Why do language models perform worse for morphologically complex languages?

Catherine Arnett
Department of Linguistics
University of California San Diego
ccarnett@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
University of California San Diego
bkbergen@ucsd.edu

Abstract

Language models perform differently across languages. It has been previously suggested that morphological typology may explain some of this variability (Cotterell et al., 2018). We replicate previous analyses and find additional new evidence for a performance gap between agglutinative and fusional languages, where fusional languages, such as English, tend to have better language modeling performance than morphologically more complex languages like Turkish. We then propose and test three possible causes for this performance gap: morphological alignment of tokenizers, tokenization quality, and disparities in dataset sizes and measurement. To test the morphological alignment hypothesis, we present MorphScore, a tokenizer evaluation metric, and supporting datasets for 22 languages. We find some evidence that tokenization quality explains the performance gap, but none for the role of morphological alignment. Instead we find that the performance gap is most reduced when training datasets are of equivalent size across language types, but only when scaled according to the so-called "byte-premium"—the different encoding efficiencies of different languages and orthographies. These results suggest that languages of particular morphological types are not intrinsically advantaged or disadvantaged in language modeling. Differences in performance can be attributed to disparities in dataset size. These findings bear on ongoing efforts to improve performance for low-performing and under-resourced languages.

1 Introduction

An enduring goal in NLP is to develop language-general systems that achieve equal performance on all languages (Bender, 2011). Yet to date performance on languages other than English and a small number of high-resource languages remains extremely poor (Joshi et al., 2020; Ranathunga and de Silva, 2022; Søgaard, 2022; Atari et al., 2023; Ramesh et al., 2023). This has been attributed to a lack of research on non-English languages (Blasi et al., 2022), a lack of training data, and the possibility that evaluations are skewed towards high-resource languages (Choudhury, 2023).

Beyond these systemic biases, it's also possible that certain linguistic features lead to higher or lower language modeling performance. Specifically, it has been proposed that languages with more complex morphology are harder to model (Cotterell et al., 2018; Park et al., 2021). Languages with more inflectional classes are morphologically more complex, and thus harder to predict. This can be described in terms of enumerative complexity (Ackerman and Malouf, 2013).

Greater morphological complexity may lead to worse language model performance, as morphologically rich languages tend to have a large number of very infrequent word forms produced by combinations of morphemes, which leads to data sparsity (Shin and You, 2009; Bender, 2011; Botev et al., 2022). This claim finds empirical support in Gerz et al. (2018a), who demonstrated over a sample of 50 languages that morphologically rich (agglutinative) languages performed worse than less morphologically rich (fusional) languages. In the current work (§3), we replicate this analysis and extend it to much larger transformer models, both in monolingual and multilingual settings. We, too, find a robust performance gap between agglutinative and fusional languages.

This effect is surprising, as there are reasons to think that agglutinating languages should be *easier* for language models to learn. In studies on first language acquisition, children are observed to acquire more complex morphological systems earlier, especially systems that are uniform and transparent (Dressler, 2010). This may be due to

All code and data for this paper available below.
https://osf.io/jukzd/?view_only=3d0d491d24074215a0ab81f72a693c16

Proceedings of the 31st International Conference on Computational Linguistics, pages 6607–6623
January 19–24, 2025.

the fact that the form-meaning correspondences in these systems are more transparent, and thus more informative (Slobin, 1973, 2013, 2001; Dressler, 2010). By adulthood, there are no observed cross-linguistic differences in the level of acquisition of different languages according to morphological typology. Therefore, there is no linguistic evidence that would predict that any language should be harder to learn than any other language.

Identifying the causes for this performance gap could permit improved performance for morphologically rich languages (which are often low-resource) and reduce the performance inequity, potentially enabling users and researchers to be better able to use and do research on language models (Khanuja et al., 2023) in their own languages. We evaluate three possible explanations.

Hypothesis 1: Tokenization is not Morphologically Aligned

When the token boundaries for a given word line up with its morpheme boundaries, that tokenization is morphologically aligned. For example, the word 'books' in English is composed of the root 'book' and the plural morpheme '-s'. A morphologically aligned tokenization would be ['book', 's']. By contrast, ['boo', 'ks'] and ['b', ooks'] would be morphologically misaligned tokenizations.

Morphological alignment of the tokenizer – or lack thereof – could impact language modeling performance, especially for morphologically rich languages. For these languages, relatively frequent morphemes are combined to create a large number of unique word forms, which may be rare or completely novel. If the tokenizer does not segment words along morphological boundaries, it may be difficult for the language model to efficiently learn and represent the structure of the language. Additionally, this may be further exaggerated for morphologically complex languages, which tend to have longer words.

This hypothesis would predict that agglutinative languages have less morphologically aligned tokenizers than fusional languages and that morphological alignment negatively correlates with metrics of language model performance. To test this hypothesis, Section 4.1 introduces **MorphScore**, tokenizer evaluation for morphological alignment in 22 languages. To our knowledge, this is the first such multilingual evaluation for morphological alignment of tokenizers.

Hypothesis 2: Tokenization is Worse

Second, morphologically rich languages might tend to engender lower quality tokenizations. There is no current consensus on how to evaluate intrinsic tokenization quality (Zouhar et al., 2023; Chizhov et al., 2024). But compression is one of the most widely used metrics (Gallé, 2019; Rust et al., 2021, *inter alia*). It is usually measured as sequence length – the number of tokens needed to encode a sequence – or corpus token count (CTC; Schmidt et al., 2024). Better compression has been linked to better language modeling performance because it allows for more language data to fit into a fixed sequence length (Gallé, 2019; Liang et al., 2023; Dagan et al., 2024; Goldman et al., 2024); however, there is some evidence to suggest that compression is not directly linked to performance (Deletang et al., 2024; Schmidt et al., 2024).

Agglutinative languages might have worse compression on average because words tend to be longer (Fenk-Oczlon and Fenk, 1999; Berg et al., 2022) and there are more unique word forms (Sandra, 1994). In Turkish, for example, a single root may have millions of unique word forms (Hakkani-Tür et al., 2002). It is therefore less likely that the tokenizer will store whole words in its vocabulary, instead representing words using multiple tokens. This in turn may lead to worse compression and thus worse performance. If we find worse compression for agglutinative languages than for fusional languages, this may indicate that suboptimal compression is related to the performance gap.

Another proposed metric of tokenization quality is Rényi entropy (Zouhar et al., 2023), which measures how evenly distributed token frequencies are over the whole vocabulary, penalizing very high- and very low-frequency tokens. Rényi entropy has been shown to be predictive of downstream task performance (ibid). Because of their larger number of low-frequency word forms, it is possible that agglutinative languages have higher numbers of low-frequency tokens (specific inflectional forms) or higher numbers of high-frequency tokens (very high frequency morphemes used in many different word forms) than fusional languages. Therefore if agglutinative languages display worse (higher) Rényi entropy than fusional languages, this could indicate that inefficient token frequency distribution contributes to the performance gap.

In Section 5, we collect both compression and Rényi entropy and test whether agglutinating lan-

guages have worse compression and Rényi entropy, which would suggest that aspects of tokenization quality are driving the performance gap.

Hypothesis 3: Less Training Data

The role of data quantity for pre-trained language models is uncontroversial: the more, the better. In some cases, increasing data can improve performance more than increasing model size (Hoffmann et al., 2024). Western European high-resource languages tend to be less morphologically rich, and correspondingly, many morphologically rich languages are low-resource. Morphologically rich languages have less annotated data (Botev et al., 2022) and are less well researched. According to a survey by Blasi et al., despite having more speakers than most European languages, morphologically complex languages like Bengali, Swahili, and Korean have only a small number of studies. German, Romanian, French, and Italian have been better studied, despite having many fewer speakers (Blasi et al., 2022, Table 2). Therefore, data scarcity may be driving the observed performance gap between agglutinative languages and fusional languages.

Furthermore, recent work has shown that there are disparities in the number of bytes needed to convey the same amount of information in different languages (*byte premium*; Arnett et al., 2024a), due to orthographic encoding and linguistic reasons. Morphologically rich languages are more often written with non-Latin scripts, which require more bytes to be represented in common encoding standards like UTF-8. Morphologically rich languages also have longer words, which may amplify the effect. Byte premiums may thus exacerbate the data scarcity problem, and agglutinative languages may be trained on effectively less data even than it currently seems. Section 6 asks whether monolingual language models trained on byte-premium-scaled text demonstrate the previously observed performance gap.

2 Background

2.1 Morphological Typology

The field of morphological typology seeks to categorize languages according to their word formation strategies (Brown, 2010). Some languages primarily use words composed of a single morpheme or a small number of morphemes. Other languages incorporate many morphemes into a single word. This paper focuses on two types of languages: fusional and agglutinative languages. Fusional languages tend to encode multiple morpho-syntactic features into a single morpheme, where agglutinative languages tend to use different morphemes to represent each feature (Plank, 1999; Haspelmath, 2009; Dressler, 2010). As a result, agglutinative languages also tend to be polysynthetic (having words composed of many individual morphemes; Baker, 1996). For example, Turkish has separate plural and accusative morphemes, but in English, the root, tense, number, and person may all be loaded onto a single morpheme (Exs. (1) and (2)).

(1) tarla-lar-ı (Turkish)
field-PL-ACC
(Plank, 1999)

(2) are
be-PRES.2PL (English)

Typological categorization is much more complex than this binary categorical distinction. In order to connect this work with previous studies, it is helpful to use a very coarse view of morphological type; however, these properties are gradient. Languages may have both fusional and agglutinative properties (and properties of other morphological types, too). See Plank (1991; pp. 11-16) for discussion on this point.

2.2 Morphologically Aligned Tokenization

There is an area of active research on the relationship between morphological alignment of tokenizers and how it relates to language model performance. Work in this area often stems from the assumption that morphologically aligned tokenization is the gold standard for tokenization (Hofmann et al., 2022; Bauwens and Delobelle, 2024; Libovický and Helcl, 2024, *inter alia*). Morphologically-aware or aligned tokenization has been argued to lead to more meaningful tokens, which in turn leads to better performance (Banerjee and Bhattacharyya, 2018; Klein and Tsarfaty, 2020; Tan et al., 2020; Hofmann et al., 2021, 2022; Minixhofer et al., 2023; Bauwens and Delobelle, 2024). There is empirical evidence from several languages to support this claim, e.g. English (Jabbar, 2023), Korean (Lee et al., 2024) , Latvian (Pinnis et al., 2017), Arabic (Tawfik et al., 2019), Japanese (Bostrom and Durrett, 2020), Hebrew (Gueta et al., 2023), Kinyarwanda (Nzeyimana and Niyongabo Rubungo, 2022), and Uyghur (Abulimiti and Schultz, 2020). However, these efforts are

limited by the availability of morphologically annotated datasets (Minixhofer et al., 2023), which are often only available for a small number of relatively high-resource languages.

Some evidence also exists to the contrary (Zhou, 2018; Minixhofer et al., 2023; Gutierrez-Vasques et al., 2023), even for some of the same languages. Work on German and Czech (Macháček et al., 2018), Nepali, Sinhala, and Kazakh (Saleva and Lignos, 2021), Korean (Choo and Kim, 2023), Turkish (Kaya and Tantuğ, 2024), and Spanish (Arnett et al., 2024b) did not show any benefit of morphologically aligned tokenization. This is consistent with other work, e.g. Uzan et al. (2024), showing that BPE, which generally performs best on metrics such as compression, has the least morphologically meaningful tokens compared to other tokenization algorithms.

3 Evidence for a Performance Gap

This section describes three analyses that show lower performance for agglutinative languages. Previous analyses, which demonstrated evidence for the performance gap between fusional and agglutinative languages, all had significant confounds. We extend previous work by additionally controlling for amount of training data and extending to models which–as they are much larger and use the transformer architecture–better represent the state of the field.

3.1 Reanalysis of Gerz et al. (2018a)

Gerz et al. (2018b) analyzed a multilingual LSTM trained on 50 languages and found that fusional languages categorically outperformed agglutinative languages. This seminal finding is nevertheless limited in ways. It did not control for the number of training tokens, which was different for each language. We addressed this in a replication of the analysis on the original data, fitting a full linear model in R with morphological type and number of training tokens as fixed effects, predicting perplexity. We fit a reduced model with only number of training tokens as a fixed effect. An ANOVA showed that the full model explained more variance in the data than the reduced model ($F(3, 45)=5.221$, $p=0.004$)). After controlling for number of training tokens, there is still a significant effect of morphological type, where agglutinative languages had higher perplexities than fusional languages.

3.2 Multilingual Models

Evidence from Gerz et al. (2018a) comes from just one model. To extend this work, we test a number of more contemporary multilingual language models, including XGLM (Lin et al., 2022), BLOOM (Le Scao et al., 2023), mT0 (Muennighoff et al., 2023), MaLA (Lin et al., 2024), and LLaMA2 (Touvron et al., 2023).

We test these models across a variety of benchmarks: commonsense reasoning benchmark scores from XStoryCloze (Lin et al., 2022), XCOPA (Ponti et al., 2020), XNLI (Conneau et al., 2018), Wikipedia (Guo et al., 2020), and XWinograd (Muennighoff et al., 2023) reported in the BigScience BLOOM evaluation results[1] and the SIB-200 benchmark (Adelani et al., 2024), as reported in the release paper.

We combine all of the benchmark scores into one dataset. All scores are on a scale between 0 and 1. We use language family information from the WALS database (Dryer and Haspelmath, 2013) and annotate the morphological type according to grammars and linguistic articles about each language. For each language model, we calculate the proportion of training data for each language according to reported data quantities in tokens or bytes. If languages were upsampled for model training, we include upsampled proportions.

We fit a full linear mixed effects model in R (Bates, 2010) predicting benchmark score with morphological type and language family as fixed effects and model and task as random effects. We fit a reduced model that is the same as the full model, except without morphological type as a predictor. We run an ANOVA to compare model fit. We find that the full model (with morphological type as a fixed effect) explains more variance than the reduced model ($\chi^2(3) = 149.16$, $p<0.001$). Even after controlling for amount of training data, language family, model, and benchmark task, there is still a significant effect of morphological type, where fusional languages show better performance than agglutinative languages.

3.3 Monolingual Models

Both of the previous analyses measure performance of multilingual models. None of these models had controlled or balanced amounts of training data for the languages they were trained on. This intro-

[1] https://huggingface.co/datasets/bigscience/evaluation-results

duces a confound, because European languages are typically both higher resource and fusional. The lower-resource languages in this sample were more likely to be agglutinative. In this final analysis of the performance gap, we compare performance of a suite of 1,989 monolingual models from Chang et al. (2023), covering 252 languages, which were trained on matching numbers of tokens. For each language, there are up to 12 models, with up to three different model sizes and four different training corpus sizes. The three model sizes were tiny (4.6M parameters), mini (11.6M parameters), and small (29.5M parameters). The four dataset sizes were low-resource (1M tokens), medlow-resource (10M tokens), medhigh-resource (100M tokens), and high-resource (1B tokens). Perplexities were calculated using 500k held-out tokens. We use the same language family and morphological type data as in §3.2.

We use perplexity as a metric of performance. This is the only existing evaluation metric for all the languages represented by these models.

We fit a full linear regression with morphological type, model size, and dataset size as predictors. We also fit a reduced model with only model size and dataset size as predictors. We use an ANOVA to compare the fit of these two models. We find that morphological type explains variance above and beyond the other two predictors ($\chi^2(3) = 28.809$, p<0.001). We also fit full and reduced models with language family as an additional predictor. Even after accounting for language family, morphological type still explains additional variance ($\chi^2(3) = 3.3324$, p=0.02).

Morphological type is predictive of performance after controlling for model size and data amounts, which supports the other analyses.

3.4 Interim Discussion

Using both perplexities and benchmark scores as evaluation metrics, and evaluations from monolingual and multilingual models, we found a robust performance gap between agglutinative languages and fusional languages. This evidence amplifies prior work by using more evaluation metrics for more languages, with more contemporary multilingual and monolingual models trained with balanced training data.

The following sections test three factors that may be driving this gap, corresponding to the three hypotheses above: morphological alignment of the tokenizers (§4), tokenization quality (§5), and disparities in data measurement (§6).

4 H1: Morphological Alignment

Does differential morphological alignment of tokenizers in languages with more or less complex morphology explain their performance gap? We present a new evaluation framework, called MorphScore, which permits a comparison of morphological alignment across tokenizers and languages. We evaluate monolingual tokenizers for 22 languages and analyze the relationship between MorphScore and morphological type. Code and datasets for MorphScore are available on GitHub: `https://github.com/catherinearnett/morphscore`.

4.1 MorphScore: Evaluating Morphological Alignment of Tokenizers

Calculating MorphScore. To evaluate a tokenizer's MorphScore for each word in a test set, we assign a value of `1` if the tokenizer places a token boundary at the morpheme boundary of interest, regardless of other token boundaries. We assign a value of `0` if there is not a token boundary at the morpheme boundary of interest. We exclude items which contain no token boundaries (i.e. the entire word form is in the tokenizer's vocabulary), so as not to penalize the tokenizer for not segmenting the word. MorphScore is the mean of the assigned values across the dataset for a given language. See Table 1 for examples.

Languages. MorphScore uses datasets of morphologically annotated words. We created datasets for 22 languages, which are listed in Appendix A. Half are agglutinative languages and half are fusional, according to grammars and descriptions of the languages. Language selection was also balanced for resource level, where about half of the languages of each morphological type are higher-resource, and half lower-resource. The sample was designed to be as diverse as possible in terms of language family and writing system, given the other constraints. Note that all fusional languages in the sample are Indo-European, which reflects the distribution of fusional languages in the world's languages, but not all Indo-European languages are fusional (e.g. Armenian). Among the Indo-European languages, two are from the Indic branch (Gujarati, Urdu) and a variety of subgroups are represented: Slavic (Bulgarian, Slovenian, Croatian), Baltic (Lithuanian), Hellenic (Greek), Armenian,

Germanic (Swiss German, Icelandic), and Celtic (Irish). The other language families represented in the sample are Japonic, Koreanic, Dravidian, Kartvelian, Austronesian, Turkic, Niger-Congo, and Uralic, as well as an isolate (Basque).

Datasets. Each dataset is composed of words with their morpheme boundary annotations from Universal Dependencies[2] (UD) or UniMorph[3]. Words in MorphScore do not contain any umlaut or suppletion and the whole word form can be composed of the lemma and the morpheme (or the two morphemic units annotated). Most of the datasets only had one morpheme boundary annotation per word, with the exception of the Korean datasets. For Korean, when multiple morpheme boundaries were annotated, we chose the left-most boundary. We deduplicated items, and chose a random sample of 2000 for sets where there were more than 2000 items. We only included languages with at least 100 items.

4.2 Tokenizers

We use the monolingual tokenizers from Chang et al. (2023), which are from the same models used for perplexities in §3.3. Each tokenizer is a SentencePiece (Kudo and Richardson, 2018) tokenizer with a vocabulary size of up to 32k. Each tokenizer is trained on 10k lines of text randomly sampled from the model training data.

4.3 Results

We evaluate the tokenizers on their corresponding MorphScore dataset. MorphScores are reported in full in Table 3 in Appendix A. In order to address Hypothesis 1, we first conduct a two sample *t*-test to evaluate whether agglutinative languages have lower MorphScores than fusional languages. This would be consistent with the explanation that tokenizers are more likely to fail to align token boundaries with morpheme boundaries in agglutinative languages. To the contrary, we find that agglutinative languages have higher MorphScores (M=66.3%) than fusional languages (M=53.3%), a significant difference (t(20.874)=2.950, p=.008).

We also tested for a negative correlation between MorphScore and perplexity, such that better MorphScores were correlated with better performance. We fit a linear regression between the variables, but found no significant correlation (F(1, 13)=0.323, p=0.580).

4.4 Discussion

One possible explanation for this result is that words in agglutinative languages are on average being segmented into more tokens, making it more likely that a token boundary will fall on a morpheme boundary. This in turn could be driven by word length, as agglutinative languages tend to have longer words. It could also be due to a higher number of token boundaries per word (fertility), as higher fertility means that as there are more token boundaries, it becomes more likely that one of the token boundaries would fall on a morpheme boundary due to chance. Upon analysis, we found that agglutinative languages indeed had longer words (t(29,923)=18.222, p<0.001) and more tokens per word (t(37,375)=34.27, p<0.001). We fit a linear regression with number of tokens per word, word length in characters, and morphological types as predictors for MorphScore. We found that fertility and word length are both negatively correlated with MorphScore (χ^2(1)=61.457, p<0.001; χ^2(1)=364.03, p<0.001; respectively); however, the effect sizes were extremely small with an adjusted $R^2 = 0.021$. Given these small effects, longer words or higher fertility cannot explain the greater than 20% higher MorphScores for agglutinative languages.

In order to mitigate concern about the choice to exclude one-token words from the calculation of MorphScore, we also calculate MorphScore such that a one-token word is counted as correct. Agglutinative languages still had higher MorphScores than fusional languages (t(18.874) = 2.393, p = 0.027). Furthermore, we found no difference in the absolute number of one-token words (t(19.867) = -0.768, p = 0.452) nor in the proportion of one-token words (t(17.014) = -0.577, p =0.572) between agglutinative and fusional languages.

These results are inconsistent with Hypothesis 1; morphological tokenizer alignment (as measured by MorphScore) is higher for agglutinative languages rather than lower, and this effect cannot be explained by higher fertility or longer word length.

5 H2: Tokenization Quality

We next evaluate whether tokenization quality can explain the performance gap between agglutinative and fusional languages. We use two metrics of tok-

[2] https://universaldependencies.org/
[3] https://unimorph.github.io/

Language	Word	Source	Segmentation	Score
Basque	aldiz	morphemic	aldi + z	
		Tokenizer 1	[‘al’, ‘diz’]	0
		Tokenizer 2	[‘aldi’, ‘z’]	1
Croatian	suučesnika	morphemic	suučesnik + a	
		Tokenizer 1	[‘su’, ‘uče’, ‘s’, ‘nika’]	0
		Tokenizer 2	[‘su’, ‘u’, ‘če’, ‘s’, ‘nika’]	0
Icelandic	samráðs	morphemic	samráð + s	
		Tokenizer 1	[‘samráð’, ‘s’]	1
		Tokenizer 2	[‘samráðs’]	exclude
Greek	Αδριανής	morphemic	Αδριανή + ς	
		Tokenizer 1	[‘Α’, ‘δ’, ‘ριανής’]	0
		Tokenizer 2	[‘Α’, ‘δρ’, ‘ιανή’, ‘ς’]	1

Table 1: Example items with morphemic segmentations and tokenizations with MorphScores according to their morphological alignment.

enization quality: compression and Rényi entropy. To achieve sufficient statistical power, we use the same tokenizers as the previous section, but add all the languages from Chang et al. (2023) with FLORES datasets and for which we have morphological type labels, for a total of 63 languages. Perplexities for each tokenizer come from Chang et al. (2023).

5.1 Compression

We use corpus token count (CTC; also known as sequence length) as our measure of compression. CTC (Schmidt et al., 2024) is the number of tokens it takes to encode a text. Lower CTC indicates better compression, which is thought to have various effects on performance, cost, and inference time. If a tokenizer encodes a given text with more tokens, this will mean more sequences in order to pass the text through a language model. Each sequence, thus, will contain less information. This leads to higher training cost and slower inference (Song et al., 2021; Petrov et al., 2023; Yamaguchi et al., 2024) and worse model performance (Gallé, 2019; Liang et al., 2023; Goldman et al., 2024).

We calculate CTC based on FLORES-200 (NLLB Team et al., 2022) by encoding the text for each language with its respective tokenizer and counting the sequence length, not including beginning- and end-of-sequence tokens. FLORES offers parallel texts for each language, meaning that each text contains the same content, and sequence lengths should be comparable between languages.

5.2 Rényi entropy

Rényi entropy has been proposed as a metric of tokenization quality, as it measures the distribution of token frequencies over the tokenizer vocabulary, penalizing low- and high-frequency tokens. It has been shown to correlate with downstream performance (Zouhar et al., 2023).

Rényi entropy might also capture undesirable tokenizer properties that could be causing the performance gap. Agglutinative languages have longer words (Fenk-Oczlon and Fenk, 1999; Berg et al., 2022) and more unique word forms (Sandra, 1994). This means that a tokenizer with a fixed vocabulary size will necessarily use shorter tokens on average for an agglutinative language than for a fusional language[4]. Shorter tokens will have higher frequencies on average (Berg et al., 2022), and these tokens will carry less information, as the meaning of a word is distributed over more tokens.

We calculate Rényi entropy from the FLORES dataset for each language using `tokenization-scorer`[5] (Zouhar et al., 2023) with the recommended setting (α=2.5).

5.3 Results

Agglutinative languages have higher CTC (worse compression) than fusional languages (t(85.944)=2.507, p=0.014). On average, their sequences are 3.5% longer. However, there

[4] As there has not been previous empirical evidence to support this point, we test this. We use the same tokenizers as in the previous section and tokenize all the FLORES datasets for which we have corresponding monolingual tokenizers. We then calculate mean token length for the FLORES dataset. The mean token length for fusional languages was 2.92 characters and the mean token length for agglutinative languages was 3.25. This difference is statistically significant (t(68.36) = 3.236, p = 0.002).

[5] https://github.com/zouharvi/tokenization-scorer

is no correlation between CTC and perplexity (linear regression; F(1, 190)=2.05, p=0.154). This indicates that compression, at least measured in this way, does not explain the performance gap.

There is also a difference in Rényi entropy between agglutinative and fusional languages. Agglutinative languages have worse (higher) Rényi entropy (M=0.547) than fusional languages (M=0.488; t(150.53)=5.168, p<0.001).

In order to test whether Rényi entropy can help explain the performance gap, conduct a Likelihood Ratio Test comparing two linear mixed effects models. The full model predicts perplexity from morphological type, Rényi entropy, and CTC as fixed effects, with model size as a random intercept. We then fit a reduced model, removing morphological type as a fixed effect. We compare the models with an ANOVA and find that morphological type explains additional variance above and beyond the other predictors (χ^2(3)=29.464, p<0.001). This indicates that Rényi entropy does not explain all of the variance significantly explained by morphological type. A variance partitioning analysis using the `partR2` package in R (Stoffel et al., 2024) produces an R^2 for morphological type of 0.100 and for Rényi entropy of 0.030, while the full model R^2 is 0.144. Therefore the vast majority of the variance is still explained by morphological type. This suggests that Rényi entropy could explain only a small part of the performance gap.

5.4 Discussion

These results are inconsistent with the hypothesis that tokenizer compression explains poorer language modeling performance for agglutinative languages. Other results, e.g. Deletang et al. (2024); Schmidt et al. (2024), also show a lack of relationship between compression and language model performance. While compression indicates how much information can be represented in a fixed sequence length, the effect of compression may be outweighed by other features of a particular tokenizer, or language models may be able to overcome suboptimal tokenization. This is an area for further research, as it remains unclear what the best criteria are for intrinsic evaluation of tokenizers (Zouhar et al., 2023; Chizhov et al., 2024).

6 H3: Data Measurement Disparities

The final hypothesis for the performance gap is disparities in training data.

The monolingual models used in §3.3 and §5 were designed to be trained on comparable amounts of training data with comparable tokenizers (Chang et al., 2023). Nevertheless, there are differences in performance between languages. Chang et al. (2024) trained a similar suite of models (the Goldfish models), taking into account the byte premiums for each language.

Byte premiums (Arnett et al., 2024a) are the ratio of the number of bytes it takes to represent a content-matched text in different languages. For example, a text in a language with a byte premium of 3 relative to English will be three times larger in bytes than the content-matched English text file. One of the major contributors to byte premiums is the writing system used by a language. Latin characters are represented with a single byte in UTF-8 encoding. In the most extreme cases, characters for scripts like Khmer take three bytes per character, not including diacritics. As a result, some languages have byte premiums of up to 5 relative to English.

This has implications for many things, including how much text tokenizers are trained on. Most training data can be measured in number of tokens, but this isn't the case for tokenizer training data, as the tokenizer hasn't been trained yet. The Goldfish tokenizers and models are trained on byte-premium-scaled text quantities, which was designed to reduce the effects of the data measurement disparities between languages.

In this section, we test whether taking byte premiums into account can reduce or completely eliminate the performance gap. We annotate 154 languages for morphological type and use the same procedure as in §3 to test for the performance gap with the Goldfish models.

6.1 Results

The Goldfish models exhibit numerically higher perplexity for agglutinative (M=143.62) than fusional languages (M=132.63), but this difference is not statistically significant (t(137.36)=1.180, p=0.077). Therefore, after taking byte premiums into effect, the Goldfish models do not exhibit the same performance gap that was demonstrated in previous research and in Section 3 above.

We tested whether there was a relationship between byte premium and morphological type, and found that there was a marginally significant difference between byte premiums for agglutinative and fusional languages (*t*-test; t(157.9)=1.960,

p=0.0518).

6.2 Discussion

The results show that after taking into account byte premiums, there is no difference in performance according to morphological typology. Thus, accounting for byte premiums by scaling training data reduces most of the variance previously accounted for by morphological type. This suggests, therefore, that differences that seemed to be driven by morphological typology are actually being driven by disparities in dataset size measurement.

7 Discussion

We find that byte premiums explain the largest portion of the performance gap, which means that cross-lingual differences in text encoding size can explain these particular, previously documented performance differences. The results do not support the idea that some languages are harder to model than others, but it does seem that languages need to be treated differently, e.g. by scaling data quantities. This result can be used to inform how much data should be used to train tokenizers and language models, especially in low-resource or multilingual settings. By not taking byte premiums into account, we may be disadvantaging languages which are historically under-represented in the field, even when resources for them do exist.

While these results may be surprising based on the NLP literature, these results are consistent with evidence from language acquisition work, which has not shown any cross-linguistic differences in learnability of languages. These results are unsurprising from an empirical perspective, as work in Linguistics and NLP has consistently shown that more data will always facilitate better learning, irrespective of the complexity of the language.

There do seem to be limits to the learnability of linguistic systems. There are some language systems that linguistic theory predicts are impossible for humans to learn. Recent work has shown that language models are less successful at learning those languages, compared to existing and possible linguistic systems (Kallini et al., 2024). Therefore, we do not predict that these results hold for systems that are more complex than any attested natural language.

The results relating to Rényi entropy do suggest that there may be differences in tokenization which could be affecting performance, however more work is needed on this topic.

8 Conclusion

This paper first presented new evidence consistent with a performance gap between languages of different morphological types. We presented and tested three hypotheses as to the cause(s) for this performance gap: morphological alignment of the tokenizer, tokenization quality, and measurement disparities of dataset size. We found that while there was evidence that tokenization quality (as measured by Rényi entropy) plays a small role, dataset size seems to explain a large portion of the performance gap. After scaling training data according to byte premiums – a measure of how many bytes it takes to represent text in different languages – the performance gap goes away.

To do this work, we also created MorphScore, which is an evaluation method that can be used to evaluate the morphological alignment of tokenizers. We release the datasets needed to evaluate MorphScore in 22 languages: https://github.com/catherinearnett/morphscore.

These results raise questions about other unintended differences in the way languages are treated that could lead to differences in performance between languages. This is a critical issue for achieving language-general NLP systems and making language models perform equitably. While it does not seem that morphological typology is the primary reason for the observed performance gap, the initial observation led to greater understanding of crosslinguistic NLP. It is important to keep evaluating the dimensions along which languages vary and considering whether language technologies, such as LLMs, introduce inequalities between languages. We have yet to fully understand all the ways in which English-centric practices in NLP may have impeded progress for language models in other languages.

Limitations

For all of the analyses, we were limited by the number of languages for which we had morphological type annotations. These annotations are time-consuming and are themselves limited by the resources available, namely grammars and linguistic descriptions. The number of languages in the MorphScore analysis is even more limited. Having more annotations and datasets included in this work would make the analyses more reliable. This

is an important place for expansion in future work.

In the MorphScore datasets, we were also limited by the type of existing data. There were differences in domain and breadth in the Universal Dependencies and UniMorph datasets for each language. There are also different numbers of items in each dataset for each language. This means some languages will have more diversity among the items and there will be more statistical power than others, therefore the treatment of each language was not the same, which could introduce uncontrolled variance. Additionally, the morpheme boundaries that are annotated for different languages was not consistent. Some boundaries were inflectional and some were derivational. If there existed large datasets of both inflectional and derivational morphologically annotated words in a wide range of languages, this would have improved the robustness of the MorphScore results.

Finally, because the annotations in UD and UniMorph chose only one boundary (or, if there were multiple boundaries, we chose one), we can only evaluate whether the token boundaries align with the morpheme boundary we chose. We did this to limit confounds, as all but one dataset had one annotated boundary per word. Additionally, agglutinative languages would have more morpheme boundaries per word, which could skew results. However, there was no controlled selection process for which morpheme boundary was used for the MorphScore analysis, therefore this could have also affected results.

In the analysis of the Goldfish models, the evidence that byte premiums account for the performance gap is supported by a marginally significant difference between byte premiums according to their morphological type. It is possible that with an even larger sample of languages, the effect would instead meet the standard threshold for significance. We argue that in conjunction with the other results, it still demonstrates that taking byte premiums into account significantly reduces the performance gap.

Acknowledgments

We would like to thank the UC San Diego Social Sciences Computing Facility Team for the use of the Social Sciences Research and Development Environment (SSRDE) cluster. Experiments were conducted using hardware provided by the NVIDIA Corporation as part of an NVIDIA Academic Hardware Grant. We would also like to than the Language and Cognition Lab for their helpful input, especially Tyler Chang and James Michaelov.

References

Ayimunishagu Abulimiti and Tanja Schultz. 2020. Building language models for morphological rich low-resource languages using data from related donor languages: the case of Uyghur. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 271–276, Marseille, France. European Language Resources association.

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Aranes, Glyd Jun and Zeman, Dan . 2021. Ud cebuano-gja. Accessed: 2024-09-04.

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Koldo Gojenola, and Larraitz Uria. 2015. Automatic conversion of the basque dependency treebank to universal dependencies. In *Proceedings of the fourteenth international workshop on treebanks an linguistic theories (TLT14)*, pages 233–241.

Catherine Arnett, Tyler A. Chang, and Benjamin Bergen. 2024a. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 1–9, Torino, Italia. ELRA and ICCL.

Catherine Arnett, Pamela D Rivière, Tyler Chang, and Sean Trott. 2024b. Different tokenization schemes lead to comparable performance in Spanish number agreement. In *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–38, Mexico City, Mexico. Association for Computational Linguistics.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Mark Baker. 1996. *The polysynthesis parameter*. Oxford Studies in Comparative Syntax. Oxford University Press.

Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded

subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 55–60, New Orleans. Association for Computational Linguistics.

Douglas M Bates. 2010. lme4: Mixed-effects modeling with R.

Thomas Bauwens and Pieter Delobelle. 2024. BPE-knockout: Pruning pre-existing BPE tokenisers with backwards-compatible morphological semi-supervision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5810–5832, Mexico City, Mexico. Association for Computational Linguistics.

Jatayu Baxi and Brijesh Bhatt. 2021. Morpheme boundary detection & grammatical feature prediction for Gujarati : Dataset & model. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 369–377, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Emily M Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6.

Thomas Berg, Peter Zörnig, and Charlotte Lehr. 2022. The effects of type and token frequency on word length: a cross-linguistic study. *Glottotheory*, 13(2):173–209.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, et al. 2017. The hindi/urdu treebank project. In *Handbook of Linguistic Annotation*. Springer Press.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Georgie Botev, Arya D. McCarthy, Winston Wu, and David Yarowsky. 2022. Deciphering and characterizing out-of-vocabulary words for morphologically rich languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5309–5326, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Dunstan Patrick Brown. 2010. Morphological typology. In *Handbook of Linguistic Typology*, pages 487–503. Oxford University Press.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2023. When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages. *arXiv preprint arXiv:2311.09205*.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2024. Goldfish: Monolingual Language Models for 350 Languages. *arXiv preprint arXiv:2408.10441*.

Pavel Chizhov, Catherine Arnett, Elizaveta Korotkova, and Ivan P. Yamshchikov. 2024. BPE Gets Picky: Efficient Vocabulary Refinement During Tokenizer Training. *Preprint*, arXiv:arXiv:2409.04599. Preprint.

Sanghyun Choo and Wonjoon Kim. 2023. A study on the evaluation of tokenizer performance in natural language processing. *Applied Artificial Intelligence*, 37(1):2175112.

Monojit Choudhury. 2023. Generative AI has a language problem. *Nature Human Behaviour*, 7(11):1802–1803.

Jayeol Chun, Na-Rae Han, Jena D. Hwang, and Jinho D. Choi. 2018. Building Universal Dependency treebanks in Korean. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Gautier Dagan, Gabriel Synnaeve, and Baptiste Roziere. 2024. Getting the most out of your tokenizer for pre-training and domain adaptation. In *Forty-first International Conference on Machine Learning*.

Gregoire Deletang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. 2024. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*.

Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The Universal Dependencies treebank for

Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 33–38, Valencia, Spain. Association for Computational Linguistics.

Kaja Dobrovoljc and Nikola Ljubešić. 2022. Extending the SSJ Universal Dependencies treebank for Slovenian: Was it worth it? In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 15–22, Marseille, France. European Language Resources Association.

Wolfgang U Dressler. 2010. A typological approach to first language acquisition. *Language acquisition across linguistic and cognitive systems*, 52:109–124.

Matthew S. Dryer and Martin Haspelmath. 2013. WALS Online (v2020.3).

Gertraud Fenk-Oczlon and August Fenk. 1999. Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology*, 3:151–177.

Matthias Gallé. 2019. Investigating the effectiveness of BPE: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018a. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018b. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. 2024. Unpacking Tokenization: Evaluating Text Compression and its Correlation with Model Performance. *arXiv preprint arXiv:2403.06265*.

Eylon Gueta, Omer Goldman, and Reut Tsarfaty. 2023. Explicit Morphological Knowledge Improves Pretraining of Language Models for Hebrew. *arXiv e-prints*, pages arXiv–2311.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. Languages through the looking glass of BPE compression. *Computational Linguistics*, 49(4):943–1001.

Dilek Z Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2002. Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities*, 36:381–410.

Martin Haspelmath. 2009. An empirical test of the Agglutination Hypothesis. *Universals of language today*, pages 13–29.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2024. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Haris Jabbar. 2023. MorphPiece: Moving away from Statistical Language Representation. *arXiv preprint arXiv:2307.07262*.

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.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. *arXiv preprint arXiv:2401.06416*.

Yiğit Bekir Kaya and A Cüneyd Tantuğ. 2024. Effect of tokenization granularity for Turkish large language models. *Intelligent Systems with Applications*, 21:200335.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Septina Dian Larasati, Vladislav Kuboň, and Daniel Zeman. 2011. Indonesian morphology tool (morphind): Towards an indonesian corpus. In *Systems and Frameworks for Computational Morphology: Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011. Proceedings 2*, pages 119–129. Springer.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jungseob Lee, Hyeonseok Moon, Seungjun Lee, Chanjun Park, Sugyeong Eo, Hyunwoong Ko, Jaehyung Seo, Seungyoon Lee, and Heuiseok Lim. 2024. Length-aware byte pair encoding for mitigating oversegmentation in Korean machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2287–2303, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Jindřich Libovickỳ and Jindřich Helcl. 2024. Lexically grounded subword segmentation. *arXiv preprint arXiv:2406.13560*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. MaLA-500: Massive Language Adaptation of Large Language Models. *arXiv preprint arXiv:2401.13303*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Irina Lobzhanidze. 2022. *Finite-State Computational Morphology*. Springer.

Teresa Lynn and Jennifer Foster. 2016. Universal Dependencies for Irish. In *Proceedings of the Second Celtic Language Technology Workshop*, Paris, France.

Dominik Macháček, Jonáš Vidra, and Ondřej Bojar. 2018. Morphological and language-agnostic word segmentation for NMT. In *International Conference on Text, Speech, and Dialogue*, pages 277–284. Springer.

Kosuke Matsuzaki, Masaya Taniguchi, Kentaro Inui, and Keisuke Sakaguchi. 2024. J-UniMorph: Japanese morphological annotation through the universal feature schema. In *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 7–19, Mexico City, Mexico. Association for Computational Linguistics.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. CompoundPiece: Evaluating and improving decompounding performance of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 343–359, Singapore. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Hyunji Hayley Park, Katherine J Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36, pages 36963–36990. Curran Associates, Inc.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 Shared Task on Morphological Reinflection: Generalization Across Languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural machine translation for morphologically rich languages with improved subword units and synthetic data. In *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20*, pages 237–245. Springer.

Frans Plank. 1991. Of abundance and scantiness in inflection: A typological prelude. *Paradigms: the economy of inflection*, pages 1–39.

Frans Plank. 1999. Split morphology: How agglutination and flexion mix. *Linguistic Typology*, 3:279–340.

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.

Prokopis Prokopidis and Haris Papageorgiou. 2014. Experiments for dependency parsing of Greek. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 90–96, Dublin, Ireland. Dublin City University.

Loganathan Ramasamy and Zdeněk Žabokrtský. 2012. Prague dependency style treebank for Tamil. In *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1888–1894, İstanbul, Turkey.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Mohammad Sadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. 2013. Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 306–314, Atlanta, Georgia. Association for Computational Linguistics.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic parsed historical corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984, Istanbul, Turkey. European Language Resources Association (ELRA).

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Jonne Saleva and Constantine Lignos. 2021. The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Online. Association for Computational Linguistics.

Dominiek Sandra. 1994. The morphology of the mental lexicon: Internal word structure viewed from a psycholinguistic perspective. *Language and cognitive processes*, 9(3):227–269.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. *arXiv preprint arXiv:2402.18376*.

Hyopil Shin and Hyunjo You. 2009. Hybrid n-gram probability estimation in morphologically rich languages. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pages 511–520. Waseda University.

Kiril Simov, Petya Osenova, Alexander Simov, and Milen Kouylekov. 2005. Design and implementation of the bulgarian hpsg-based treebank. *Journal of Research on Language and Computation. Special Issue*, pages 495–522.

Dan I Slobin. 1973. Cognitive prerequisites for the development of grammar. In *Studies of child language development*, pages 175–208. Holt, Rinehart, & Winston.

Dan I Slobin. 2001. Form-function relations: how do children find out what they are? *Language acquisition and conceptual development*, 3:406.

Dan I Slobin. 2013. Crosslinguistic evidence for the language-making capacity. In *The crosslinguistic study of language acquisition*, pages 1157–1256. Psychology Press.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece Tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin A. Stoffel, Shinichi Nakagawa, and Holger Schielzeth. 2024. partr2: Partitioning r2 in glmms. Https://CRAN.R-project.org/package=partR2.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online. Association for Computational Linguistics.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ahmed Tawfik, Mahitab Emam, Khaled Essam, Robert Nabil, and Hany Hassan. 2019. Morphology-aware word-segmentation in dialectal Arabic adaptation of neural machine translation. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 11–17, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Omri Uzan, Craig W Schmidt, Chris Tanner, and Yuval Pinter. 2024. Greed is All You Need: An Evaluation of Tokenizer Inference Methods. *arXiv preprint arXiv:2403.01289*.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, Valletta, Malta.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON*

Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 1–39, Online. Association for Computational Linguistics.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024. An Empirical Study on Cross-lingual Vocabulary Adaptation for Efficient Generative LLM Inference. *arXiv preprint arXiv:2402.10712*.

Marat M. Yavrumyan and Anna S. Danielyan. 2020. Universal Dependencies and the Armenian Treebank. *Herald of the Social Sciences*, 2:231–244.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Giulio Zhou. 2018. Morphological zero-shot neural machine translation.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A MorphScore

Table 2 reports the languages represented in MorphScore, their writing systems, language families, morphological types, and the number of items in each dataset.

The sources used for each language are:

- Bulgarian: UD_Bulgarian-BTB train split (Simov et al., 2005)
- English: UD_English-GUM train split (Zeldes, 2017)
- Spanish: UD_Spanish-AnCora train split (Taulé et al., 2008)
- Greek: UD_Greek-GUD train split (Prokopidis and Papageorgiou, 2014)
- Persian: UD_Persian-PerDT train split (Rasooli et al., 2013)
- Japanese: (Matsuzaki et al., 2024)
- Korean: UD_Korean-Kaist train split (Chun et al., 2018)
- Turkish: UniMorph (Pimentel et al., 2021)
- Indonesian: UD_Indonesian-GSD (Larasati et al., 2011)
- Hungarian: UD_Hungarian-Szeged train split (Vincze et al., 2010)
- Urdu: UD_Urdu-UDTB train split (Palmer et al., 2009; Bhat et al., 2017)
- Slovenian: UD_Slovenian-SSJ train split (Dobrovoljc et al., 2017; Dobrovoljc and Ljubešić, 2022)
- Tamil: UD_Tamil-TTB train split (Ramasamy and Žabokrtský, 2012)
- Georgian: UD_Georgian-GLC test split (Lobzhanidze, 2022)
- Armenian: UD_Armenian-BSUT train split (Yavrumyan and Danielyan, 2020)
- Irish: UD_Irish-IDT train split (Lynn and Foster, 2016)
- Icelandic: UD_Icelandic-Modern train split (Rögnvaldsson et al., 2012)
- Gujarati: UniMorph (Baxi and Bhatt, 2021)
- Kurdish: UniMorph (Kirov et al., 2018)
- Cebuano: UD_Cebuano-GJA test split (Aranes, Glyd Jun and Zeman, Dan , 2021)
- Basque: UD_Basque-BDT train split (Aranzabe et al., 2015)
- Zulu: UniMorph (Vylomova et al., 2020)

Full MorphScore results for the tokenizers from Chang et al. (2023) are reported in Table 3.

Language	ISO 639-3	Writing Sys. (ISO 15924)	Lang. Family	Morph. Type	Num. Items
Armenian	hye	armn	Indo-European	agglutinative	2000
Basque	eus	latn	Basque	agglutinative	2000
Bulgarian	bul	cyrl	Indo-European	fusional	2000
Cebuano	ceb	latn	Austronesian	agglutinative	131
English	eng	latn	Indo-European	fusional	2000
Georgian	kat	geor	Kartvelian	agglutinative	200
Greek	ell	grek	Indo-European	fusional	112
Gujarati	guj	gujr	Indo-European	fusional	547
Hungarian	hun	latn	Uralic	agglutinative	2000
Icelandic	isl	latn	Indo-European	fusional	1852
Indonesian	ind	latn	Austronesian	agglutinative	1552
Irish	gle	latn	Indo-European	fusional	1877
Japanese	jpn	jpan	Japonic	agglutinative	2000
Korean	kor	hang	Koreanic	agglutinative	2000
Northern Kurdish	kmr	latn	Indo-European	fusional	319
Persian	pes	arab	Indo-European	fusional	2000
Slovenian	slv	latn	Indo-European	fusional	2000
Spanish	spa	latn	Indo-European	fusional	2000
Tamil	tam	taml	Dravidian	agglutinative	884
Turkish	tur	latn	Turkic	agglutinative	2000
Urdu	urd	arab	Indo-European	fusional	1649
Zulu	zul	latn	Niger-Congo	agglutinative	2000

Table 2: Languages for which we created morphological datasets and evaluated MorphScore.

Lang	Lang. Name	MorphScore	Morph. Type
hye_armn	Armenian	0.634	agg
eus_latn	Basque	0.724	agg
bul_cyrl	Bulgarian	0.584	fus
ceb_latn	Cebuano	0.806	agg
eng_latn	English	0.781	fus
kat_geor	Georgian	0.660	agg
ell_grek	Greek	0.586	fus
guj_gujr	Gujarati	0.347	fus
hun_latn	Hungarian	0.739	agg
isl_latn	Icelandic	0.574	fus
ind_latn	Indonesian	0.708	agg
gle_latn	Irish	0.468	fus
jpn_jpan	Japanese	0.691	agg
kor_hang	Korean	0.692	agg
kmr_latn	Kurdish	0.202	fus
pes_arab	Persian	0.345	fus
slv_latn	Slovenian	0.650	fus
spa_latn	Spanish	0.592	fus
tam_taml	Tamil	0.435	agg
tur_latn	Turkish	0.591	agg
urd_arab	Urdu	0.747	fus
zul_latn	Zulu	0.541	agg

Table 3: MorphScore results from Section 4.