# Multilingualism Encourages Recursion: a Transfer Study with mBERT

**Andrea Gregor de Varda**
University of Milano-Bicocca
a.devarda@campus.unimib.it

**Roberto Zamparelli**
CIMeC – University of Trento
roberto.zamparelli@unitn.it

## Abstract

The present work constitutes an attempt to investigate the relational structures learnt by mBERT, a multilingual transformer-based network, with respect to different cross-linguistic regularities proposed in the fields of theoretical and quantitative linguistics. We pursued this objective by relying on a zero-shot transfer experiment, evaluating the model's ability to generalize its native task to artificial languages that could either respect or violate some proposed language universal, and comparing its performance to the output of BERT, a monolingual model with an identical configuration. We created four artificial corpora through a Probabilistic Context-Free Grammar by manipulating the distribution of tokens and the structure of their dependency relations. We showed that while both models were favoured by a Zipfian distribution of the tokens and by the presence of head-dependency type structures, the multilingual transformer network exhibited a stronger reliance on hierarchical cues compared to its monolingual counterpart.

## 1 Introduction

Massively Multilingual Models (MMMs) are neural networks that can perform a NLP task in multiple languages, relying on a shared set of parameters. At the time of writing, the state-of-the-art performance of MMMs is achieved by transformer-based models such as multilingual BERT (mBERT, Devlin et al., 2019), XLM (Conneau and Lample, 2019), and XLM-R (Conneau et al., 2020a). They are usually derived from monolingual language models, trained simultaneously on multilingual text in up 104 languages without major architectural changes nor any reliance on explicit cross-lingual signal. The practical need for MMMs in NLP is undisputed: they drastically reduce resource and maintenance requirements with respect to multiple monolingual models, and benefit in particular low- and mid-resource languages (Dufter and Schütze, 2020). MMMs reach impressive performance levels in zero-shot cross-lingual transfer, enabling the fine-tuning of a model on supervised data in a set of $N$ languages $\{L_i\}_{i=1\dots N}$ and its application to a different language $L_{N+1}$, with no additional training[1]. Zero-shot cross-lingual transfer has been shown to be effective across a variety of tasks and languages (Dufter and Schütze, 2020; Liu et al., 2020; Pires et al., 2019; Wu and Dredze, 2019; see Doddapaneni et al., 2021 for a review), and, although performance levels tend to be higher for typologically similar languages, it yields surprising results in languages written in different scripts (Pires et al., 2019) and with little (Karthikeyan et al., 2020) or no (Conneau et al., 2020b; Wang et al., 2019) vocabulary overlap. The distribution of resources available for NLP researchers in the world's languages is extremely skewed, with only a small subset of them being represented in the evolving language technologies (Joshi et al., 2020). MMMs constitute an attempt to mitigate the effects of this uneven allocation of resources by leveraging the knowledge that can be shared across languages.

Besides the obvious practical advantages that MMMs can bring to the NLP community, the nature of the cross-linguistic information extracted by these models is of high theoretical interest from a linguistic standpoint, and can contribute to the domain of artificial intelligence research in relevant subfields such as representation learning and interpretability. A modest but growing body of findings suggests that the structure of the representation space that MMMs exploit is multilingual in nature (Pires et al., 2019; Wu and Dredze, 2019; Hu et al., 2020; Liu et al., 2020; although see Dhar and Bisazza, 2021 for opposite conclusions). For instance, syntactic trees can be retrieved from mBERT's intermediate representational subspaces, with these subspaces being approximately shared

[1] $\{L_i\}_{i=1\dots N}$ and $L_{N+1}$ are typically resource-rich and resource-poor languages, respectively.



*Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP (SIGTYP 2022)*, pages 1 - 10
July 14, 2022 ©2022 Association for Computational Linguistics

across languages (Liu et al., 2020). If MMMs learn universal patterns which generalize across languages, the structure of the representations they induce could inform us of the presence of latent regularities in different language spaces. Furthermore, the benefits of the study of the MMMs' behaviour extend to the domain of representation learning, a subfield of AI research focusing on the development of computational representations and the analysis of their properties (Bengio et al., 2013). While the present study will not analyze the internal states of the networks, it will be possible to draw conclusions on the generality of their learned representations through non-parametric probing, by directly examining their behaviour in response to non-linguistic input. More precisely, we will compare the suitability of the representational formats induced by mono- and multilingual models with respect to different properties that are desirable from a linguistic perspective.

The present work aims to analyze the generalizations that BERT and mBERT induced from natural language data in a set of transfer learning experiments. The use of transfer learning methods to shed light on the relational structures learned by neural networks has been recently adopted for monolingual models (Papadimitriou and Jurafsky, 2020). Here, we extended the transfer approach to a multilingual setting, and compare the performance of mBERT and its monolingual counterpart in generalizing their native task (i.e. masked language modelling) to artificial languages that display different degrees of structural similarity with natural languages. We wish to highlight three main differences between our paradigm and the methodologies Papadimitriou and Jurafsky proposed.

1. **Cross-lingualism.** The most significant contribution of our study consists of the transposition of Papadimitriou and Jurafsky's paradigm to a multilingual setting.

2. **Direction of the transfer.** Papadimitriou and Jurafsky (2020) evaluated the performances of several LSTM models trained on non-linguistic data and transferred zero-shot to a natural language corpus in Spanish. We invert the direction of the transfer, testing the pre-trained multilingual model on artificial corpora derived from formal grammars. This choice is desirable for three reasons: first, it allows to test the model once for each experimental condition, and not in different languages. Second, it frees us from the need to train several models – one for each artificial corpus – since we can leverage one single multilingual pre-training. Third, it lets us draw conclusions on the structural generalizations which have been directly induced from natural language data. In the other direction, the models could have extracted helpful generalizations from the artificial dataset which might still not have been visible when looking at natural language alone.

3. **Neural architecture.** Papadimitriou and Jurafsky (2020) have employed LSTM models for all their experiments; however, transformers are gaining increasing popularity in NLP research and applications, and achieve state-of-the-art results across different downstream tasks. Most MMMs are built as transformer architectures, and mBERT is an instance of this class. Hence, our study can be informative also in terms of model comparison. Note that moving to a bidirectional transformer requires a different approach to calculating sequence-level performance, one which eliminates the randomness from the masking process. Our approach, which we name *iterative token-level cloze task* (ITCT) is detailed in Section 2.

Our experiment focused on the Zipf's law and hierarchy, which have been considered as universal linguistic features (Zipf, 1935; Chomsky, 1957). We evaluated the transfer performances of the two transformers in four corpora, characterized by increasing statistical and structural consistency with natural languages. The models were tested on (a) a RANDOM corpus, composed by sequences of tokens sampled from a uniform distribution, (b) a ZIPFIAN corpus, where the tokens were extracted from a Zipfian distribution, (c) a FLAT BRACKETS corpus, composed of sequences of matching parentheses with crossed dependencies, and (d) a NESTED BRACKETS corpus, consisting of paired symbols nested hierarchically. To anticipate the results, we found that both models showed higher performance scores in the ZIPFIAN compared to the RANDOM condition, and in the FLAT BRACKETS as opposed to the ZIPFIAN corpus, while only the multilingual model showed a significant performance advantage in the comparison between the FLAT BRACKETS and the NESTED BRACKETS

corpora. We conclude that while mathematical regularities and pairwise head-dependent relationships are detected across model types, the multilingual input favours the reliance on structural cues, and specifically on balanced constituent structures, a hallmark of theoretical linguistic formalisms.

## 2 Methods

As a testing procedure, we froze all mBERT's weights setting it in evaluation mode, and assessed its structural knowledge by studying its predictive ability. To do so, we employed a non-parametric evaluation procedure, which we named *iterative token-level cloze task* (ITCT). The ITCT consists of an adaptation of mBERT's native functionality, i.e. masked language modelling (MLM). The main difference consists in the fact that while in MLM the model has to predict the tokens corresponding to the masks applied to a randomly selected subpart of the input (15% of the tokens in the sentence), in the ITCT all the tokens are masked iteratively. This mitigates the aleatory dimension in the selection of the tokens that are masked, and provides an index of the predictability of a sequence, where each token has to be predicted by the model given the whole remaining context. After freezing its weights[2], at the first timestep $t_0$ the model is presented with the input sequence where the first token is masked, i.e. substituted with a mask token, and two special characters – [CLS] and [SEP] – are appended to the beginning and the end of the sequence, to mark the sequence boundaries. The model then predicts the original token relying upon the right context, and the hidden vector corresponding to the masked token is passed through a softmax over the vocabulary, in order to assign it a probability. At $t_1$, the mask is moved from the first to the second token, and now mBERT's prediction is conditioned by both the right and the left contexts. The process is repeated until the end of the sequence, with the number of timesteps $N$ being equal to the length of the tokenized input (see Figure 1). The mean probability assigned to the masked tokens across all the timesteps is taken as an index of the overall predictability of the sequence. Note that a proper sequence probability metric would require a multiplicative chain rule of the kind that is applicable for auto-regressive models but not for masked language models. The notion of average probability does not correspond to any well-defined notion in probability theory, but it serves the purpose of comparing different structural configurations in the context of our study[3]. The predictions of mBERT were compared with the ones produced by its monolingual counterpart; this comparison provides a crucial element for distinguishing which generalizations are driven by the multilingual input, and which can be extracted by the same model from monolingual data.

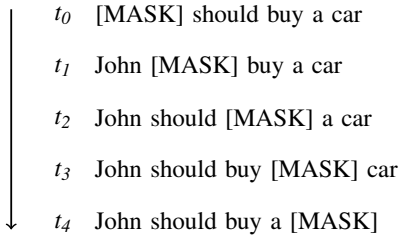


Figure 1: Unfolding of the iterative token-level cloze task for every timestep $t$ in a sample sentence.

## 3 Data

The corpora on which our analyses were performed were created in a way such that the sequence length varied within each condition, but was identical across all conditions, both in terms of the number of sequences and of the number of tokens within each sequence; they all shared a 50,000 three-letter tokens vocabulary. These design choices were made in order to license pairwise comparisons at the sequence level, so that the difference in probability assigned by the models to a given item of a corpus could be compared with the probability assigned to the corresponding item in the other datasets. This approach allowed us to rule out the effects of intervening variables such as vocabulary and length, so that the differences in the models' predictions could be driven only by structural differences between the corpora. The models were tested on 1,000 sequences in each corpus; within each condition, the mean sequence length was 9.90, with a standard deviation equal to 14.74.

### 3.1 Nested brackets

A NESTED BRACKETS corpus consisting of sequences of nested matching symbols was created

[2]Freezing the model's weights is a necessary condition for this approach, since if this procedure was implemented during training, the model trying to predict the target token at $t_1$ would have already seen it at $t_0$; at the end of the sequence, the prediction would be highly facilitated from having seen *N-1* times the target token in the same context.

[3]We thank Reviewer pYcV for bringing this issue to our attention.

to test the transfer performance of mBERT on hierarchical structures. The corpus was built from a vocabulary of 50,000 three-letter tokens, obtained from random combinations of Latin characters. The tokens were assembled into nested structures through the application of probabilistic rules, defined by a Probabilistic Context-Free Grammar (PCFG). The grammar was composed by a set of recursive rules of the form in (1):

(1) $S \rightarrow tok_i\ S\ tok_i\ S$ [P1]

Where S denotes the start symbol, $tok_i$ a given terminal symbol sampled from the vocabulary, and P1 the probability assigned to the application of the rule. Rules of this form are said to be recursive since the same non-terminal symbol S appears on both sides of the formula, which enables it to be reapplied to its own output. The rule in (1) allows for both right and central recursion, since the non-terminal symbol S is rewritten into itself both within a pair of terminal symbols and in the rightmost part of the formula. The probabilities in P1 followed a Zipfian distribution, so that the terminal symbols were distributed accordingly in the corpus; their distribution summed up to 0.4. This set of rules was complemented by the rule in (2), where the start symbol was rewritten into the empty string ε. The probability assigned to this rule was higher than the sum of all the previous rules, in order to contain the growth of the tree depth. Empty sequences were removed from the corpus.

(2) $S \rightarrow \varepsilon$ [0.6]

The most prominent feature of the sequences generated by this grammar is that the pairwise dependency arcs instantiated between tokens never cross (see Figure 2). In other words, the pairing between tokens can only be nested hierarchically within the overlying dependency relations. This condition is equivalent with respect to this property to the nested parentheses corpus created by Papadimitriou and Jurafsky (2020), although in their work they created the structured sequences with a stack-based grammar, designed to either open a new bracket or close the last one that had been opened at each timestep.

### 3.2 Flat brackets

A FLAT BRACKETS corpus was created in order to isolate the effects of non-nested dependency pairing from the presence of hierarchical structures

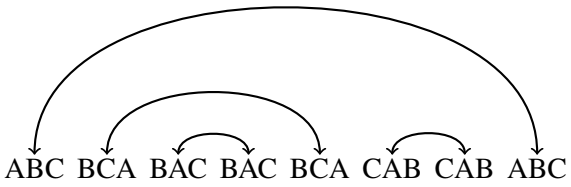


Figure 2: Example of a NESTED BRACKETS sequence.

in the transfer performances. The corpus was derived by randomly shuffling the tokens of each item of the NESTED BRACKETS corpus, a process that creates structures where the dependencies do not necessarily nest, and the pairing arcs instantiated within an entry may cross (see Figure 3). Differently from Papadimitriou and Jurafsky (2020), who created a novel corpus for this condition without any reference to the hierarchical one, we adopted a procedure that kept constant the length of the sequences, and the identity of the tokens within them. We maintain that our methodology licences more meaningful comparisons between the two corpora, since the only difference between them is the hierarchical property of recursive nesting.

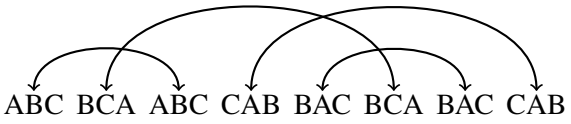


Figure 3: Example of a FLAT BRACKETS sequence.

### 3.3 Zipf's corpus

The ZIPF'S CORPUS was created in order to evaluate whether BERT and mBERT's performances were affected by the mathematical distribution of the token frequencies in the corpus. The sequences were constructed so that their length had to coincide with the one of the corresponding entry in the previous corpora. For each item in the NESTED BRACKETS corpus, we sampled a number of tokens coinciding with its length from a Zipf's distribution, and conjoined them to form a sequence of tokens. In the creation of this corpus no dependency relation was explicitly encoded. We remark that this corpus is similar to the FLAT BRACKETS corpus, with the only difference that the tokens are not repeated twice within each sequence, and so

| | | | | Model | | | |
|---|---|---|---|---|---|---|---|
| | | | | BERT | | mBERT | |
| Corpus | Zipf | Pairing | Nesting | Mean | SD | Mean | SD |
| Random corpus | | | | 0.0121 | 0.0137 | 0.0094 | 0.0135 |
| Zipf's corpus | ✓ | | | 0.0253 | 0.0457 | 0.0250 | 0.0525 |
| Flat brackets | ✓ | ✓ | | **0.6784** | 0.1780 | 0.6353 | 0.1558 |
| Nested brackets | ✓ | ✓ | ✓ | 0.6576 | 0.1677 | **0.6417** | 0.1536 |

Table 1: Featural summary of the structural and mathematical properties of the four corpora, and descriptive statistics of the results of the transfer. The best performances for each model are highlighted in bold.

there are no structural correspondences between tokens.

### 3.4 Random corpus

In order to define a baseline for the evaluation of the networks' predictions, we constructed a RANDOM CORPUS where the tokens composing the sequences were sampled from a uniform distribution. As for the ZIPF'S CORPUS, the length of each sequence matched the length of the corresponding entry in the other corpora.

## 4 Models and experimental setup

All our experiments were performed employing BERT's native masked language modelling component. The configuration of the model was left unaltered with respect to Devlin et al.'s (2019) release. In particular, we relied on the monolingual and multilingual models derived from $\text{BERT}_{\text{BASE}}$, which is composed of 12 layers, 12 self-attention heads, and a hidden size of 768; the overall network comprises 110M parameters. The networks did not undergo any fine-tuning nor adaptation process, as they were employed as out-of-the-box masked language models. As mentioned above, BERT and mBERT only differ in their vocabulary and the weights learned during training, sharing an identical configuration both in terms of architectural choices and learning objectives. While BERT was pre-trained on 800M words of the monolingual BooksCorpus (Zhu et al., 2015) and 2,500 words of the English Wikipedia, mBERT was trained on the entire Wikipedia dump of 104 languages. The two models rely on different tokenizers, each comprising a separate WordPiece vocabulary (Wu et al., 2016). The different tokenizations of the input sequences do not allow us to directly compare the raw probabilities assigned by the two models to a given sequence; for this reason, we will not consider the absolute item-wise difference in the models' predictive performance, but rather the pattern of results between the experimental conditions. This comparative approach is also needed in light of the models' vocabulary. If we simply compared probabilities across models and conditions, our results might be biased by the fact that some of the three-letter tokens might be assigned different probabilities depending on whether they form English meaning-bearing vocabulary items (e.g. "for") or not (e.g. "zyi"). However, since we only compare conditions within models, and the two conditions of main interest (i.e. FLAT and NESTED BRACKETS) are composed by the same tokens in different arrangements, our contrasts are robust with respect to this possible confound.

## 5 Results

Table 1 reports the mean and the standard deviation of the average probabilities assigned by BERT and mBERT to the token sequences in the four corpora considered in the study, along with a schematic summary of the structural features characterizing each corpus. In line with our expectations, both models assigned on average higher probabilities to the correct tokens in the ZIPF'S CORPUS than in the RANDOM CORPUS, despite a low absolute difference in the scores (0.0132 for the monolingual and 0.0156 for the multilingual model). The substantial increase in performance was obtained with the transition from the ZIPF'S CORPUS to the FLAT BRACKETS corpus, with an average improvement of 0.6531 for BERT and 0.6103 for mBERT in the metric. Interestingly, the two networks started diverging in their behaviour with the subsequent step of the structural hierarchy. While the target tokens of the recursive structures in the NESTED BRACKETS corpus were associated with higher probabilities by the multilingual model, monolingual BERT

| | | Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | BERT | | | mBERT | | |
| Corpus 1 | Corpus 2 | $t$ | $p$ | $d$ | $t$ | $p$ | $d$ |
| Random corpus | Zipf's corpus | 8.9740 | $\ll .001$ | 0.3918 | 9.2296 | $\ll .001$ | 0.4069 |
| Zipf's corpus | Flat brackets | 116.9351 | $\ll .001$ | 5.0255 | 117.4448 | $\ll .001$ | 5.2476 |
| Flat brackets | Nested brackets | -12.7494 | $\ll .001$ | -0.1205 | 3.2662 | 0.0011 | 0.0415 |

Table 2: Pairwise comparisons of the results of the transfer to the four corpora. The comparisons are made exclusively between the corpora that are adjacent in the hierarchy of structuredness. The reported $p$-values are uncorrected, but all the contrasts remain significant when a Bonferroni correction is applied to the $\alpha$ threshold (0.05/3 = 0.0166).

showed no facilitation induced by the presence of nested dependencies. On the contrary, the best performing condition for monolingual BERT was the transfer to the FLAT BRACKETS corpus. While mBERT's results reflected a positive association with the hierarchy of levels of structure that characterizes our four corpora, its monolingual counterpart showed an inverted trend in the last two conditions.

We tested the statistical significance of this dissociation through a set of paired samples $t$-tests between the mean probability assigned by the models to the sequences of two corpora. We performed the tests only on the pairs of corpora that were adjacent in the structural hierarchy; this choice led us to three comparisons which we summarized in Table 2. The first two columns of the table specify the two conditions being contrasted; the following three columns report the $t$ statistic, the associated $p$-value, and Cohen's $d$ as a measure of the effect size for BERT; the last three columns indicate the same statistical indexes for mBERT. As can be evinced by the table, the first two contrasts (RANDOM CORPUS-ZIPF'S CORPUS and ZIPF'S CORPUS-FLAT BRACKETS) are highly significant for both models, with an increment in performance attested by the positive sign of the $t$ and $d$ statistics. The effect size associated with the comparisons is modest in the first case and extremely high in the second, with no considerable differences between the models. Nonetheless, as shown in the third row of the table, the pairwise contrast between the FLAT BRACKETS corpus and the NESTED BRACKETS corpus supports the observation of a dissociation between the results of BERT and mBERT. Indeed, while for both models the performance in the NESTED BRACKETS and the FLAT BRACKETS corpus is significantly different, the direction of such difference is the opposite, as shown by the sign of the $t$ statistic and the Cohen's $d$. While the effect sizes associated with such contrasts are negligible, both dissociations are statistically significant.

## 6 Discussion

In discussing the present findings, we begin by focusing on the commonalities in the networks' output, and conclude by commenting on the dissociation in their results on the two corpora with token pairing. First, both models showed a preference for sequences where the mathematical distribution of the tokens resembled their empirical distribution in natural languages. We believe that the higher average predictability that characterized the ZIPF'S CORPUS when compared with the RANDOM CORPUS reflects a tendency of the networks to expect a non-uniform distribution of the tokens in input that is coherent with the data on which the pre-training had been performed. Then, we maintain that the substantial gain in performance on the FLAT BRACKET corpus is to be attributed to the paired correspondences between tokens, which in turn might mimic head-dependency type structures in natural language corpora. Arguably, the most surprising result that we obtained is the dissociation between BERT and mBERT's results in the transfer to the two corpora characterized by token pairing. While the multilingual model was facilitated in its native task by the presence of nested structures – although with a minimal effect size –, the same improvement in performance is not found in its monolingual counterpart. On the contrary, the strongest transfer performance is achieved by BERT in the FLAT BRACKETS corpus. These results suggest that the presence of multiple languages in the input during pre-training leads the models to rely on more structured grammatical abstractions. Obviously, this finding does not imply that mBERT does not capture the paired relationships with crossing

arcs; the biggest progress is undoubtedly obtained when these simpler one-to-one correspondences are included in the input. Nonetheless, it seems that when more complex structures are instantiated between the tokens in the sequences, the multilingual model is able to capture these configurational regularities, and exploit them in order to make stronger predictions regarding the masked input. This difference should be attributed to the nature of the input that the networks had been presented with during pre-training, since under every other aspect except vocabulary and training set size (e.g. architectures, training regimes, objective functions) the models were identical.

### 6.1 Follow-up analyses

While the results of mBERT can be given a straightforward interpretation, the fact that the monolingual model showed a preference for the non-recursive corpus needs to be explained. Indeed, if its results had been exclusively driven by the absence of a hierarchical bias, we would have expected no significant difference between its performance scores in the transfer on the two corpora characterized by token pairing. What we found was instead a clear, significant preference for the FLAT BRACKETS corpus, that cannot be explained in terms of sequence length or identity of the tokens, since all these low-level factors were maintained unaltered in the two corpora (see Section 3). Without any clear *a priori* expectation on the factors that might have driven this effect, we inspected the ninety-ninth percentile of the sequences that showed the highest difference between the probability assigned by BERT to the flat and the nested structures. In other words, we computed the difference of the ITCT scores assigned by the model to the nested and the corresponding flat sequences (henceforth $\Delta$ score), and selected the first 10 sequences after ranking them in descending order. We remind the reader that the sequences in the two conditions comprised the same tokens, assembled hierarchically and projectively in the NESTED SEQUENCES corpus, and randomly shuffled in the FLAT SEQUENCES corpus. The most salient property of the items with the highest $\Delta$ score that we derived was that 80% of them were four-token sequences, with the form *abab* in the FLAT BRACKETS condition and either *aabb* or *abba* in the NESTED BRACKETS condition. We reasoned that a property that distinguishes these three classes of sequences is the presence of identical adjacent tokens, which characterizes both forms of the FLAT BRACKETS corpus, but not the *abab* sequence in the NESTED BRACKETS condition. We speculate that the models – and in particular the monolingual one – might not expect the same token to appear in two immediately adjacent positions within a given sentence, and that this tendency might have driven the higher performance scores of the monolingual model on the FLAT BRACKETS corpus. For this hypothesis to have a plausible theoretical ground, we needed to assess whether the contiguous repetition of identical tokens is indeed a rare phenomenon in natural language. While it is well known that lexical repetition is common at the discourse level – words that have entered the discourse have a higher reuse probability than lexical frequency (Heller et al., 2010) –, the probability of reoccurrence of the same token in two contiguous positions has not been assessed through corpus studies. To do so, we counted the number of such instances of repetition in four corpora of 1M tokens, derived from Wikipedia dumps in three languages (English, Chinese, and Finnish) belonging to three different language families (Indoeuropean, Sino-Tibetan, and Uralic). We tokenized each corpus with mBERT's WordPiece tokenizer, removed punctuation and unknown characters, and counted the number of occurrences of a given token at index $i$ and $i$+1. Perhaps surprisingly, we found that in two out of three corpora the probability of having the same token $k$ at index $i$ and $i$+1 was lower than chance (i.e., lower than the probability of having a random token sampled from the corpus' vocabulary; see Table 3). These results support the idea that the juxtaposition of identical tokens is indeed an unusual occurrence in natural language.

Once we verified the low frequency of identical adjacent tokens in three natural languages, we needed to evaluate whether subsequences of this kind had an actual effect on the models' predictions. In order to test this hypothesis, we ran four linear regression models (one for each considered corpus $\times$ model combination) with the score assigned to each sequence as a dependent variable, and the amount of identical adjacent tokens as a predictor. More precisely, we employed as independent variable the ratio of token reoccurrences over the total amount of token pairs in the sequence.[4]

[4]We chose not to employ the raw amount of adjacent pairs as a regressor in order to mitigate the effects of sequence

| Language | Family | Repetitions | Vocabulary | P repetitions | P random |
|---|---|---|---|---|---|
| English | Indoeuropean | 19 | 24,443 | $1.9^{-5}$ | $\mathbf{4.1^{-5}}$ |
| Chinese | Sino-Tibetan | 1,891 | 9,604 | $\mathbf{1.8^{-3}}$ | $1.0^{-4}$ |
| Finnish | Uralic | 21 | 17,546 | $2.1^{-5}$ | $\mathbf{5.7^{-5}}$ |

Table 3: Probability of adjacent tokens repetitions and chance level sampling. The probability of the repetitions was computed by dividing the raw count of the repetitions by the number of tokens in the corpus (1M), while the probability of sampling a random token from the vocabulary was obtained dividing 1 by the word types in the corpus.

In line with our expectations, we found a negative, highly significant effect of the ratio of reoccurrences on BERT's scores in the FLAT BRACKETS condition (B = -0.1958, $t$ = -16.098, $p \ll 0.001$, $R^2$ = 0.206). A similar pattern of results was found in the NESTED condition (B = -0.1667, $t$ = -7.949, $p \ll 0.001$), although the model explained a limited amount of variance ($R^2$ = 0.06). Crucially, we found no significant effect of the adjacent pairs ratio on the results of the multilingual model neither in the FLAT BRACKETS (B = -0.0214, $t$ = 1.795, $p$ = 0.073, $R^2$ = 0.003) nor in the NESTED BRACKETS condition (B = 0.0312, $t$ = 1.577, $p$ = 0.115, $R^2$ = 0.002). These results corroborate our previous suspicion concerning the different performance patterns of the two transformers models. More precisely, they suggest that for the monolingual model, local heuristics relying on linear order prevail, hiding to the model the structural cues that are instantiated in the nested brackets corpus. Conversely, the cross-lingual model seems to be able to rely on more abstract structural features and to exploit them in its predictive behaviour, while not being influenced by shallow linguistic factors such as token repetition.

## 7 Conclusion

The mathematical regularities of the pre-training input seemed to have been absorbed to the same extent by the monolingual and the multilingual transformer models, since the ZIPF'S CORPUS exerted a similar facilitation effect with respect to the RANDOM CORPUS across model types. Furthermore, token pairing appears to have elicited a much stronger advantage in the predictive task we employed in our study, suggesting that regardless of the mono- or multilingual nature of the input the pre-training procedure had induced a strong structural transfer towards non-hierarchical and non-projective structures. We agree with Papadimitriou and Jurafsky's (2020) conclusion that this facilitation emphasizes the importance of pairing, head-dependency type structures in the linguistic embeddings of neural language models. In addition, our results extend the previous findings to the generalizations employed by pre-trained transformer models, and validate the methodological choice of inverting the direction of the transfer. More importantly, the difference in BERT and mBERT's performances when recursion is implemented in the data suggests that the high surface inconsistency of the input the multilingual model is exposed to during pre-training promotes stronger structural generalizations. This finding directly answers a question raised in the Introduction, concerning the relevance of this study with respect to the domain of representation learning. We noted that the comparison between mono- and multilingual models on linguistic and non-linguistic tasks could have allowed us to draw conclusions on the aptness of their induced representational formats with respect to different properties that are desirable from a linguistic perspective. While the behaviour of the monolingual model seems to be influenced by other non-structural congruences with the pre-training input (such as the presence of adjacent paired tokens), this experiment suggests that multilingual representations are more deeply aligned with the structures posited in theoretical linguistics, showing a hierarchical bias when transferred zero-shot to non-linguistic input.

## 8 Limitations and further directions

While our results provide empirical evidence for a higher structural awareness in mBERT as opposed to BERT, the generalizability of our findings to natural language is yet to be assessed. In the present paper we employed artificial languages in order to maximize the experimental control over the input; we leave to future research an evaluation of the structural biases operating in mono- and multilingual models in a more naturalistic setting.

length.

## References


Y. Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828.

Noam Chomsky. 1957. *Syntactic structures*. De Gruyter Mouton.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Prajit Dhar and Arianna Bisazza. 2021. Understanding cross-lingual syntactic transfer in multilingual recurrent neural networks. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 74–85, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2021. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Jordana Heller, J Pierrehumbert, and David N Rapp. 2010. Predicting words beyond the syntactic horizon: Word recurrence distributions modulate on-line long-distance lexical predictability. *Architectures and Mechanisms for Language Processing (AMLaP)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

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

K Karthikeyan, Wang Zihan, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2020. Exploring fine-tuning techniques for pre-trained cross-lingual models via continual learning. *arXiv preprint arXiv:2004.14218*.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning music helps you read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839.

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.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

George Kingsley Zipf. 1935. *The psycho-biology of language: An introduction to dynamic philology*, volume 21. Psychology Press.