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

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

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 (Kreutzer et al., 2022), 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*.

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 subject-before-verb order in Greek: the structural forms parallel to English.

The effect we showcase here demonstrates the type of fluency that can be lost with multilingual training — something that current evaluation methods miss. Our proposed method can be expanded, without the need for manual data collection, 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.

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), and proposed methods for more equitable models (Ansell et al., 2022; Pfeiffer et al., 2022; Ogueji et al., 2021; Ògúnrẹ̀mí and Manning, 2023; Virtanen et al., 2019; Liang et al., 2023, inter alia). 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 variable construction in the target language which can take two structural surface forms: one which is parallel to English (S_{parallel}) and one which is not ($S_{\text{different}}$). Surface forms parallel to English are those which mirror English structure.

Once we have identified such a construction in our target language, we can ask: are multilingual models biased towards S_{parallel}? We can use syntactic treebank annotations to pick out sentences that exhibit the structures S_{parallel} or $S_{\text{different}}$, and put these extracted sentences into two corpora, C_{parallel} and $C_{\text{different}}$. We then calculate a ratio $r_{\texttt{model}}$ for each model: the average probability of a sentence in C_{parallel} divided by the average probability of a sentence in $C_{\text{different}}$ according to the model. Our experimental question then boils down to asking if $r_{\texttt{multi}}$ is significantly larger than $r_{\texttt{mono}}$. To get an estimation of $P_{\texttt{model}}(x)$, we can extract the prob-

Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023), pages 143–146
May 6, 2023 ©2023 Association for Computational Linguistics

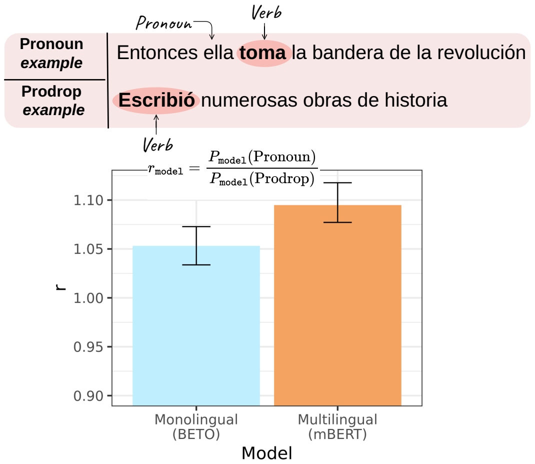

Figure 1: Results from our experiment on the Spanish GSD treebank, along with two examples from the treebank to illustrate S_{parallel} (with pronoun) and $S_{\text{different}}$ (pro-drop). Error bars represent 95% bootstrap confidence intervals.

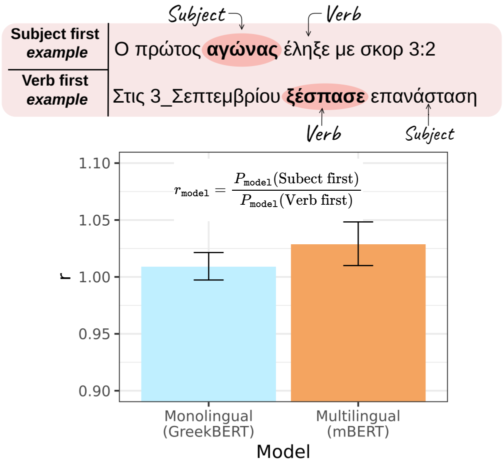

Figure 2: Results from our experiment on the Greek Dependency Treebank, along with two examples from the treebank to illustrate S_{parallel} (Subject-Verb) and $S_{\text{different}}$ (Verb-Subject). Error bars represent 95% bootstrap confidence intervals.

ability of *one word* w in each sentence that best represents the construction, and approximate the probability of x with $P(w_x|x)$. Using a carefully chosen word as a proxy for the probability of a construction is a methodological choice also made in reading time psycholinguistics experiments (Levy and Keller, 2013).

2.1 Case Study: Spanish Pro-drop

For our Spanish case study, we examine the feature of whether the subject pronoun is realized. 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, even adding expletive syntactic subjects as in "**it** is raining". We extract C_{parallel} (with subject pronoun) and $C_{\text{different}}$ (dropepd subject pronoun) from the Spanish GSD treebank (De Marneffe et al., 2021). We take all sentences with a pronoun dependent of the root verb and add them to C_{parallel} (283 sentences) and all sentences where there is no `nsubj` relation to root verb and add them to $C_{\text{different}}$ (2,656 sentences), ignoring some confounder constructions. We always pick the main root verb of the sentence as our logit word w.

2.2 Case Study: Greek Subject-Verb order

For our Greek case study, we examine the feature of 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_{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 extract C_{parallel} (Subject-Verb order, 1,446 sentences) and $C_{\text{different}}$ (Verb-Subject order, 425 sentences) from the Greek Dependency Treebank (Prokopidis and Papageorgiou, 2017). We define w to be the first element of the subject and verb: This first element is closer to the surrounding context, and so gives us a word-order-sensitive measurement of how the subject-verb construction is processed within the context.

3 Results

Results are shown in Figures 1 and 2, showing for both of our case studies that multilingual BERT has a greater propensity for preferring English-like sentences which exhibit S_{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, $p < 0.05$).

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, 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 Pre-trained 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.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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.

Roger P. Levy and Frank Keller. 2013. Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, 68(2):199–222.

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.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. *arXiv preprint arXiv:2301.10472*.

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.

Tolúlọpẹ́ Ògúnrẹ̀mí and Christopher D. Manning. 2023. Mini but Mighty: Efficient multilingual pretraining with linguistically-informed data selection.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

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.