# Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models

**Isabel Papadimitriou*** and **Kezia Lopez*** and **Dan Jurafsky**
Computer Science Department
Stanford University
{isabelvp,keziakl,jurafsky}@stanford.edu

## Abstract

While multilingual language models can improve NLP performance on low-resource languages by leveraging higher-resource languages, they also reduce average performance on all languages (the 'curse of multilinguality'). Here we show another problem with multilingual models: grammatical structures in higher-resource languages bleed into lower-resource languages, a phenomenon we call *grammatical structure bias*. We show this bias via a novel method for comparing the fluency of multilingual models to the fluency of monolingual Spanish and Greek models: testing their preference for two carefully-chosen variable grammatical structures (optional pronoun-drop in Spanish and optional Subject-Verb ordering in Greek). We find that multilingual BERT is biased toward the English-like setting (explicit pronouns and Subject-Verb-Object ordering) as compared to our monolingual control. With our case studies, we hope to bring to light the fine-grained ways in which dominant languages can affect and bias multilingual performance, and encourage more linguistically-aware fluency evaluation.

## 1 Introduction

Multilingual language models share a single set of parameters between many languages, opening new pathways for multilingual and low-resource NLP. However, not all training languages have an equal amount, or a comparable quality of training data in these models. In this paper, we investigate if the hegemonic status of English influences other languages in multilingual language models. We propose a novel method for evaluation, whereby we ask if model predictions for lower-resource languages exhibit structural features of English. This is similar to asking if the model has learned some languages with an "English accent", or an English *grammatical structure bias*.

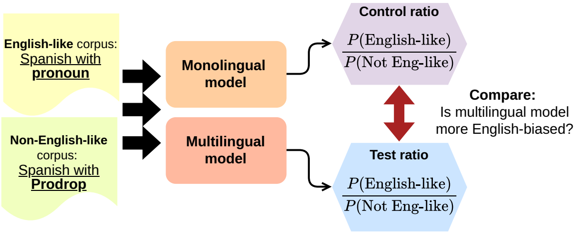
English-like corpus:
Spanish with pronoun
Non-English-like corpus:
Spanish with Prodrop
Monolingual model
Multilingual model
Control ratio
$\frac{P(\text{English-like})}{P(\text{Not Eng-like})}$
Compare:
Is multilingual model more English-biased?
Test ratio
$\frac{P(\text{English-like})}{P(\text{Not Eng-like})}$

Figure 1: Our method for evaluating English structural bias in multilingual models. We compare monolingual and multilingual model predictions on two sets of natural sentences in the target language: one which is structurally parallel to English, and one which is not.

We demonstrate this bias effect in Spanish and Greek, comparing the monolingual models BETO (Cañete et al., 2020) and GreekBERT (Koutsikakis et al., 2020) to multilingual BERT (mBERT), where English is the most frequent language in the training data. We show that *mBERT prefers English-like sentence structure in Spanish and Greek* compared to the monolingual models. Our case studies focus on Spanish pronoun drop (prodrop) and Greek subject-verb order, two structural grammatical features. We show that multilingual BERT is structurally biased towards explicit pronouns rather than pro-drop in Spanish, and putting the subject before the verb in Greek: the structural forms parallel to English.

While the effect we showcase here will likely not impact performance on the downstream classification tasks used to evaluate multilingual models (Hu et al., 2020), it demonstrates the type of fluency that can be lost when co-training with a language dominant in the training data — something that current evaluation methods miss. In fact, though we choose two clear-cut syntactic features to investigate, there are many less-measurable features that make language production fluent: subtle lexical choice, grammatical choice, and discourse expression, among many others. With this paper, beyond showing a trend for two specific grammatical fea-

[*] Equal contribution

| $S_{\text{parallel}}$: **English-like structure** | $S_{\text{different}}$: **Different structure** |
|---|---|
| **Spanish explicit pronoun** (pron in orange, verb in blue) | **Spanish prodrop** (verb in blue) |
| Yo volveré para averiguarlo | Jamás dan soluciones y siempre [. . . ] |
| I will return to figure it out | [They] Never give solutions and always [. . . ] |
| El 2004 , ella hizo doblaje a el Inglés [. . . ] | Jugó de centrocampista en el Real Zaragoza |
| In 2004, she did dubbing to English [. . . ] | [He/She/You] Played as a midfielder in Real Zaragoza |
| Ella decide pasar sus vacaciones en la granja | Habita en Perú . |
| She decides to spend her vacation in the country | [He/She/You] Lives in Peru |
| **Greek Subject-Verb** (subject in orange, verb in blue) | **Greek Verb-Subject** (subject in orange, verb in blue) |
| Πηγές της Αντιπολίτευσης αναφέρουν ότι [. . . ] | Το σκορ του αγώνα άνοιξε ο Γουέν Ρούνι |
| Sources of the Opposition mention that [. . . ] | The score of the game opened Wayne Rooney |
| Σε άλλες πλευρές ο ποταμός κυλά από ψηλούς βράχους | Εδώ πρέπει να γίνουν μεγαλύτερες προσπάθειες. |
| On other sides, the river flows from tall boulders | Here must happen bigger efforts |
| Η εκπαίδευση και η μόρφωση απέκτησαν επιτέλους προτεραιότητα | Απασχόληση στο εξωτερικό ψάχνουν οι μισοί Έλληνες σε παραγωγική ηλικία |
| Training and education have finally acquired priority | Employment in foreign countries search half of Greeks |

Table 1: Examples from our dataset for $S_{\text{parallel}}$ and $S_{\text{different}}$ in Spanish and Greek, along with roughly word-by-word gloss translations in English. In all cases, we've underlined $w(x)$, the word we use to represent the construction in our calculations. For presentation reasons, these examples are not randomly picked and have been chosen to be significantly shorter than the average sentence in our datasets.

tures, we wish to highlight the problem of language dominance in multilingual models, and also call for more evaluations focused on fluency.

Our proposed method can be expanded without having to manually collect data to any language with a syntactic treebank and a monolingual model. Since our method focuses on fine-grained linguistic features, some expert knowledge of the target language is necessary for evaluation. Though translated and automatically-curated multilingual evaluation has hugely helped the development of multilingual NLP, fluency evaluation — which requires some linguistic expertise to set up — has been missing from the multilingual NLP literature. Our work bridges this gap by proposing fluency testing for multilingual models.

Our work builds off of a long literature on multilingual evaluation which has until now mostly focused on downstream classification tasks (Conneau et al., 2018; Ebrahimi et al., 2022; Clark et al., 2020; Liang et al., 2020; Hu et al., 2020; Raganato et al., 2020; Li et al., 2021). With the help of these evaluation methods, research has pointed out the problems for both high- and low-resource languages that come with adding many languages to a single model (Wang et al., 2020; Turc et al., 2021; Lauscher et al., 2020, inter alia). Methods for creating more equitable models have been proposed, through identifying or reserving language-specific parameters for each language (Ansell et al., 2022; Pfeiffer et al., 2022), through training models without tyoplogically distant languages that dominate the training data (Ogueji et al., 2021; Virtanen et al., 2019), as well as through adding model capacity (Conneau et al., 2019; Xue et al., 2021; Lepikhin et al., 2021). We hope that our work can add to these analyses and methodologies by pointing out issues beyond downstream classification performance that can arise with multilingual training, and aid towards building and evaluating more equitable multilingual models.

## 2 Method

Our method relies on finding a construction in the target language which can take two structural surface forms: one which is parallel to English ($S_{\text{parallel}}$) and one which is not ($S_{\text{different}}$). Surface forms parallel to English are those which mirror English structure. For example, English has strict Subject-Verb-Object word order, and so a *parallel* structure in another language is one where the verb and its arguments appear in this order, while a *different* structure is one where the verb appears before the subject.

Once we have identified such a construction in our target language, we can ask: are multilingual models biased towards $S_{\text{parallel}}$? For a native speaker of the target language, structural, semantic, and discourse features determine whether they will use $S_{\text{parallel}}$ or $S_{\text{different}}$ in a given context — with the alternative option usually being less fluent. We assume that a BERT-sized monolingual model in the target language will have a sufficiently accurate representation of this fluent variation between $S_{\text{parallel}}$ and $S_{\text{different}}$ without being influenced by other languages. Therefore, to understand if multilingual models have an English structural bias, we now just have to answer: do multilingual models prefer $S_{\text{parallel}}$ over $S_{\text{different}}$ *more* than the fluent distribution defined by a monolingual model?

### 2.1 Collecting model judgements

By design, both $S_{\text{parallel}}$ and $S_{\text{different}}$ are constructions that occur naturally in the target language. Therefore, we should be able to use the syntactic treebank annotations to pick out sentences that exhibit the structures $S_{\text{parallel}}$ or $S_{\text{different}}$. We can put these extracted sentences into two corpora, $C_{\text{parallel}}$ and $C_{\text{different}}$. Note that the sentences in $C_{\text{parallel}}$ and $C_{\text{different}}$ are unrelated and not paired, and that the two corpora can have different sizes. Crucially, we have to use natural sentences for both of our corpora: we cannot artificially alter sentences from $S_{\text{parallel}}$ to $S_{\text{different}}$, or use templates to create sentences. This is because our evaluation is about the subtelties of fluency, and altered or templated stimuli are not naturally produced and are often awkward, confounding any effect we might want to measure.

Each model gives us a ratio $r_{\texttt{model}}$: the average probability of a sentence in $C_{\text{parallel}}$ divided by the average probability of a sentence in $C_{\text{different}}$ according to the model, that is:

$$r_{\texttt{model}} = \frac{\sum_{x \in C_p} P_{\texttt{model}}(x)\ /\ |C_p|}{\sum_{x \in C_d} P_{\texttt{model}}(x)\ /\ |C_d|} \qquad (1)$$

We want to compare judgements on these corpora from two models: a monolingual model `mono` and a multilingual model `multi`. Our experimental question then boils down to asking if $r_{\texttt{multi}}$ is significantly larger than $r_{\texttt{mono}}$.

How can we calculate $P_{\texttt{model}}(x)$ for a given sentence $x$? Looking at model judgements over long natural sentences introduces a lot of noise that is unrelated to the structural construction in question, reducing the statistical power of our experiment. Furthermore, since we are looking at encoder-only bidirectional models, there is no canonical or controlled way of extracting the probability of a whole sentence. To get a better model judgement for each sentence, we can extract the probability of *one word* in each sentence that best represents the construction. For example, if we are looking at pronoun drop, it makes sense to use main verb of the sentence as the target word, as this is the syntactic head of the pronoun that is present or dropped.

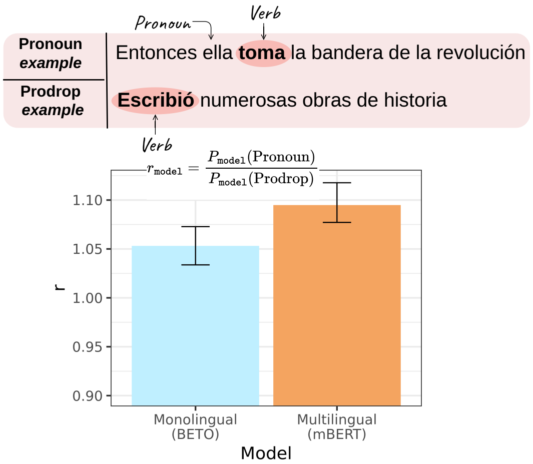

Figure 2: Results from our experiment on the Spanish GSD treebank, along with two examples from the treebank to illustrate $S_{\text{parallel}}$ (with pronoun) and $S_{\text{different}}$ (pro-drop). We compare model logits for the main verb of the sentence, which is bold and highlighted in the examples. Error bars represent 95% bootstrap confidence intervals. We find that $r_{\texttt{mono}}$ is significantly smaller than $r_{\texttt{multi}}$ (bootstrap sampling, $p < 0.05$).

If we define $w$ to be a function that returns the structurally-relevant word from each sentence, we can approximate $P_{\texttt{model}}(x)$ in Eq. (1) with $P_{\texttt{model}}(w(x)|x)$. The probability $P(w(x)|x)$ is much simpler to define for BERT-style models: it is simply the logit of the word $w(x)$ when we encode the sentence $x$ using `model`.

Extending our fluency evaluation to a new language requires three language-specific steps: (1) decide on an appropriate construction with two structural forms $S_{\text{parallel}}$ and $S_{\text{different}}$, (2) decide on an appropriate $w(x)$: which word in each structural form can represent the form, and (3) use treebank annotations to pull out sentences which exhibit $S_{\text{parallel}}$ or $S_{\text{different}}$, and identify the relevant word. We detail these steps for our two case studies.

### 2.2 Case Study: Spanish Pro-drop

In Spanish, the subject pronoun is often dropped: person and number are mostly reflected in verb conjugation, so the pronoun is realized or dropped depending on semantic and discourse factors. English, on the other hand, does not allow null subjects except in rare cases, and expletive syntactic subjects are even added when there is no clear subject (like in "There is..." sentences). For our Spanish experiment, we define $S_{\text{parallel}}$ to be sentences which have the subject pronoun of the main verb, as is necessary in English, and $S_{\text{different}}$ to be pro-drop sentences which have a main verb with no realized subject. We define $w$ to be the main verb of the sentence, which is always present in our extracted examples.

To extract our corpora $C_{\text{parallel}}$ and $C_{\text{different}}$, we use the Spanish GSD treebank from the Universal Dependencies dataset (De Marneffe et al., 2021). We ignore all sentences not verb-rooted (i.e. noun phrases), those rooted with "haber" (which when used in its copula-like existential-presentative form cannot take an explicit subject, translating to "There is" in English), and those using the impersonal-"se" passive construction (e.g. "se nos fue permitido", "it was permitted of us"). We then take all sentences with a pronoun subject (a pronoun dependent of the root verb) and add them to $C_{\text{parallel}}$ and all sentences where there is no `nsubj` relation to root verb and add them to $C_{\text{different}}$. We always pick the main root verb of the sentence as our $w$. We collect 283 sentences in $C_{\text{parallel}}$ and 2,656 sentences in $C_{\text{different}}$.

### 2.3 Case Study: Greek Subject-Verb order

English is a fixed word order language: with few exceptions, the order of a verb and its arguments is Subject-Verb-Object. Greek, on the other hand, has mostly free word order (Mackridge, 1985), meaning that the verb and arguments can appear in any order that is most appropriate given discourse context. For our experiment, we define $S_{\text{parallel}}$ to be cases in Greek when the subject precedes the verb, as is the rule in English. $S_{\text{different}}$ is then the cases when the verb precedes the subject, which almost never happens in English. We define $w$ to be the first element of the two: the subject when the subject comes first or the verb when the verb comes first. Using this $w$, we can capture the model's judgement on argument order.

To extract our corpora $C_{\text{parallel}}$ and $C_{\text{different}}$, we use the Greek Dependency Treebank, the Universal Dependencies treebank for Greek (Prokopidis and Papageorgiou, 2017). We take all sentences where the main verb has a lexical subject, and we add to $C_{\text{parallel}}$ if the subject appears before the verb and to $C_{\text{different}}$ if it appears after. We collect 1,446 sentences in $C_{\text{parallel}}$ and 425 sentences in $C_{\text{different}}$.

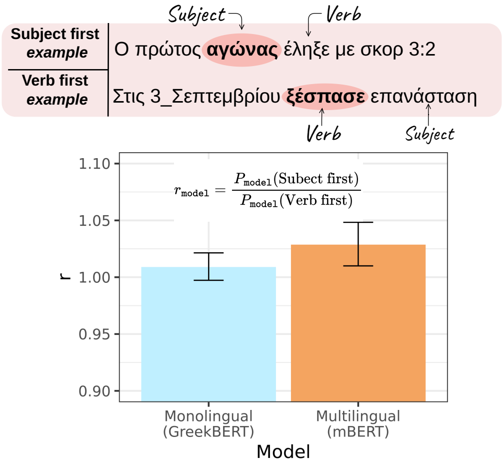

Figure 3: Results from our experiment on the Greek Dependency Treebank, along with two examples from the treebank to illustrate $S_{\text{parallel}}$ (Subject-Verb) and $S_{\text{different}}$ (Verb-Subject). We measure and compare model logits for the bold/highlighted words: the subject in subject-verb sentences and the verb in verb-subject sentences. Error bars represent 95% bootstrap confidence intervals. $r_{\texttt{mono}}$ is significantly smaller than $r_{\texttt{multi}}$ (bootstrap sampling, $p < 0.05$).

## 3 Results

Results are shown in Figures 2 and 3, showing for both of our case studies that multilingual BERT has a greater propensity for preferring English-like sentences which exhibit $S_{\text{parallel}}$. Multilingual BERT significantly prefers pronoun sentences over pro-drop compared with monolingual BETO (bootstrap sampling, $p < 0.05$), and significantly prefers subject-verb sentences over verb-subject sentences over GreekBERT (bootstrap sampling, $p < 0.05$).

## 4 Discussion

In this paper, we proposed fluency evaluation as a further way of understanding the curse of multilinguality: what can be lost when we train many languages together. The discrepancies that we point out in these experiments are not going to seriously affect multilingual LM performance, especially in the more coarse-grained classification tasks that are most commonly used for evaluation. But, as we demonstrate here, not all levels of language learning can be evaluated from such datasets. We hope the case studies in this paper can inspire more fine-grained evaluation of multilingual models, so that we understand the "accent"-like effects of hegemonic languages more fully.

## 5 Limitations

This study is meant to highlight the kinds of modeling flaws that have so far gone undetected and that can arise for lower-resource languages in multilingual models. However, our study does not focus on languages that are truly low-resource, and in fact, as designed it could not do so: our methodology relies on having an available monolingual model, which of course requires a large amount of training data. This is because our method requires a control: we can only judge multilingual models against what we can believe to be a non-biased language model in the language. There are ways to test for fluency in low-resource languages that would not require a monolingual model as a control, but would require dataset collection in the target language for features that reflect fluency and linguistic acceptability (similar to what the CoLA dataset achieves for English (Warstadt et al., 2019)). We hope our study can create motivation for such work in linguistically-aware, fine-grained multilingual evaluation for languages of all resource levels.

Our experiments focus on BERT-style models, since this is mostly the size of model available for monolingual, non-English models (in our case BETO and GreekBERT). However, it is not necessary from these experiments that our findings extrapolate to larger models that are commonplace at the time of writing.

Lastly, though these effects that we measure are sometimes discourse-dependent, we can only get isolated sentences from the UD treebanks to input to our models. However, we do not expect that having more context should favor one model more than another for our evaluation. Since this work compares models on the same inputs, we did not consider this a significant confounder.

## References

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

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

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299.

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.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, pages 110–117.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

P. Mackridge. 1985. *The Modern Greek Language: A Descriptive Analysis of Standard Modern Greek*. Oxford University Press.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. *arXiv preprint arXiv:2205.06266*.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal dependencies for greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. The Association for Computational Linguistics.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *CoRR*, abs/2106.16171.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL-HLT*.