A Linguistic Diversity Score for NLP Data Sets

Anonymous ACL submission

Abstract

TODO

1 Introduction

Data sets for training and testing NLP models are increasingly multilingual and aiming at broad linguistic coverage. They are often said to contain *typologically distinct* or *diverse* languages, including some low-resource ones.

The degree of linguistic diversity is typically described as the number of languages included in the data set and, less often, as the number of language families to which these languages belong. Both of these numbers do indeed indicate the level of linguistic diversity: the more languages and families, the more diversity. But how much diversity do we need? How can we define desired or optimal diversity to set as a goal when composing data sets? These question are typically not addressed in multilingual NLP, but they are important because we want to be able to assess whether our methods and approaches generalise well across languages without the need to test them on each single language (even if we had the necessary data for all languages).

With this paper, we would like to start a discussion on defining optimal diversity and quantifying the degree to which multilingual NLP data sets capture it. We propose an initial measure that can score linguistic coverage in NLP data sets by estimating how much of the full diversity spectrum is covered. For this, we need a simple scalable method to describe and compare languages, ideally a numeric 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 desired diversity and a method to compare the actual diversity with the desired one.

We propose to use text statistics as a simple quantitative descriptor. To represent and quantify desired diversity, we propose to compare all data sets against a predefined sample of languages chosen to maximize geographic and phylogenetic diversity. We consider the 100-language-sample (100L) defined by WALS (Comrie et al., 2013) to be such a sample. Since text data are needed for our language attributes and they are not easily accessible for all the languages in the 100L sample, we create a new corpus aiming to cover the 100L sample — the 100L corpus (100LC). 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 Jaccard index suitable for comparing measures.

The result of our study is a new technique to estimate linguistic diversity of a data set, which NLP researchers can easily apply and use it as a complement to existing techniques in order to make better informed choices when designing a multilingual data set. Representing the full spectrum of linguistic diversity is a way to better 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 linguistic diversity in data sets relies on comparable descriptions of languages: given their descriptions, 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 REFs.

Typically, we regard languages that belong to the same family to be similar. To know which language belongs to which family, we turn to two most popular authorities: WALS and Glottolog (Dryer and Haspelmath (2013)), which will sometimes give us different answers because they do not fully agree on the phylogeny of languages. Language families can also be a too coarse for 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] WALS is an atlas of worldwide linguistic diversity that describes the structural features and geographic locations of overall 2676 languages (Dryer and Haspelmath (2013)). 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* and it is calculated as entropy of feature values (average per language).[2] Moran (2016) compose a sample of 10 maximally diverse languages selected from language clusters made with WALS and AUTOTYP features. Other work in NLP uses grammatical features (usually termed *typological*) for other purposes such as improving model performance (Ponti et al., 2019) or predicting the features Bjerva, but not so much for sampling.

Finally, languages can be described using features derived from various text statistics. This could be the type-token ratio (TTR) or unigram entropy of a text, whose values have been shown to reflect morphological complexity REFs. Languages with similar values of TTR or entropy can be considered similar in this sense. Many other methods have been proposed for assessing linguistic complexity using text statistics REFs and they can all, 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 by far.

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; **?**) by focusing specifically on diversity assessments and data-independent scores for language comparison.

3 100LC Corpus as a Diversity Benchmark

The 100LC corpus is an ongoing collection effort for texts written in languages which are part of the WALS one hundred language sample. The editors selected this sample as an orientation for contributors of chapters. Contributors were asked to at least cover 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 from Glottolog 3.3 (Hammarström et al., 2018).

The 100LC corpus is comprised of existing text resources, e.g., Project Gutenberg,[3] Open Subtitles (Lison and Tiedemann, 2016), The Parallel Bible Corpus (Mayer and Cysouw, 2014), the Universal Declaration of Human Rights,[4] and extended with manually collected translations, transcriptions, and

[1] An alternative database is AUTOTYP REF.

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

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

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

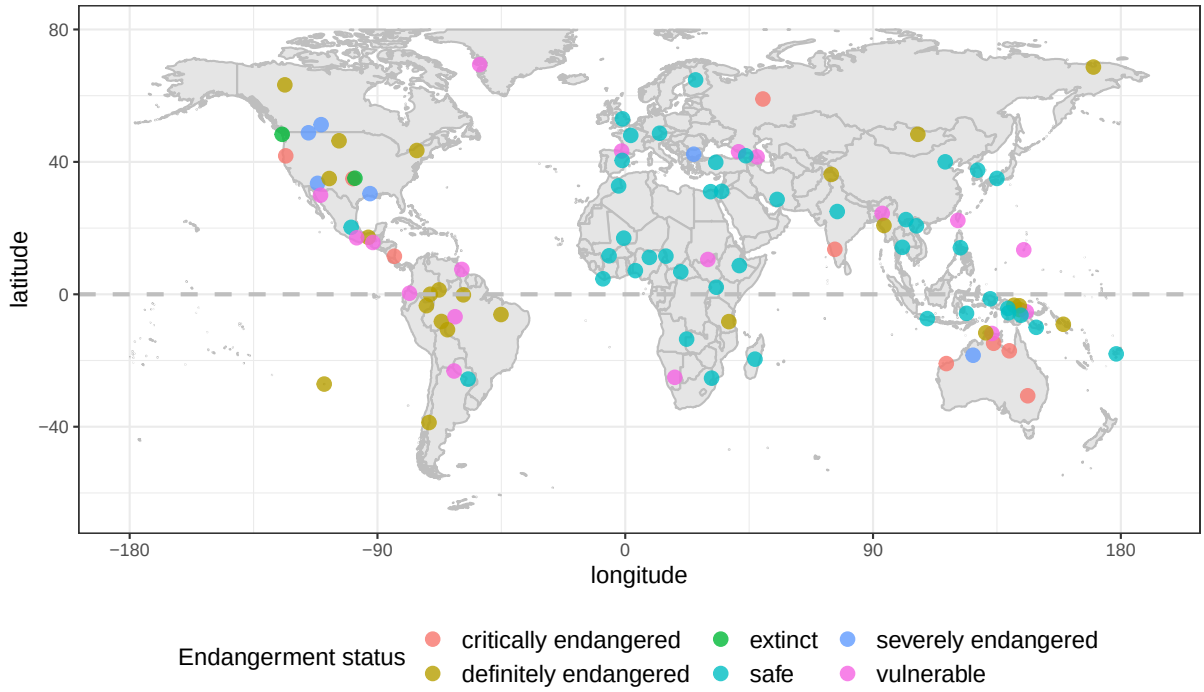

Figure 1: 100LC language sample and endangerment status

grammatical annotations from sources of language documentation and description.

The original texts are annotated to various degrees – from no annotation at all to several layers of phonemic/morphological annotations including translations into a meta language. Texts of various modes (spoken, written) and genres (conversation, technical, (non-)fiction) are included. The collection aims at capturing cross-linguistic diversity in terms of languages and their modalities and genres.

Due to the fact that the WALS sample includes minority 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, e.g. languages represented in Project Gutenberg and OpenSubtitles, we randomly sample 50k word tokens from chunks of contiguous text. This procedure allows us to build corpora of comparable sizes for cross-linguistic comparison. Taken together, 100LC currently contains more than 100 million word tokens from genealogically and geographically diverse languages (see Table 1 for an 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	84	137 000 000	15

Table 1: Overview of basic statistics of the 100LC (as of November 2021).

4 Comparing Data Sets with Jaccard Similarity

Our goal is to estimate linguistic diversity of a data set with respect to some desired diversity. Our score is thus a comparison between two data sets, where one of them represents the desired diversity. In our case, the WALS-SC is always taken as the desired diversity and all other data sets are compared with it.

To compare two data sets, we consider scaled distributions 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 WALS-SC data set. The values of the numerical attribute (one measurement per language) are on

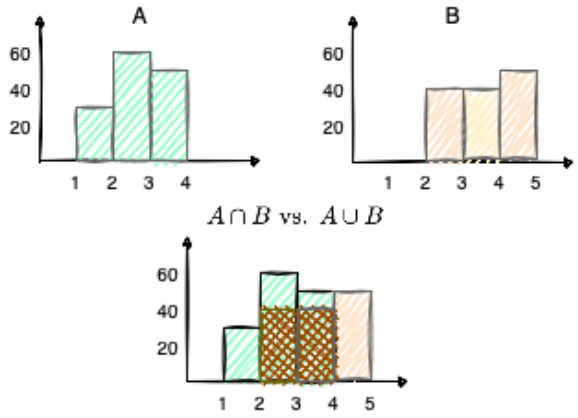

Figure 2: A toy example of comparing sets of measures with Jaccard index.

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. More narrow bins can 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 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.[5] We normalise the counts by multiplying 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. 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 explanations.

Once we have represented our two sets in this way, we compare them using generalised Jaccard similairy. 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).

Original Jaccard index REF compares two sets, but its generalised versions are suitable for comparing sets of measurements, which is exactly what we need. We thus use the *minmax* version of the score (J_{mm}), initially proposed by Tanimoto for comparing vectors of binary values and generlised to weighted vectors by Grefenstette. In this 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 the weight.

More 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 enumerator 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 our 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:

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

Higher values of J_{mm} indicate more similarity between A and B, and, indirectly, better coverage

[5] Diversity is not entirely independent of the size, but we will leave these considerations on the side for now.

Samples	MWL	H	TTR
500 tokens vs. max.	0.95	0.91	0.88
2K tokens vs. max	0.97	0.93	0.94

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

of linguistic diversity in A.

5 Mean Word Length as a Language Attribute

Here we 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.[6] Good candidates for such 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 REF. While this growth can be predicted, it makes the measure very 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 too simplistic given the ongoing discussion on the status of words as linguistic units REF and several factors that can influence their length independently of the structural properties of languages. We argue that it is a practical yet 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 and what can be done about them in Section 8.

We take words to be sequences of Unicode characters delimited by spaces or other conventional delimiters defined by common multilingual tokenisers. We split words into sequences of characters[7] and take the length of character sequences as word length. We apply this same definition to all scripts (we come back to this point in Section 8).

Word length is related to the structure of language 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 can relate longer words to (poly)synthetic languages vs. shorter words in analytic languages. Finally morphological fusion in combination with rich morphology can lead to middle-length words. EXs The interrelatedness between morphological types and other elements of grammar REFs 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: like all *good codes* language makes frequent messages shorter and infrequent longer. This relation has been known at least since Zipf (1949) and rather well studied in more recent work (Grzybek, 2007; Piantadosi et al., 2011) OTHER?. In this way, word length is indirectly related to common text-based measures of language complexity such as unigram entropy and TTR, which both rely on word frequency: languages with higher entropy and TTR have more rare words and these rare words can be expected to be relatively long. If we reverse this relationship, we can expect to see higher entropy and TTR having seen longer words.

This brings us to an important advantage of word length over other measures: it manifests itself in very small samples of text and remains stable across different sizes. Mean word length in a sample of contiguous text of only 500 tokens gives us already a very good estimation of the mean word length of the whole text. To see this, consider Figure 3, which shows the values of mean word length for the WALS-SC languages on random samples of length 500, 2000 and 10000 tokens (values at 10000 are almost identical to overall values).

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

[6] 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 we use). Then, the comparison would be performed using more general methods for external cluster validation Halkidi.

[7] We regard complex Unicode characters as single units, which means that dependent classes of characters are merged with their hosts.

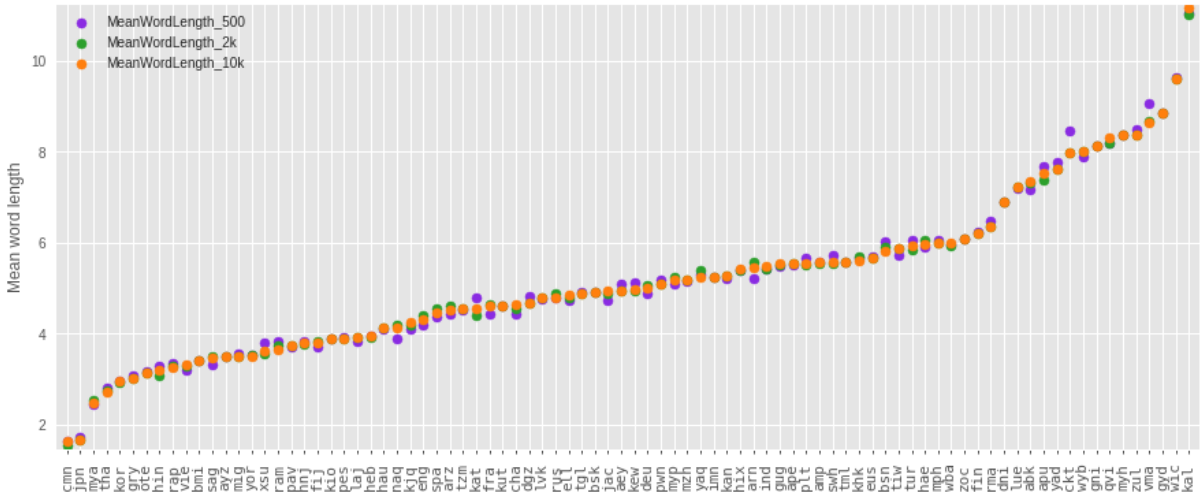

Figure 3: Mean word length measures at different text sizes in WALS-SC. The languages on the x-axis are sorted according to the increasing value calculated on the biggest sample (10K). The values in the two smaller samples (2K and 500) depart very little from the main trend.

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.

Being a text feature, the median 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.[8] 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 3 and discussed in more detail in Section 7.

Descriptions of the data sets often did not include all the information that was needed for our comparison. In particular, the number of language families was often not stated. To add this information, we extracted language names from the data files, converted these names into ISO codes manually, then retrieved the corresponding families from the Glottolog database (top level family). All 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 codes led to some changes in the number of languages, compared to those cited in the data descriptions. For instance, mBERT training data has 97 distinct languages, not 104 as mentioned in the original description.

Sampling from NLP data sets Since our numerical attribute (median 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 (or paragraphs) 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 can be just 200-300 tokens long.

Word and character segmentation We tokenise all the collected samples into word-level tokens using the Python library Polyglot REF. 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, 2018).[9]

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,

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

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

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 3 lists all the reviewed data sets together with some information about WALS-SC. For each data set, we list the main references, the number of languages and families.[11] We summarise the goals and criteria used for language selection (if available) in the fourth column. Finally, we report two diversity scores. TI in the fifth column is the typology score reported by Ponti et al. (2020) or by the creators of the data sets (Ruder et al., 2021). Our Jaccard minmax score (J_{mm}) is reported in the last column.

Comparing the data sets, we see that Universal Depenendencies agree 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 samples, such as XNLI and XCOPA get a better score than the Bible 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 TyDaQA and XQuAD get relatively low scores by both approaches, despite the careful language selection in TyDiQA.

Figure 4 allows us to see 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 high mean word length score much better on J_{mm} than those that remain completely on the low-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 distribution (TyDiQA). The data set that agrees least with WALS-SC is EXTREME, with 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, even though most of them are likely to be low-resource languages.

8 Discussion

So far, we have highlighted the advantages of our proposal for assessing linguistic diversity in NLP data sets by comparing the distributions of mean word length. Here we turn to its limitations.

8.1 Orthographic Incomparability

The main issue with word length as we define it 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. Words in languages with logographic scripts will tend to be shorter due to this fact, regardless of the structure of the language. On the other hand, the length will be longer in languages with alphabetic script, but with non-transparent orthography (e.g. words in English and French might appear longer than they are structurally). In this paper, we take a practical decision to neglect these factors. They can, in principle, conflate languages with different structures in the same word length bins (which might be the reason for the peaks that we see in some data sets), but in practice, we see the languages rather well discriminated without additional efforts to distinguish between such cases.

For charting the true distribution of word lengths across languages, script normalisation would be needed. While this remains beyond the scope of the current paper, we envision several possibilities for obtaining more comparable word lengths. The first option 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 calculated over its output. This approach is currently not feasible since the state-of-the art g2p performance depends considerably on the type of the script. At the moment, g2p processing would introduce more confusion than normalization. However, the work on broad multilingual coverage of g2p models is ongoing (Ashby et al., 2021) and one might expect to see better solutions in the future. At the same time, a stronger international standardisation of the phonemic tran-

[11] Multilingual data sets are currently released and updated at a fast pace. New versions of these data sets might be released after this paper is submitted causing some differences in the cited numbers.

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	NA	0.53
mBERT (GitHub repo[10])	97*	15*	Top 100 size of Wikipedia plus Thai and Mongolian	NA	TODO
XTREME-(R) (Ruder et al., 2021; Hu et al., 2020)	50	14	Diversity	0.42	TODO
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	0.39	0.58
XCOPA (Ponti et al., 2020)	11	11	Max diversity	0.41	0.58
TyDiQA (Clark et al., 2020)	11	10	Typological diversity	0.41	0.52
XQuAD (Artetxe et al., 2020; Rajpurkar et al., 2016)	12*	6*	Extension to new languages	0.36	0.36
WALS-SC (this paper)	86	51	WALS 100L sample coverage	–	–

Table 3: 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).

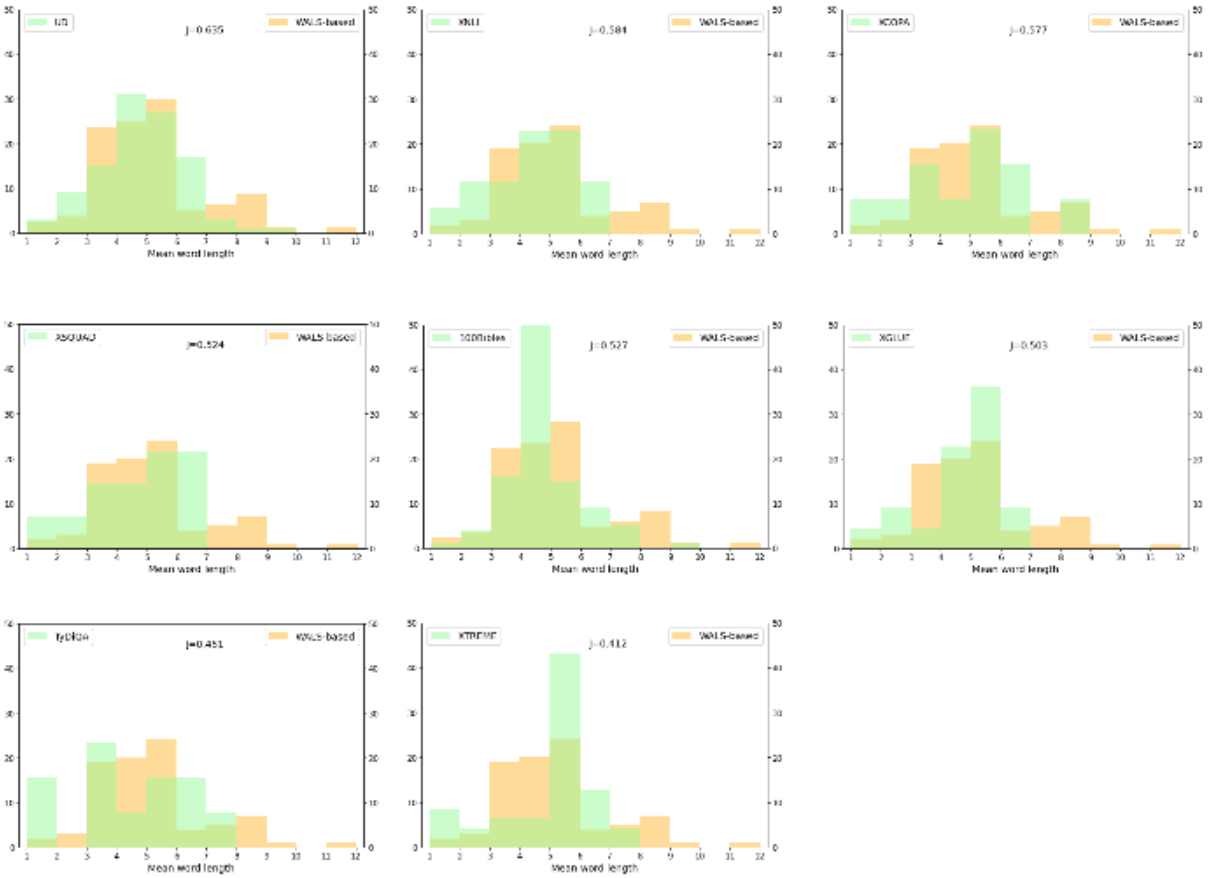

Figure 4: Union and intersection between 100LC and NLP data sets DRAFT

scription would be needed in order to obtain actually comparable measures. Current practices are still rather varied (e.g. whether one uses narrow or broad transcription, or something in between).

As an initial assessment of what we would gain with 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 (pho, 2010). For each language in our mini-corpus (22 languages), we calculate the mean word length in two versions: orthographic and phonemic.[12] We then perform a correlation test between these two variables and obtain the Spearman rank correlation of $\rho = 0.7$. We then manually inspect the disagreement between the two measures and find four languages whose mean word length changes considerably: it decrease for Burmese (from 10.22 to 8.15) and French (from 4.55 to 3.18); it increases for Korean (from 2.85 to 6.56) and Japanese (from 1.59 to 3.77). The values do not change much for the other languages (see APPENDIX). Such studies of a broader scope, together with various quantitative studies of the scripts (Sproat and Gutkin, 2021), might lead to better comparability of word lengths.

Another potential solution for cross-linguistic incomparability of word lengths would be via subword tokenisation. Gutierrez-Vasques et al. (2021) show that text entropy converges across languages after a small number of BPE merges. The merged vocabulary at this level might consist of comparable subword units across languages. It is clear how this would solve the problem of overestimating word length in alphabetic scripts, but it is still unclear how it would help with more logographic scripts. Curiously, the languages with these scripts converge to the same entropy values like all the other languages without any specific processing.

8.2 Conflating Linguistic Structure

The other limitation of relying on word length only is the fact that languages can be structurally different while belonging to the same word length bin. This could another 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 find 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 structural phenomena more directly.

[12] for those languages where both broad and narrow rtranscriptions are available, we take the narrow version.

9 Conclusion

References

2010. The principles of the international phonetic association (1949). *Journal of the International Phonetic Association*, 40(3):299–358.

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.

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.

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.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.

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.

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.

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.

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.

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.

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

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).

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.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation.

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.

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.

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