sets of measures with the minmax version of the Jaccard index.

Taken together, WALS-SC currently contains more than 100 million word tokens from genealogically and geographically diverse languages written in fifteen different scripts (see Appendix A for an overview). Given that its language sample is recommended by an influential team of typologist, we regard this data set as an initial approximation of the overall linguistic diversity, which can be improved in the future.

4 Comparing Data Sets with Jaccard Similarity

Our goal is to estimate the linguistic diversity of a data set with respect to some desired diversity (represented in our study by WALS-SC). Our score is thus a comparison between two data sets. We compare scaled distributions of the values of a numerical attribute as shown in Figure 2. The upper part of the figure shows (constructed) examples of

[3] In the final version, the link to the shared repository will be provided here.

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

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

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 WALS-SC data set. 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.[6]

Once we have represented our two sets in this way, we compare them using a generalised version of Jaccard similarityy. 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 ($x \in L$), y is its corresponding measurement ($y \in \mathbb{R}$) and the range of the index i is $1 \ldots |X|$. 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:

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

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. In the example in Figure 2, this gives the following vectors:
$\mathbf{a}$: $a_1 = 0, a_2 = 30, a_3 = 60, a_4 = 50, a_5 = 0$
$\mathbf{b}$: $b_1 = 0, b_2 = 0, b_3 = 40, b_4 = 40, b_5 = 50$

With these weights, we obtain the following similarity score:

$$J_{mm}(\mathbf{a}, \mathbf{b}) = \frac{0+0+40+40+0}{0+30+60+50+50} = \\ = \frac{80}{190} = 0.42 \quad (3)$$

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.

5 Mean Word Length as a Language Attribute

We now turn to the question of how to define and calculate a numerical attribute for calculating Jaccard minmax similarity. This needs to be one number that tells us something about the structural properties of each language.[7] Good candidates for such

[6] Another way to normalise the counts would be to divide them by the size of the set, but we chose the first option in order to preserve the notion of *number of languages*, which is helpful for the subsequent explanations.

[7] More generally, multiple attributes can be used too. In this scenario, languages would be embedded in a multidimensional space and clustered (instead of mono-dimensional bins that

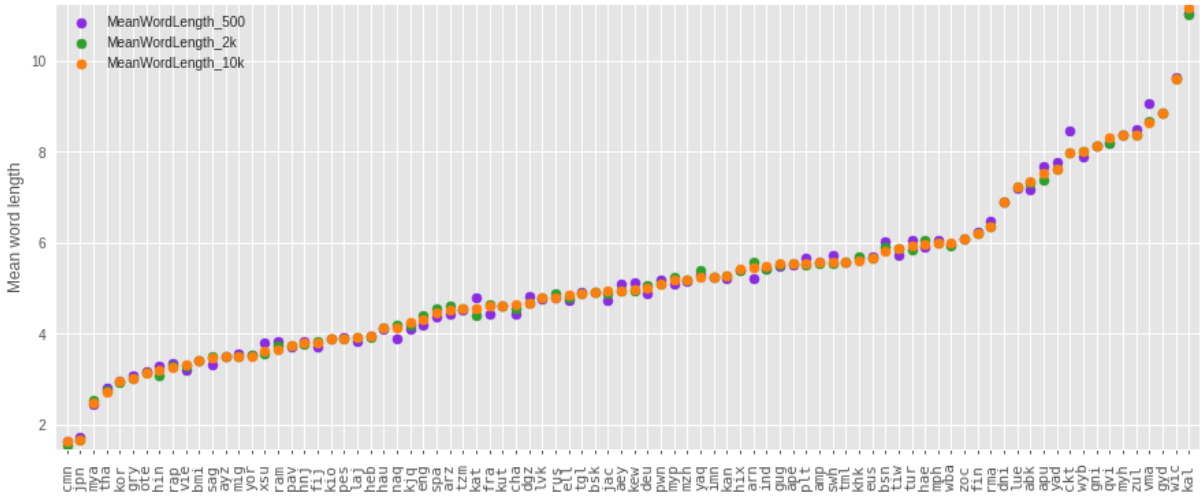

Figure 3: Mean word length measures at different text sizes in WALS-SC. The languages on the x-axis are sorted according to the increasing value calculated on the biggest sample (10K). The values in the two smaller samples (2K and 500) depart very little from the main trend.

attributes are diversity indices derived from typological databases and various complexity measures calculated from text (see Section 2). A limitation of the measures proposed before is that they all require considerable resources: either a detailed grammar description or a relatively big sample of text necessary to collect comparable statistics. In particular, token-to-type ratio (TTR) and text entropy are known to grow as a function of text size (Tweedie and Baayen, 1998; Bentz et al., 2017). While this growth is predictable, it makes the measure dependent on the data size.

What we propose instead is to measure the *mean word length* as a single attribute that differentiates between languages. This approach might appear simplistic given the ongoing discussion on the status of words as linguistic units (Haspelmath, 2017; Wray, 2015). In an NLP setting, we argue that word length is still a practical and meaningful measure that can be easily calculated and applied to any language, regardless of the size of the available resources. We come back to the limitations of this approach in Section 8.

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 split words into sequences of characters and take the length of character sequences as word length.[8] We apply this same definition to all scripts (see Section 8).

Word length is related to the structures of languages in several ways. The most prominent relation holds between word length and 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. Along another dimension of morphological diversity, we find longer words in (poly)synthetic languages vs. shorter words in analytic languages. Finally, morphological fusion in combination with rich morphology can lead to middle-length words. The interrelatedness between morphological types and other elements of grammar, e.g., word order (Sinnemäki, 2010; Ehret and Szmrecsanyi, 2016; Futrell et al., 2015), makes word length a more global attribute describing indirectly other properties of languages beyond morphology.

The relation between word length and word frequency follows from communicative efficiency of language (Zipf, 1949; Grzybek, 2007; Piantadosi et al., 2011; Bentz and Ferrer Cancho, 2016) connecting word length to unigram entropy and TTR, which both rely on word frequency.

This brings us to an important advantage of word length over other text statistics: 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. To

we use). Then, the comparison would be performed using more general methods for external cluster validation (Halkidi et al., 2001).

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

see this, consider Figure 3, which shows the values of the mean word length for the WALS-SC languages on random samples of the length 500, 2000 and 10000 tokens (values at 10000 are almost identical to overall values). Appendix C shows that languages are almost identically ranked with all the sample sizes.

Being a text feature, the mean word length can be calculated without matching languages in the sample to linguistic databases, which is very convenient for automatic screening of large samples. We can tell how diverse our samples are even if we do not know exactly what languages they contain.

6 Tests: Data and Methods

We calculate the Jaccard minmax score for a number of popular data sets in NLP.[10] Without attempting to provide an exhaustive evaluation, we review data sets that are 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 7.

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 database (top level family). Therefore, 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.

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 the 104 as mentioned in the original description.

Sampling from NLP data sets Since our numerical attribute (mean word length) can be calculated on small samples, we take a single random sample for each data set considered. 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.

Word and character segmentation We tokenise all the collected samples into word-level tokens using the Python library Polyglot (Al-Rfou, 2015). If a resulting token does not contain any alphanumeric characters, we discard it as punctuation. All the remaining tokens are further segmented into characters using the Python library `segments` (Moran and Cysouw, 2018).

Bin width We set the bin width for calculating J_{mm} to 1. This is a rather coarse level of granularity, which helps smaller samples get better scores and also accommodate some noise that can be found in such diverse samples. In addition to this, we also tried 0.5 as the width. We do not report the latter results, but the main trends did not change.

7 Findings: How Linguistically Diverse are NLP Data Sets?

Table 1 lists all the reviewed data sets together with some information about WALS-SC.

Comparing the data sets, we see that the Universal Dependencies data set agrees the most with WALS-SC, showing thus more diversity than usually believed. On the other hand, the coverage of the Bible 100 corpus is surprisingly low given the fact that the majority of its languages are non-Indo-European. Some much smaller language samples, such as XNLI and XCOPA get a better score than the Bible 100 sample.

If we compare our scores to Ponti et al. (2020), we see considerable agreement, but also some differences. Our score ranks XNLI and XCOPA higher, while TyDiQA and XQuAD get relatively low scores by both approaches, despite the careful language selection in TyDiQA.

Figure 4 shows where the data sets diverge the most. The main difference is whether a data set includes languages with long words or not (mean length > 8). Those samples that contain at least some languages with long words score much better on J_{mm} than those that remain completely on the short-middle side. Given the relationships between word length and the structure of language (discussed in Section 5), we believe this is just. The second important factor is a strong peak of the distribution indicating a bias towards one of the length bins (Bible100 and XGLUE). The third factor is a different ("wrong") shape of the distri-

[10] In the final version, the link to the shared code will be provided here.

Name and main references	N(L)	N(F)	Criteria / goal	TI	J_{mm}
Universal Dependencies (UD) (Nivre et al., 2020)	106*	20*	Bias towards Eurasia recognised but not intended	–	0.63
Bible 100 (Christodouloupoulos and Steedman, 2015)	103*	30*	Majority non-Indo-European	–	0.52
mBERT (GitHub repo[9])	97*	15*	Top 100 size of Wikipedia plus Thai and Mongolian	–	0.56
XTREME (Hu et al., 2020)	40	14	Diversity	0.42	0.41
XGLUE (Liang et al., 2020; Wang et al., 2019)	19	7*	–	–	0.50
XNLI (Conneau et al., 2018; Bowman et al., 2015; Williams et al., 2018)	15	7*	Span families, include low resource languages	0.39	0.58
XCOPA (Ponti et al., 2020)	11	11	Max diversity	0.41	0.57
TyDiQA (Clark et al., 2020)	11	10	Typological diversity	0.41	0.45
XQuAD (Artetxe et al., 2020; Rajpurkar et al., 2016)	12*	6*	Extension to new languages	0.36	0.52
WALS-SC (this paper)	86	51	WALS 100L sample coverage	–	–

Table 1: Multilingual NLP data sets with more than 10 languages in comparison to WALS-SC. N(L): the number of languages in the data set. N(F): the number of families to which the languages belong. TI: typology index by Ponti et al. (2020). J_{mm}: Jaccard minmax similarity (this paper).

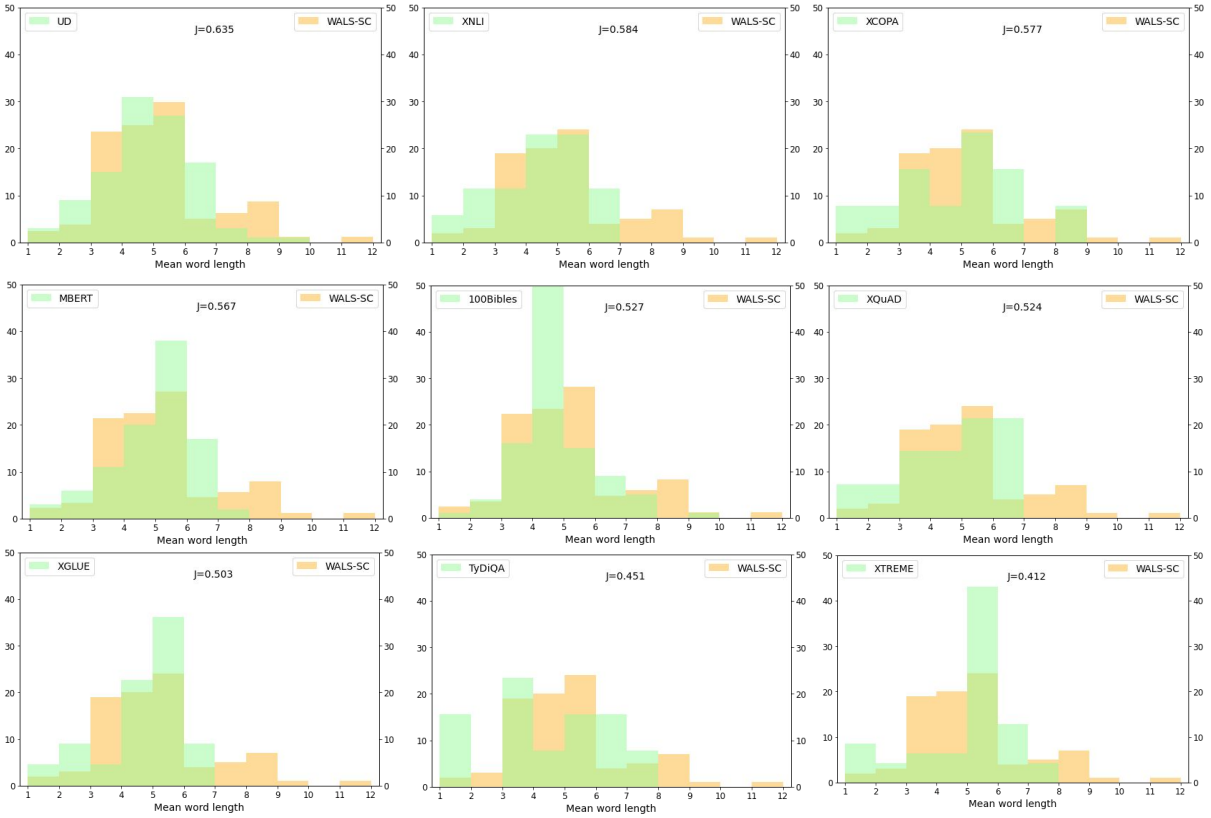

Figure 4: Union and intersection between the distributions of the mean word length in WALS-SC and NLP data sets.

bution (TyDiQA). The data set that agrees least with WALS-SC 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 WALS-SC 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.

8 Discussion

We have highlighted the advantages of our proposal for assessing linguistic diversity in NLP data sets by comparing the distributions of mean word length. Now we turn to this measure's limitations.

The main issue with word length as defined above is that Unicode characters represent different linguistic units, from low-level representations close to sounds in alphabetic scripts to high-level meaningful units in logographic scripts. This can, in principle, lead to overestimating or underestimating word length in some languages. While we can compare and score NLP data sets without knowing the true distribution of word lengths, a better estimation of this value would provide more sound and more interpretable measures of linguistic diversity.

For charting the true distribution of word lengths across languages, script normalisation would be needed. One way to approach this task would be to replace orthography with phonemic transcription. In this scenario, text samples from NLP data sets would be pre-processed with a grapheme-to-phoneme (g2p) model and the word length would be measured on its output. This approach is currently not feasible since the state-of-the art g2p performance depends considerably on the type of the script (Ashby et al., 2021). At the moment, g2p processing would introduce more confusion than normalization. However, the work on broad multilingual coverage of g2p models is ongoing and one might expect to see better solutions in the future. At the same time, a stronger international standardisation of the phonemic transcription would be needed in order to obtain actually comparable measures (Moran and Cysouw, 2018). Current practices are still rather varied (e.g., whether one uses narrow or broad transcription).

As an initial assessment of what would change if we used phonemic transcription, we have created a small parallel corpus of transcriptions of the short story *The North Wind and the Sun*, which is traditionally used for illustrating the sounds of various languages. For each language in our corpus (21 languages), we calculate the mean word length in two versions: orthographic and phonemic. We then perform a correlation test between these two variables and obtain a Spearman rank correlation of $\rho = 0.66$. This is a relatively strong correlation, but still indicating considerable differences. The list of languages and their mean word lengths is given in Appendix B. Such studies of a broader scope, together with various quantitative studies of the scripts (Sproat and Gutkin, 2021), will lead to better comparability of word lengths.

Another limitation of relying on word length is the fact that languages can be structurally different while belonging to the same word length bin even with normalised orthography. This could be a reason why several data sets have strong peaks in the middle of the word length distribution. Combining several numerical attributes with word length would be a way to obtain more nuanced language descriptions. For this, one would need to define attributes that are mutually independent, while all the text-based measures proposed so far are rather strongly correlated (see Section 2). Future work in this direction would need to address internal word structure more directly.

9 Conclusion

We have shown that NLP data sets can be assigned a linguistic diversity score by comparing their distribution of word lengths with the distribution found in the WALS-SC data set, which we compile as an initial capture of the overall linguistic diversity. The scores assigned by our method (J_{mm}) largely agree with previously proposed scores while providing a more fine-grained comparison and an initial upper bound. One finding that comes out of our analysis is that languages on the high end of the mean word length scale (> 8) are poorly represented even in the most diverse data sets. Thus, the most important challenge for future work on capturing full linguistic diversity is to broaden the diversity spectrum, which will require annotating and processing more low-resource languages of certain types.

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.

Lucas F.E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor, and Winnie Yan. 2021. Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 115–125, 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.

Chris Bentz and Ramon Ferrer Cancho. 2016. Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, pages 1–4. University of Tübingen.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. 2017. The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19(6):275.

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.

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

Katharina Ehret and Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In *Complexity, isolation, and variation*, pages 71–94. de Gruyter.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Quantifying word order freedom in dependency corpora. In *Proceedings of the third international conference on dependency linguistics (Depling 2015)*, pages 91–100.

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

Peter Grzybek, editor. 2007. *Contributions to the Science of Text and Language: Word Length Studies and Related Issues: 2nd, rev. paperback ed.* Kluwer, Dordrecht, NL.

Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. 2001. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145.

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 datasetfor cross-lingual pre-training, 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 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 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

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.

Kaius Sinnemäki. 2010. Word order in zero-marking languages. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 34(4):869–912.

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.

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

Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.

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.

George K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley.

A WALS-SC overview

Genre	Langs	Tokens	Scripts
conversation	7	5000	1
fiction	12	36 000 000	7
grammar	5	700	1
non-fiction	67	101 000 000	13
professional	39	80 000	15
Total	86	137 000 000	15

Table 2: Overview of basic statistics of the WALS-SC (as of November 2021).

B Orthographic vs. Phonemic Mean Word Length

	ISO396-3	MWL Orth	MWL Phon	Trans. type
1	aey	5.21	5.5	unk
2	arn	4.81	4.65	narrow
3	cmn	1.59	4.44	unk
4	deu	5	4.35	narrow
5	ell	4.62	4.23	unk
6	eng	4.19	3.46	narrow
7	eus	5.3	4.98	narrow
8	fra	4.55	3.18	broad
9	hau	3.8	4.07	narrow
10	heb	6.62	6.57	unk
11	hin	3.53	3.93	narrow
12	ind	5.92	5.25	unk
13	jpn	1.59	3.77	unk
14	kat	5.99	6.32	narrow
15	kor	2.85	6.56	unk
16	mya	10.22	8.15	unk
17	pes	3.99	5.03	unk
18	spa	4.62	4.36	narrow
19	tha	3.25	3.03	unk
20	tur	6.74	7.02	broad
21	vie	3.24	3.87	unk

Table 3: A comparison of the orthographic and the phonemic word length. For those languages where both broad and narrow transcriptions are available, we took the narrow version.

C Language Rank Correlation with Different Sample Size

To make sure that the stability across different sample sizes suggested by Figure 3 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 4 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 4: Spearman rank correlation showing how much rankings of languages change with text measures taken on random samples of different size.