The Effects of Corpus Choice and Morphosyntax on Multilingual Space Induction

Vinit Ravishankar[§] **Joakim Nivre**[†]
[§]Department of Informatics, University of Oslo
[†]RISE Research Institutes of Sweden
[†]Dept. of Linguistics and Philology, Uppsala University
[§]vinitr@ifi.uio.no [†]joakim.nivre@ri.se

Abstract

In an effort to study the inductive biases of language models, numerous studies have attempted to use linguistically motivated tasks as a proxy of sorts, wherein performance on these tasks would imply an inductive bias towards a specific linguistic phenomenon. In this study, we attempt to analyse the inductive biases of language models with respect to natural language phenomena in the context of building multilingual embedding spaces. We sample corpora from 2 sources in 15 languages and train language models on pseudo-bilingual variants of each corpus, created by duplicating each corpus and shifting token indices for half the resulting corpus. We evaluate the cross-lingual capabilities of these LMs, and show that while correlations with language families tend to be weak, other corpus-level characteristics, such as type-token ratio, tend to be more strongly correlated. Finally, we show that multilingual spaces can be built, albeit less effectively, even when additional destructive perturbations are applied to the training corpora, implying that (effectively) bag-of-words models also have an inductive bias that is sufficient for inducing multilingual spaces.

1 Introduction

A variety of proxies and analytical methods have been used to study the inductive biases of language models towards natural language. This work includes targeted syntactic evaluation (Gulordava et al., 2018; Linzen et al., 2016), language model responses to formulaic synthetic languages (Ravfogel et al., 2019; White and Cotterell, 2021), as well as attempts to correlate differences in language modeling performance to language features over a wide range of languages (Cotterell et al., 2018).

In this paper, we combine two strands that have, of late, been fairly active research threads. The first of these concerns the inductive biases of language models towards languages that exhibit a specific grammar; the second addresses the inductive biases of these models towards multilingualism, which in this context refers to a model's ability to build a multilingual space (rather than distinct monolingual spaces), when trained on corpora consisting of text in multiple languages.

Prior work in this domain is focused on either a) quantifying language model performance across a variety of languages, or b) studying the effects of different architectural components on the quality of the induced multilingual space. We attempt to unite the two strands of research by studying transformer-based masked language models in an effort to quantify the extent to which the grammar of the language being modelled affects the model's ability to build a multilingual space. We use Dufter and Schütze's (2021) metrics, namely word translation and sentence retrieval, as a proxy for the utility of this space. Our main findings are:

- Masked language models are capable of building multilingual spaces even when destructive perrturbations, like lemmatisation and shuffling, are applied to the training corpora.
- Multilingual performance is only weakly correlated with languages and language families.
- Multilingual performance correlates better with corpus-level statistics like type-token ratio, and the frequency of *hapax legomena*.

2 Related Work

Language modelling There has been a considerable amount of research addressing inductive biases that language models may have towards specific grammatical patterns, or towards natural languages with specific structures. An early study by Cotterell et al. (2018) demonstrates, over 21 languages, that certain languages are harder to model than others; the authors find that model performance correlates with the richness of a language's (inflectional) morphology. Later work by Mielke et al. (2019) shows

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4130–4139
December 7-11, 2022 ©2022 Association for Computational Linguistics

contradictory findings; the authors extend these experiments to 69 languages and find that morphological complexity does not correlate as strongly with performance as simpler factors like vocabulary size and sentence length do.

Other work involves studying how language modelling is affected by manually altering corpora. Ravfogel et al. (2019) train RNN-based models on English, altered to display different word orders and different degrees of morphological agreement; White and Cotterell (2021) generate corpora of natural language sentences, with constituents permuted based on Boolean switches, and show that recurrent language models show little variance in performance across word orders, compared to transformers.

Multilingualism Moving beyond monolingual language modelling, we examine the numerous works analysing what precisely multilingual language models need, in order to form an adequate multilingual space, which is quantified by measuring a model's performance on some multilingual task. Pires et al. (2019) show that subword overlap tends to improve multilingual alignment, though overlap is by no means necessary, as languages with different scripts can exist in the same multilingual space. Deshpande et al. (2021) show that while structurally similar languages do not necessarily need subword overlap, dissimilar languages rely heavily on overlap; they also show that well-aligned non-contextual word embedding spaces allow for better transfer.

On the other hand, Artetxe et al. (2020) have somewhat contradictory results, and show that neither shared vocabulary items nor joint pre-training are essential to build a multilingual encoder. K et al. (2020) and Dufter and Schütze (2021) analyse encoders from an architectural point of view. The former work shows that model depth (and not the number of attention heads) contributes to transfer performance, even when the number of parameters is kept constant. The latter points out that multilingual spaces exist because languages are forced to share parameters, and that even in the absence of shared subwords and special tokens, position embeddings play a significant role in building these spaces. Dufter and Schütze (2021) go on to show that the removal of shared position embeddings is sufficient to reduce a model's multilingual performance (as measured on word translation and sentence retrieval) to approximately random. This, we show, is not universally the case.

3 Methodology

3.1 General approach

In order to evaluate the quality of our models' multilingual spaces, we use word translation and sentence retrieval as proxy tasks; this contrasts with, for example, Deshpande et al. (2021), who use (zero-shot) transfer performance instead. We avoid this largely due to performance constraints: small models are unlikely to be parameterised enough to handle transfer.

To create synthetic multilingual (more precisely, bilingual) corpora, we follow the approach of K et al. (2020) and Dufter and Schütze (2021). Starting from a monolingual corpus, we shift the vocabulary index for every token in the original corpus up by the model's vocabulary size. For instance, the token `convenient`, with token index 42, would have a "mirror" `::convenient`, with token index 2090. This effectively gives us a parallel second half, which has the same structure as the original language, but a guarantee of no vocabulary overlap.

While this is a somewhat unrealistic simulation – after all, multilingual models are trained on languages with *different* structures – we use our formulation in order to a) have a simplified test bed where the *structure* of the language plays a role, but the *structural differences* between the two languages are ignored; and b) to avoid the complexity of the experimental space from exploding, when each language can conceivably be paired with every other language.

3.2 Data

In an effort to have a reasonably comprehensive search space of languages, we experiment over two corpora (Wikipedia and Common Crawl) and fifteen languages – namely Arabic, Czech, Danish, German, English, Spanish, Finnish, French, Hebrew, Italian, Dutch, Polish, Portuguese, Russian and Swedish. While Indo-European languages are still rather overrepresented in our data, these languages exhibit a wide range of head-depedendent entropies (Levshina, 2019). This is also part of the reason we avoid completely synthetic corpora: while it is trivial to generate synthetic corpora from some descriptive grammar, the *stochasticity* and random variation inherent to most natural languages is harder to synthetically model. Both corpora have been parsed into Universal Dependencies

Default	**Lemmatised**
he spent most of his childhood in sunamganj with his mother . david s. mack (born 1941) is an american businessman . he spent most of his childhood in sunamganj with his mother . david s. mack (born 1941) is an american businessman .	the episode be generally well receive . the software be sell and support only in japan . the episode be generally well receive . the software be sell and support only in japan .
Shuffled	**Corrupted**
most his with in of childhood spent sunamganj . mother his he s. american . born is david 1941) businessman an (mack most his with in of childhood spent sunamganj . mother his he s. american . born is david 1941) businessman an (mack	be generally . receive well episode the software be the sell in and support . japan only be generally . receive well episode the software be the sell in and support . japan only

Table 1: Sample sentences extracted from real corpora, with each of our modifications applied. Note that while the original and lemmatised corpora are sampled differently, the shuffled and corrupted corpora are modified variants of the former.

(UD) (Nivre et al., 2016, 2020; de Marneffe et al., 2021).

From each of the large corpora (Wikipedia and Common Crawl), we sample five corpora of 20k sentences for each language, with different random seeds, and split them into train and validation splits of 15k and 5k tokens, respectively. We employ a number of simple heuristics to filter out sentences that we suspect to be titles, or other noisy text. We generate two variants of each corpus: one that we tokenise with a BPE tokeniser, and another that retains UD-style tokenisation. The motivation behind this is to control for subwords: the absence of subword tokenisation is harder for our models to recover from, as they must be able to cluster tokens that have the same morphological affixes without explicit access to these affixes.

For our BPE segmented corpora, we use a model vocabulary of size 2048; this vocabulary is derived by training a fastBPE tokenizer on the respective training corpus. For UD-style tokenisation, we also use a vocab with 2048 unique tokens. We handle unknown tokens by replacing them with <unk> tokens; we also filter out sentences that have over 90% OOV tokens in the process of sentence selection, to avoid noise. As both our corpora are fairly noisy, we also apply a set of heuristics to eliminate corpus noise; for instance, we filter out sentences based on the number of title-cased tokens in them, to avoid scraping Wikipedia titles.

3.3 Perturbations

To adequately isolate the effects of word order and morphology, we apply three modifications to each combination of tokenisation method and corpus, giving us a total of $2 * 2 * 4 = 16$ corpora per language; with 15 languages and 5 seeds, this equates to $16 * 15 * 5 = 1200$ experiments in all.

Original Our original, unmodified corpus, presented with both UD- and BPE-based tokenisation.

Shuffled We modify our corpus by shuffling every sentence at a word level. Note that the shuffling procedure takes place before BPE segmentation, similar to Sinha et al. (2021). Ideally, given no word-order context, our masked language models should only be able to rely on morphological information, or bag-of-words distributions, in order to build a multilingual space. This also has a similar effect to removing positional embeddings from the transformer, as described in Sinha et al. (2021). Positional embeddings act as an ordering mechanism in masked language modelling; without them, a corpus is similar to our shuffled corpus.

Lemmatised We use the LEMMA Universal Dependencies field to generate our corpus, instead of the usual FORM field. The motivation here is to eliminate all morphological information; the difference between this and avoiding BPE tokenisation is that lemmatisation prevents unique word forms from having separate vocab indices.

Corrupted This corpus is both lemmatised and shuffled. Given this precondition, and UD-style tokenisation, there ought to be no information accessible to our model, beyond bag-of-word lemma statistics. We therefore expect word translation and sentence retrieval to be close to 0 in this setting.

3.4 Models and Evaluation

To evaluate our models' multilingual capabilities, we first train lower-capacity language models on each corpus. Each model is trained on the task of masked language modelling, on the concatenation of both halves (original and shifted) of a corpus. We use Dufter and Schütze (2021)'s BERT variant, which downsizes the original BERT model; we use

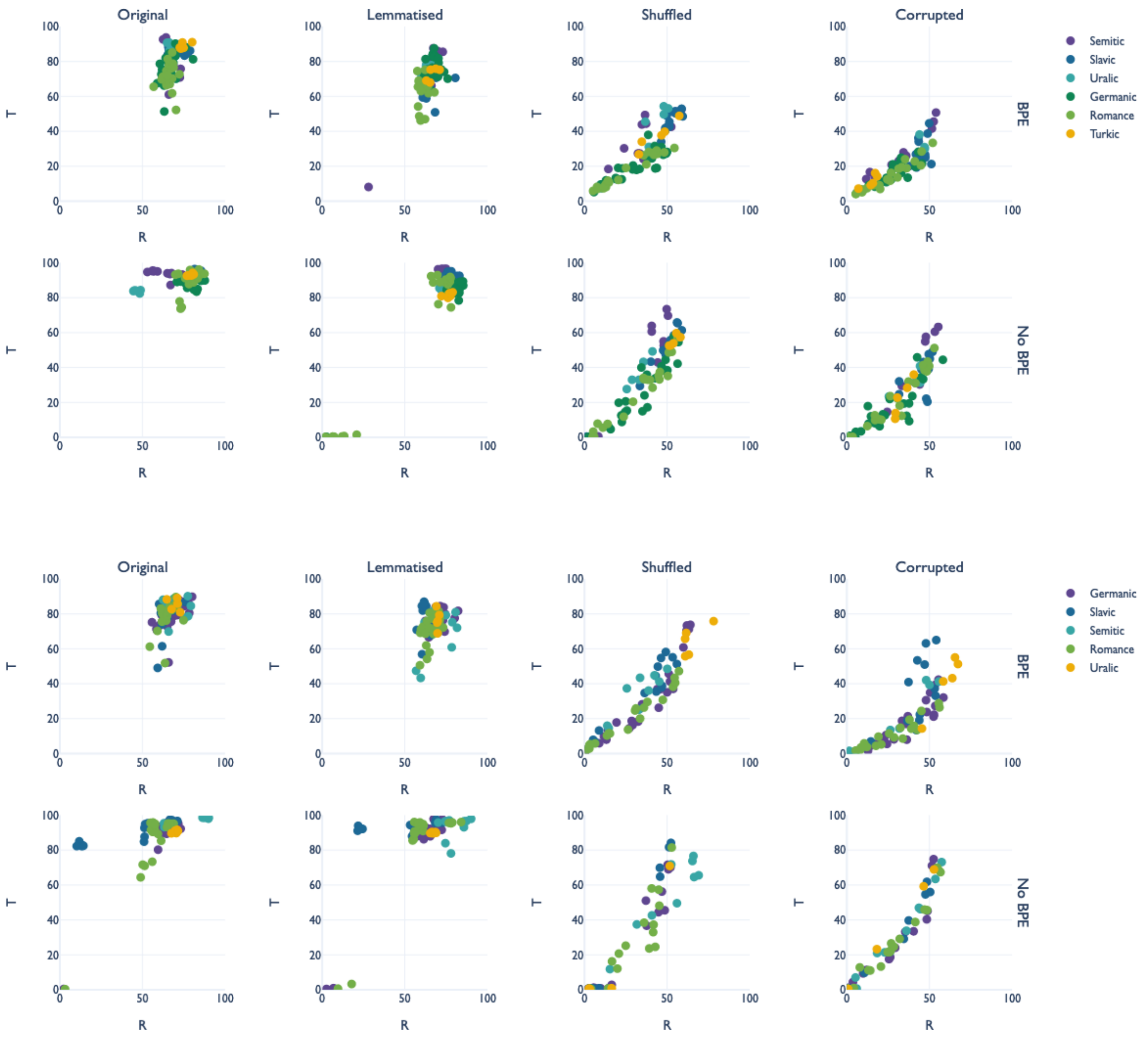


Figure 1: Results for our four perturbations, with and without BPE, with data from Common Crawl (top) and Wikipedia (bottom). Scores (sentence retrieval on the X-axis, word translation on the Y-axis) are averaged over layers 0 and 8.

single-headed, 12 layer transformer, with a head dimensionality of 64 and a feed-forward dimensionality of 256. This allows us to rapidly train a model on our corpora (in approximately 30–60 minutes per model). We set the random seed of each model to the same as the random seed used to generate the corpus we train it on; i.e. the model with seed 0, for English, is trained on the English corpus that was generated using a random seed of 0. Models are trained on V100 GPUs, each for approximately 1 hour.

Finally, we evaluate word translation and sentence retrieval scores for these models by using the deterministic gold labels, obtained by simply adding the vocab size (for translation) and by dividing the corpus into two halves and generating a sequential mapping (for retrieval). Note that this evaluation does not involve fine-tuning language models: we use the cosine similarity between either a word or a sentence and its fake parallel, for word translation and sentence retrieval resepectively. For word translation, we ensure that non-initial subwords are not included in the evaluation; while this is not ideal, none of our languages are morphologically prefixing, implying that the bulk of the semantic content is in the initial subword.

4 Results

We present results per language and experiment on Common Crawl (top) and Wikipedia (bottom) in

Figure 1. We begin by making a few general observations before moving on to study correlations with morphosyntactic and corpus factors.

'Fails' are frequent We note, first, that across most of our experiments, we have several 'fails', where our model effectively has near 0 retrieval and translation capacity. While this observation in isolation is somewhat meaningless – the model might have failed to learn effectively, either due to the random seed or due to the hyperparameters – the sheer number of experiments we run for each scenario makes these results more meaningful, when used as a comparison between training scenarios, as evidence that a certain scenario is likelier to result in a fail than another.

BPE makes word translation harder Despite controlling for non-initial subwords, using BPE tokenisation results in a drop in translation score for all our experiments. We hypothesise that this is due to common word-initial subwords being distributionally 'overloaded'; they are more likely to appear in a wider range of contexts than whole tokens are, due to the variety in consecutive subwords.

Multilingualism is robust to lemmatisation Perhaps somewhat unsurprisingly, lemmatisation does not significantly affect model scores, indicating that our model relies more on word order to build multilingual spaces. Interestingly, removing BPE segmentation results in an increase in fails on lemmatised corpora.

Bag-of-words is enough for (some) experiments Our most unexpected observation is that for both shuffling and corrupting, for both BPE and non-BPE, several experiments do appear to result in fairly successful retrieval/translation models, often with an accuracy higher than 50% on either task. This is surprising, given that a) this appears to contradict the findings of Dufter and Schütze (2021) about position embeddings being critical for multilingual spaces, and b) it implies that a simple bag-of-words model is enough to build a multilingual space. We attempt, in the following sections, to tease out what factors might enable this transfer. It is plausible that some part of this signal stems from the fact that the shuffling operation was carried out *prior* to BPE segmentation (Abdou et al., 2022); we discuss this further in Section 5.4.

5 Analysis

5.1 Clustering

In order to find potential explanations for our results, we automatically cluster our scores, using retrieval and translation scores as our cluster metrics. To determine whether either languages (given that we have five experiments per language) or language families tend to actually represent logical, meaningful clusters, we set the number of clusters to be equivalent to the number of families, and use the adjusted Rand score (Vinh et al., 2010) to measure the distance between two clusterings – clusterings based on language/family, and learnt clusterings.

We present these results in Table 2. First, clustering by language family shows little to no correlation with score-based clusters. Clusters of corpora in a single language ('language-based' clusters) are slightly clearer: while similarities are relatively low for all our BPE-based clusters, when we switch to UD tokenisation, the default and lemmatised cases begin to form more typologically relevant clusters, resembling languages. While these are by no means perfect overlaps, they are almost twice as realistic as for BPE-based tokenisation, implying that there exist language-specific features that correlate somewhat to the model's ability to form multilingual spaces. To investigate these findings in greater detail, we look for language-specific features – both corpus-specific features, and vocabulary features – and look for correlations that might explain our results.

5.2 Corpus correlations

We analyse our corpora, and measure correlations of model performance to a range of descriptive statistics, applied to the corpora that the models were trained on. For a single 'performance' metric, we follow Dufter and Schütze (2021) in defining a model's ML score as the average of its word translation and sentence retrieval scores, at layers 0 and 7. We measure correlations with:

- The number of training tokens
- The type-token ratio
- The number of one-letter types
- The number of one-letter tokens
- Average type length (in characters);
- Average token length
- Average sentence length
- Frequency of *hapax*, *dis* and *tris legomena*

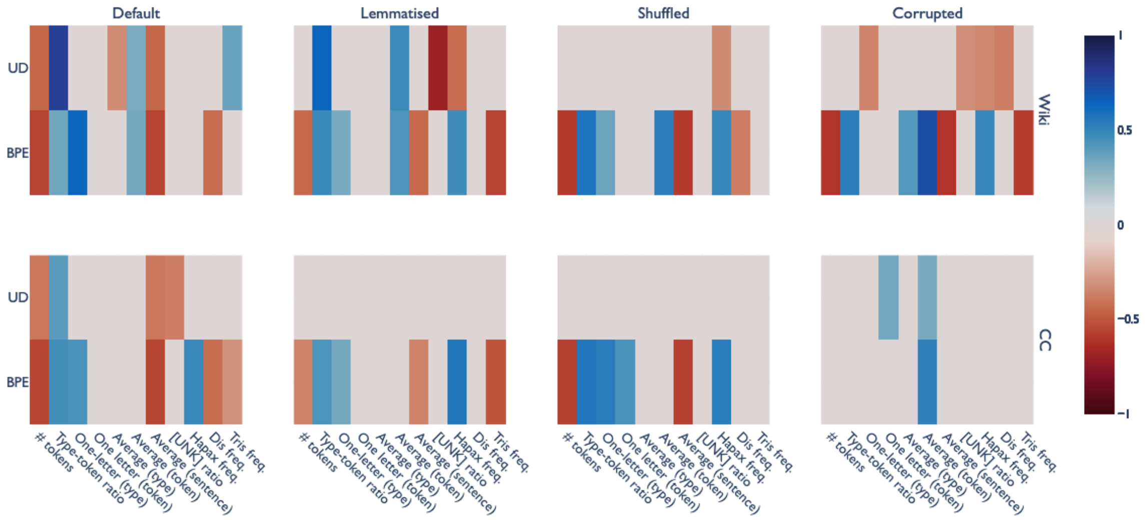


Figure 2: Spearman correlations ($\alpha = 0.001$). Greyed-out values indicate insufficient evidence.

	Language		Family	
	BPE	UD	BPE	UD
Default	0.17/0.05	0.35/0.25	0.07/0.05	0.04/0.08
Lemmatised	0.16/0.11	**0.38**/0.14	0.10/0.04	**0.14**/0.07
Shuffled	0.15/0.13	0.03/0.01	0.07/0.10	0.02/0.05
Corrupted	0.14/0.12	0.05/0.02	0.13/0.09	0.01/0.02

Table 2: Cluster similarities (adjusted Rand score) between language, or language family clusters, and k-means clustering, with a random seed of 42. Results on Wikipedia and Common Crawl are separated with a backslash.

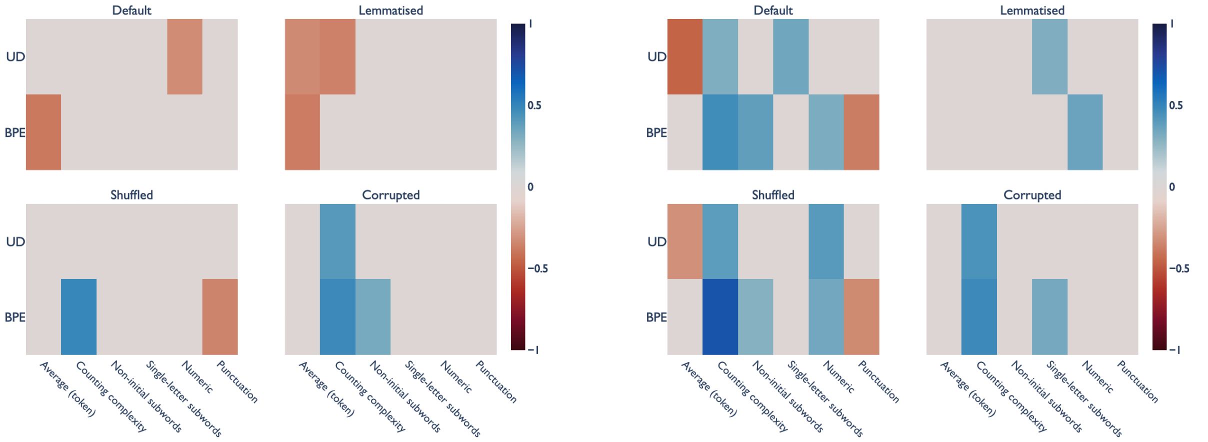
We present these statistics in Figure 2. A clear difference between doing nothing/lemmatising and shuffling/corrupting leaps out. With UD tokenisation, none of our corpus metrics correlates well with model performance, while BPE tokenisation consistently throws out a range of correlations. There is also a clear difference between Wikipedia and Common Crawl; in general, we find that correlations tend to be either weaker or less significant with Common Crawl than with Wikipedia. We hypothesise that this is due to Wikipedia being both more homogeneous and less noisy as a corpus.

Type-token ratio is a strong predictor For the default (and, to some extent, lemmatised) models, we find that type-token ratio has a strong positive correlation to ML-score (particularly retrieval), implying that lexical diversity enables better transfer. This is perhaps unsurprising – infrequent types might act as 'anchors', allowing easier transfer for their surrounding contexts. This is somewhat backed up by the disappearance of this metric in shuffled models.

Avg. token length predicts BPE performance Over our scrambled corpora, for both Wikipedia and Common Crawl,[1] it appears that average token length correlates strongly to downstream performance. The fact that this occurs for BPE tokenisation and not UD implies that this is likely a proxy for the number of BPE splits, rather than a realistic cross-linguistic measure; the more aggressive the BPE, the poorer the model. This is also somewhat backed up by the fact that the number of tokens inversely correlates to BPE performance; the shorter the average BPE split, the more the actual number of tokens in a corpus, for a given language.

Sentence length often correlates negatively This finding is consistent across all our BPE models;[1] longer sentence lengths (in tokens) imply poorer multilingual scores. This is likely at least partially related to the previous observation – the

[1] While exceptions to these observations exist, they disappear when we use a less restrictive $\alpha = 0.005$



(a) Sentence retrieval

(b) Word translation

Figure 3: Spearman correlations, with a more relaxed $\alpha = 0.01$. X-axis indicates vocabulary statistics. Y-axis indicates tokenisation method. Correlations are on Common Crawl data, with the appropriate metric averaged at layers 0 and 7.

longer the average token, the less aggressive the BPE, and the less aggressive the BPE, the shorter the average sentence.

Hapax/dis/tris ratios Results generally tend to correlate positively with the ratio of *hapax legomena* to the total number of tokens, when BPE tokenisation is used. This difference is likely due to the presence of more *morphemic* hapaxes in BPE-tokenised models: UD tokenisation is likely to result in a long tail of rarer morphological forms of rarer tokens. Curiously, this correlation, albeit weaker, is reversed for *dis* and *tris legomena*.

5.3 Vocabulary correlations

Next, we examine ML score correlations with different properties of the size 2048 UD/BPE vocabulary for each model. Note that as each model is trained with a unique corpus, each model has a unique vocabulary. Our features include:

- Average token length; for non-initial wordpieces, we do not include the length of the prefix.
- Counting complexity, using UniMorph (Kirov et al., 2020) to count the number of distinct morphological features in a given language.
- The frequency of single-letter vocab items.
- The frequency of digits in the vocab.
- The frequency of punctutation in the vocab.

We present these correlations in two heatmaps in Figures 3a and 3b. Some of our observations back up the observations in the previous section (eg. token length correlates inversely with ML score).

Counting complexity is complex Gratifyingly, the counting complexity metric (Sagot, 2013) appears to match Cotterell et al. (2018)'s observation, and is positively correlated with both retrieval and (to a larger extent) translation. Strangely, however, this correlation also appears to hold for both *corrupted* corpora; this is odd, as these corpora are lemmatised, implying the *absence* of inflectional morphology. It is plausible that this effect is still visible (albeit weakened) due to differences in the distribution of function words and stems, when compared with a language with *actual* differences in counting complexity; a language with strong case-marking, for instance, is likely to have a very different distribution of adpositions than a language without. This finding also backs up Mielke et al. (2019), who suggest that vocabulary-level measures may correlate better.

Specific tokens may act as anchors For the task of word translation, we notice that positive correlations tend to occur with the frequency of non-initial subwords, the frequency of digits, and the frequency of single-letter tokens. This effect, visible across all three categories, might indicate that these tokens act as anchors, enabling easier transfer in their contexts.

No clear patterns exist for retrieval We notice no clear factors contributing to retrieval. While the number of unused tokens does appear to correlate

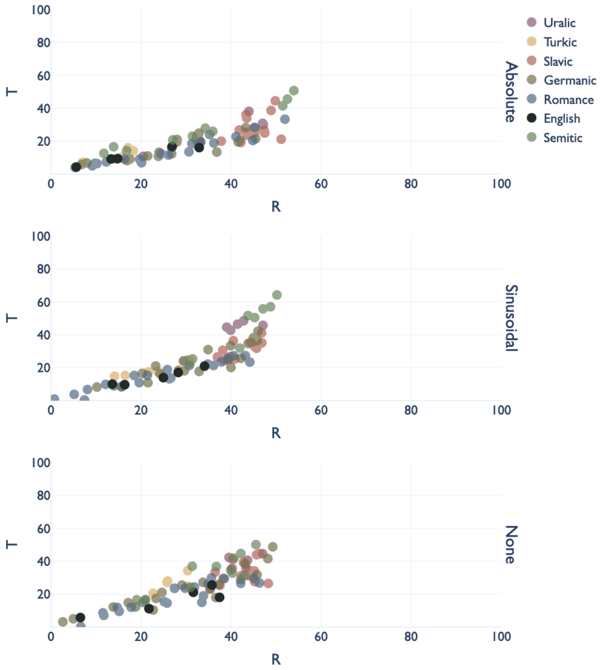


Figure 4: Retrieval/translation scores for (learnt) absolute position, (fixed) sinusoidal position and no position. English in bold black for easier comparison with Dufter and Schütze (2021).

in the lemmatised models, this is mild and is likely to be an effect of the vocab size being effectively smaller.

5.4 Ablation experiments

While somewhat tangential to our original research question, we attempted to modify the positional embedding bias in our model. Dufter and Schütze (2021) show that positional embeddings are critical to building a multilingual space; Sinha et al. (2021) show that positional embeddings are critical to building *monolingual* language models, a finding backed up in other work (Abdou et al., 2022; Papadimitriou et al., 2022), where the authors also emphasise the importance of meaningful word order. These observations are somewhat contradictory to our findings, where shuffling corpora at a token-level still allows for successful multilingual space induction.

To resolve this, we train two additional models, on a corrupted variant of Common Crawl, presented in Figure 4. The first of these has its learnt, absolute position embeddings (Devlin et al., 2019) replaced with sinusoidal embeddings, as in the original transformer paper (Vaswani et al., 2017), and the other has them removed entirely. While we would expect to see model performance drop considerably without position embeddings, this is often not the case at all; there is no real visible difference in performance across either of the tasks, implying that certain 'clues' are perhaps sufficient to build a multilingual space, even when a functional monolingual space might not exist for any of the languages.

Having said that, we note that English (annotated in black) is not one of the easier languages to build multilingual spaces for, even with absent position embeddings; as such, our English results are more similar to the results reported by Dufter and Schütze (2021).

6 Conclusion

In this work, we attempted to measure the variance in the ability of masked language models to build multilingual spaces with the underlying typology of the language. In doing so, we have shown that these models are capable of building multilingual spaces even when sentences are lemmatised and scrambled at a token level, showing that multilingualism can exist even when transformers act, functionally, like bag-of-words models. This does *not*, however, necessarily imply the ability to effectively model language (Abdou et al., 2022), but merely the ability to align two disjoint linguistic spaces.

We have also shown that, on the one hand, the ability to build a multilingual space is only weakly correlated to language (given multiple corpora) and to language family, and that, on the other hand, certain corpus-level metrics (specifically, type-token ratios and the presence of *hapax legomena*) are relatively good predictors of multilingual space quality, while others (such as the number of tokens or the average sentence length) are negatively correlated.

Our work is not without its caveats. For one, a lot of our correlating factors muddy the waters between what is an inherent property of the language itself, and what is a property of the *corpus* we use. While we use texts from the same domain in all our languages, both Wikipedia and Common Crawl are widely inconsistent across language, unless explicitly made comparable (Otero and López, 2010). Further, as discussed earlier, our scenario is not strictly realistic: first, this is a bilingual setup meant to approximate a multilingual one; second, both our languages have exactly the same structure; third, our language models are very underparameterised relative to full-scale models. It is unlikely that our observations would hold true in a real-world scenario; given, however, that our aim was to study the *inductive biases* of masked language models, using full-scale models would defeat the purpose somewhat, as the sheer volume of training data would have overridden these biases. Having said that, we present this work as an attempt to add to the often conflicting pool of papers attempting to shed some light on how language models acquire language.

Limitations

This work has several limitations, some of which we have addressed. To reiterate, in order to enable some degree of cross-linguistic diversity in this analysis, our bilingual setup is only an approximation of a true multilingual setup. Conversely, we are limited in the data we have access to: for inclusion in this study, languages had to have large and relatively noiseless dependency-parsed corpora available; as such, we are somewhat biased towards over-representing Indo-European languages.

Ethical considerations

The research presented in this work is compatible with the ACL ethics policy; the data we use is a toy subset of openly available corpora, and our models are very underparameterised, relative to the current state-of-the-art. Given the sheer number of models we train, our main experimental findings require approximately 1200 GPU hours for training, approximately equivalent to the amount of time required to train a full-scale BERT model on the same V100 GPUs.[2]

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word Order Does Matter and Shuffled Language Models Know It. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919.

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.

Ryan Cotterell, Sebastian 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.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2021. When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer. *arXiv:2110.14782 [cs]*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Philipp Dufter and Hinrich Schütze. 2021. Identifying Necessary Elements for BERT's Multilinguality. *arXiv:2005.00396 [cs]*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv:1803.11138 [cs]*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. *arXiv:1912.07840 [cs]*.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2020. UniMorph 2.0: Universal Morphology. *arXiv:1810.11101 [cs]*.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23:533–572.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Sebastian J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What Kind of Language Is Hard to Language-Model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies

[2] `https://developer.nvidia.com/blog/training-bert-with-gpus/`

v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

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.

Pablo Gamallo Otero and Isaac González López. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*, pages 21–25. Citeseer.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying arguments, bert doesn't care about word order... except when it matters. *Proceedings of the Society for Computation in Linguistics*, 5(1):203–205.

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.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages. page 11.

Benoît Sagot. 2013. Comparing complexity measures. In *Computational approaches to morphological complexity*.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Jennifer C. White and Ryan Cotterell. 2021. Examining the Inductive Bias of Neural Language Models with Artificial Languages. *arXiv:2106.01044 [cs]*.