Are Pretrained Multilingual Models Equally Fair Across Languages?

Anonymous ACL submission

Abstract

Pretrained multilingual language models can help bridge the digital language divide, enabling high-quality NLP models for lower-resourced languages. Studies of multilingual models have so far focused on performance, consistency, and cross-lingual generalization. However, with their wide-spread application in the wild and downstream societal impact, it is important to put multilingual models under the same scrutiny as monolingual models. This work investigates the group fairness of multilingual models, asking whether these models are equally fair across languages. To this end, we create a new four-way multilingual dataset of parallel cloze test examples (MozArt), equipped with demographic information (balanced with regard to gender and native tongue) about the test participants. We evaluate three multilingual models on MozArt – mBERT, XLM-R, and mT5 – and show that across the four target languages, the three models exhibit different levels of group disparity, e.g., exhibiting near-equal risk for Spanish, but high levels of disparity for German.

1 Introduction

Fill-in-the-gap cloze tests (Taylor, 1953) ask human language learners to predict what words were removed from a text. Today, language models are trained to do the same Devlin et al. (2019). This has the advantage that we can now use fill-in-the-gap cloze tests to directly compare the linguistic preferences of humans and language models, e.g., to investigate task-independent sociolectal biases (group disparities) in language models (Zhang et al., 2021). This paper presents a novel four-way parallel cloze dataset for English, French, German, and Spanish that enables apples-to-apples comparison across languages of group disparities in multilingual language models.

Language models induced from historical data are prone to implicit biases (Zhao et al., 2017; Chang et al., 2019; Mehrabi et al., 2021), e.g., as a result of the over-representation of male-dominated text sources such as Wikipedia and newswire (Hovy and Søgaard, 2015). This may lead to language models that are *unfair* to groups of users in the sense that they work better for some groups rather than others (Zhang et al., 2021). Multilingual language models can be said to be unfair to their training languages in similar ways (Choudhury and Deshpande, 2021; Anonymous, 2022; Wang et al., 2021), but this work goes beyond previous work in evaluating whether multilingual language models are *equally fair to demographic groups across languages*.

	EN	**ES**	**DE**	**FR**
WordPiece	19.7	22.0	23.6	23.1
SentencePiece	22.3	22.9	24.9	25.3

#Sentences per language: 100
#Annotations per sentence: 6
#Annotators: 240

Demographics: id_u, id_s, gender, age, first language, fluent languages, nationality, current country of residence, country of birth, time taken

Table 1: MozArt details: average number of tokens per sentence are reported using WordPiece and SentecePiece. In demographics, id_u refers to user id (anonymised) and id_s to sentence id.

To this end, we create MozArt, a multilingual dataset of fill-in-the-gap sentences covering four languages (English, French, German and Spanish). The sentences reflect diastratic variation within each language and can be used to compare biases in pretrained language models (PLMs) across languages. We study the influence of four demographic groups, i.e., the cross-product of our annotators' gender – male (*M*) or female (*F*)[1] – and first

[1]None of our annotators identified as non-binary.

language – native (*N*) or non-native (*NN*) [2] –. Table 1 presents a summary of dataset characteristics.

2 Dataset

We introduce MozArt, a dataset of parallel data in four languages (English, French, German and Spanish) with annotators' demographics. We sampled 100 sentence quadruples from the corpus provided for the WMT 2006 Shared Task.[3] This data was originally taken from the publicly available Europarl corpus (Koehn, 2005) and enhanced with word-alignments. We manually verify that sentences make sense out of context and use the data to generate *comparable cloze examples* such as:[4]

en	[MASK] that deplete the ozone layer
es	[MASK] que agotan la capa de ozono
de	[MASK], die zum Abbau der Ozonschicht führen
fr	[MASK] appauvrissant la couche d'ozone

The masked words are aligned (by one-to-one alignments) and either nouns, verbs, adjectives or adverbs.[5] We mask one word in each sentence and verify that one-to-one alignments exist in all languages. Following (Kleijn et al., 2019a), we avoid masking words that are too predictable, e.g., auxiliary verbs or constituents of multi-word expressions, or masking words that are unpredictable, e.g., proper names and technical terms.

Annotators were recruited using Prolific.[6] We applied eligibility criteria to balance our annotators across demographics. Participants were asked to report (on a voluntary basis) their demographic information regarding gender and languages spoken. Each eligible participant was presented with 10 cloze examples. We collected answers from 240 annotators, 60 per language batch, divided in four balanced demographic groups (gender × native language). We made sure that each sentence had at least six annotations. Annotation guidelines for each language were given in that language, to avoid bias and ensure a minimum of language understanding for non-native speakers. We manually filtered out spammers to ensure data quality.

The dataset is made publicly available at `github.com/anonymized` under a CC-BY-4.0 license. We include all the demographic attributes of our annotators as per agreement with the annotators. The full list of protected attributes is found in Table 1. We hope MozArt will become a useful resource for the community, also for evaluating the fairness of language models across other attributes than gender and native language.

3 Experimental Setup

Models We evaluate three PLMs: mBERT (Devlin et al., 2019), XLM-RoBERTa/XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021).[7] All three models were trained with a masked language modeling objective. mBERT differs from XLM-R and mT5 in including a next sentence prediction objective (Devlin et al., 2019). mT5 differs from mBERT and XLM-R in allowing for consecutive spans of input tokens to be masked (Raffel et al., 2020). Since mT5 is trained to reconstruct the masked-out tokens, we constrain the generation to generate single words. This enables correlation of mT5's output with our group preferences. t-SNE plots are included in Appendix B to show how languages are distributed in the PLM vector spaces.

Metrics We use several metrics to compare how the PLMs align with group preferences across languages. These include top-k precision $P@k$ with k={1, 5}, mean reciprocal rank (MRR), and two classical univariate rank correlations: Spearman's ρ (Spearman, 1987) and Kendall's τ (Kendall, 1938).

Given a set of $|S|$ cloze sentences and a group of annotators, for each sentence s, we denote the list of answers, ranked by their frequency, as $W_s = [w_1, w_2, ...]$, and the list of model's predictions as $C_s = [c_1, c_2, ...]$, ranked by their model likelihood. Then, we report $P@k = \mathbb{1}[c_i \in W_s]$ with $i \in [1, k]$, where $\mathbb{1}[\cdot]$ is the indicator function. Precision is reported together with its standard deviation, to account for the group-wise disparity in both dimensions (social groups and language):

$$\sigma_{\mathrm{gd}} = \sqrt{\frac{\sum_{i=1}^{G}(P@k_i - \overline{P@k})^2}{G}} \quad (1)$$

[2] See Schmitz (2016); Faez (2011) for discussion of the native/non-native speaker dichotomy. Participants were asked 'What is your first language?' and 'Which of the following languages are you fluent in?'. We use *native* (*N*) for people whose first language coincides with the example sentences, and non-native (*NN*) otherwise, without any sociocultural implications.

[3] https://www.statmt.org/wmt06/shared-task/

[4] For brevity, we only present noun phrases, not the full sentences.

[5] Using spaCy's POS tagger (Honnibal and Montani, 2017).

[6] `prolific.co`

[7] We use the base models available from `https://huggingface.co/models`. We report results using uncased mBERT, since it performed better on our data than its cased sibling.

where $\overline{\text{P@k}}$ is the mean value of all observations, and G the total number of groups across the dimension fixed each time i.e., $G = 4$ across social groups (*MN, FN, MNN, FNN*) and $G = 4$ across languages (EN, ES, DE, FR). We also compute the mean-reciprocal rank (MRR) of the elements of W_s with respect to the top-n ($n = 5$) elements of C_s (C_s^n):

$$\text{MRR} = \frac{1}{|S|} \sum_{s=1}^{|S|} \frac{1}{Rank_i^{C_s^n}} \tag{2}$$

Finally, we compute Spearman's ρ (Spearman, 1987) and Kendall's τ (Kendall, 1938) between W_s and C_s^5. These metrics are generally more robust to outliers.

4 Results

Following previous work on examining fairness of document classification (Huang et al., 2020; Dixon et al., 2018; Park et al., 2018; Garg et al., 2019), we focus on group-level performance differences (group disparity). We measure the group disparity as the variance in PLM's performance (P@k) across demographics (gender and native language). Table 2 shows **better precision for native speakers in German and French** (*MN*, *FN*) for P@1. In terms of group disparity, male non-natives (*MNN*) is the demographic exhibiting the highest disparity across languages in mBERT, while it is female natives (*FN*) in XLM-R. Language-wise, we see the largest group disparity with German in both models. Here, we see 3.5–4.4 between-group differences, compared to, e.g., 0.3–1.8 between-group differences for English. See Appendix A for results with P@5.

XLM-R consistently exhibits better overall performance on average, but higher between-group and between-language differences.

Figure 1 complements results from Table 2 with MRR scores and compares them to mT5. We observe a common trend that the models often underperform on non-native male speakers in all languages except for Spanish: Performance is (always) below the average, and they are the worst-off group (↓) in most of the cases. At the same time, predictions with mBERT and XLM-R seem to be biased towards native speakers because answers from *MN* and *FN* generally rank highest. Despite none of the models perform equally across groups, XLM-R shows a lower divergence across languages: Between-group differences are more than 50% smaller than with mBERT and mT5 when looking at the average MRR per language.

		P@1				
		EN	**ES**	**DE**	**FR**	P@1(σ_{gd})
MN	mBERT	13.3	12.7	11.3	10.7	12.0 (1.0)
	XLM-R	16.7	13.3	20.7	16.7	16.9 (2.6)
FN	mBERT	13.3	12.0	15.3	8.0	12.2 (2.7)
	XLM-R	16.0	15.3	24.0	17.3	18.2 **(3.5)**
MNN	mBERT	12.7	12.4	11.4	3.6	10.0 **(3.8)**
	XLM-R	15.3	13.5	15.0	11.4	13.8 (1.5)
FNN	mBERT	13.3	10.0	5.6	6.9	9.0 (3.0)
	XLM-R	20.0	14.7	13.1	12.7	15.1 (3.0)
$\overline{P@1}(\sigma_{gd})$	mBERT	13.2 (0.3)	11.8 (1.1)	10.8 **(3.5)**	7.3 (2.5)	
	XLM-R	17.0 (1.8)	14.2 (0.8)	18.2 **(4.4)**	14.5 (2.6)	

Table 2: Results on P@1 score across groups and languages, average performance in each language ($\overline{P@1}$) as well as standard deviation for group disparity (σ_{gd}). Cells with a colored background are language-wise above the average. For each model, worst group performance in terms of group disparity (highest variance) is highlighted in red.

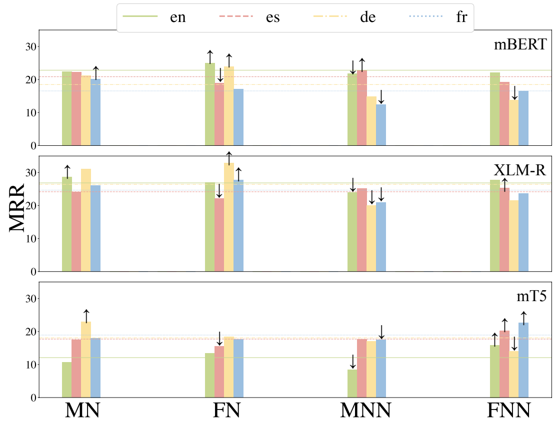

Figure 1: Average MRR (in percentage) per group in each language. Horizontal lines denote the average per language. Best-off (↑) and worst-off (↓) subgroups for each language are marked.

Table 3 gathers group level Spearman's ρ and average correlation per language. XLM-R predictions are more uniformly correlated across languages compared to mBERT, whose lexical preferences are better aligned in English and Spanish setups, and mT5, whose predictions correlate poorly with human cloze test answers. However, in line with previous results, the model exhibits bias towards male native speakers and *MNN* outlines as the worst performing group across languages, with a coefficient always below the average. Looking into the dimension of languages, German is the least aligned with human's answers in all models. See Appendix A for details on Kendall's τ.

5 Related Work

Multilingual PLMs have been analyzed in many ways: Researchers have, for example, looked at performance differences across languages (Singh

mBERT

ρ	EN	ES	DE	FR
MN	0.33 (p=0.00)	0.23 (p=0.01)	-0.14 (p=0.09)	0.10 (p=0.21)
FN	0.27 (p=0.00)	0.07 (p=0.42)	-0.01 (p=0.89)	0.14 (p=0.08)
MNN	0.30 (p=0.00)	0.16 (p=0.03)	-0.10 (p=0.23)	0.08 (p=0.32)
FNN	0.37 (p=0.00)	0.16 (p=0.06)	0.03 (p=0.69)	0.08 (p=0.30)
Avg.	0.32 (p=0.00)	0.16 (p=0.00)	-0.05 (p=0.21)	0.10 (p=0.01)

XLM-R

ρ	EN	ES	DE	FR
MN	0.45 (p=0.00)	0.46 (p=0.00)	0.35 (p=0.00)	0.48 (p=0.00)
FN	0.30 (p=0.00)	0.35 (p=0.00)	0.45 (p=0.00)	0.33 (p=0.00)
MNN	0.30 (p=0.00)	0.38 (p=0.00)	0.22 (p=0.01)	0.32 (p=0.00)
FNN	0.40 (p=0.00)	0.48 (p=0.00)	0.11 (p=0.16)	0.36 (p=0.00)
Avg.	0.36 (p=0.00)	0.41 (p=0.00)	0.28 (p=0.00)	0.37 (p=0.00)

mT5

ρ	EN	ES	DE	FR
MN	0.01 (p=0.89)	0.14 (p=0.08)	0.14 (p=0.08)	0.25 (p=0.00)
FN	-0.12 (p=0.13)	0.13 (p=0.12)	0.00 (p=0.99)	0.14 (p=0.08)
MNN	-0.10 (p=0.22)	0.12 (p=0.11)	0.03 (p=0.74)	0.11 (p=0.18)
FNN	-0.07 (p=0.41)	0.28 (p=0.00)	0.04 (p=0.58)	0.11 (p=0.16)
Avg.	-0.07 (p=0.07)	0.17 (p=0.00)	0.05 (p=0.23)	0.15 (p=0.00)

Table 3: Correlation between groups of annotators (*MN, FN, MNN, FNN*) and models' predictions, classified by language. The degree of correlation is measured with Spearman's ρ coefficient ($\rho \in [-1, 1]$). Cells highlighted in red fail to reject the null hypothesis, meaning that their difference is statistically significant (p>0.05). Groups with coloured background show a stronger correlation compared to the average in each language.

et al., 2019), looked at their organization of language types (Rama et al., 2020), used similarity analysis to probe their representations (Kudugunta et al., 2019), and investigated how learned self-attention in the Transformer blocks affects different languages (Ravishankar et al., 2021).

Previous work on fairness of multilingual models has, to the best of our knowledge, focused exclusively on task-specific models, rather than PLMs: Huang et al. (2020) evaluate the fairness of multilingual hate speech detection models, and several researchers have explored gender bias in multilingual models (Zhao et al., 2020; González et al., 2020). Dayanik and Padó (2021) consider the effects of adversarial debiasing in multilingual models.

Cloze tests were previously used in Zhang et al. (2021) to evaluate the fairness of English (monolingual) language models. In psycholinguistics, cloze tests have been performed with different age groups (Hintz et al., 2020) and native language (Stringer and Iverson, 2020), but these datasets have, to the best of our knowledge, not been used to evaluate language models.

6 Conclusion

In this paper, we present MozArt, a new multilingual dataset of parallel cloze examples with annotations from balanced demographics. This dataset is, to the best of our knowledge, the first to enable apples-to-apples comparison of group disparity of multilingual PLMs across languages. The dataset includes several demographic attributes, but we present preliminary experiments with gender and native language. We show that mBERT and XLM-R are not equally fair across languages. For example, group disparities are much higher for German (and French) than for English and Spanish. This shows the importance of evaluating fairness across languages instead of stipulating from results for a single language. We further show that both PLMs align best with the cloze test answers of female native speakers. We followed best practices for mitigating the dangers of crowdsourcing (Karpinska et al., 2021; Kleijn et al., 2019b) (see §2) and hope MozArt will be widely adopted and, over time, generate more results for other PLMs and demographic attributes.

Ethics Statement

The dataset released contains publicly available content from the proceedings of the European Parliament. Our work is based on sensitive information provided by the participants that took on our study in Prolific. The protected attributes collected are self-reported on a voluntary basis, and participants gave their consent to share them. In addition to the specific attributes analyzed in our study, which served as prescreening filters, Prolific also provides baseline data for all studies with the consent of participants to share it with researchers. For these base attributes, there might be gaps in the data because it is optional for participants to provide this information. These attributes are filled as *null* in the dataset. We performed a pilot study to determine the amount of time a task would take on average. The participants were paid based on time worked, and were given the option to opt out at any time of the study. Participants who revoked consent at any stage are not included in our study nor in the data released.

References

Anonymous. 2022. Fairness in representation for multilingual NLP: Insights from controlled experiments on conditional language modeling. In *Submitted to The Tenth International Conference on Learning Representations*. Under review.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. 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): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *AAAI-21*. AAAI, AAAI.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Erenay Dayanik and Sebastian Padó. 2021. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, 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.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Farahnaz Faez. 2011. Reconceptualizing the native/nonnative speaker dichotomy. *Journal of Language, Identity & Education*, 10(4):231–249.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

F Hintz, M Dijkhuis, V van Hoff, JM McQueen, and AS Meyer. 2020. A behavioural dataset for studying individual differences in language skills. *Scientific Data*, 7(1).

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Suzanne Kleijn, Henk Pander Maat, and Ted Sanders. 2019a. Cloze testing for comprehension assessment: The hytec-cloze. *Language Testing*, 36(4):553–572.

Suzanne Kleijn, Henk Pander Maat, and Ted Sanders. 2019b. Cloze testing for comprehension assessment: The hytec-cloze. *Language Testing*, 36(4):553–572.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. 2021. Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online. Association for Computational Linguistics.

John Schmitz. 2016. On the native/nonnative speaker notion and world englishes: Debating with k. rajagopalan. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 32:597–611.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

C. Spearman. 1987. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471.

Louise Stringer and Paul Iverson. 2020. Non-native speech recognition sentences: A new materials set for non-native speech perception research. *Behavior Research Methods*, 52(2).

Wilson L. Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, 30:415 – 433.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Assessing multilingual fairness in pre-trained multimodal representations.

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.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. 2021. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

		P@5				
		EN	**ES**	**DE**	**FR**	P@5(σ_{gd})
MN	mBERT	30.7	26.7	22.0	24.0	25.9 (3.3)
	XLM-R	39.3	30.7	34.7	32.7	34.4 (3.2)
FN	mBERT	32.0	18.7	24.7	22.0	24.4 (4.9)
	XLM-R	30.7	25.3	38.0	35.3	32.3 (4.8)
MNN	mBERT	34.0	25.9	12.1	15.0	21.8 **(8.7)**
	XLM-R	30.7	29.4	22.1	25.4	26.9 (3.4)
FNN	mBERT	32.7	25.3	16.3	16.3	22.7 (6.9)
	XLM-R	36.7	34.0	19.4	26.9	29.3 **(6.7)**
$\overline{P@5}(\sigma_{gd})$	mBERT	32.3 (1.2)	24.2 (3.1)	18.8 **(4.9)**	19.3 (3.8)	
	XLM-R	34.3 (3.8)	29.8 (3.1)	28.5 **(7.9)**	30.3 (4.1)	

Table 4: Results on P@5 score across groups and languages, average performance in each language ($\overline{P@5}$) as well as standard deviation for group disparity (σ_{gd}). Cells with a colored background are language-wise above the average. For each model, worst group performance in terms of group disparity (highest variance) is highlighted in red.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints.

A Additional results

In this section, we provide additional analysis results of the PLM's performance on MozArt. We report precision at 5 (P@5), which corresponds to the number of relevant answers among the top 5 candidates. It provides a more flexible metric for measuring model alignments with open-ended text answers, but fails to take into account the exact position within the top-k. Considering the top-5, the bias towards native speakers is diminished specially in English and Spanish, despite being *MNN* and *FNN* the worst groups –in terms of group disparity– in mBERT and XLM-R respectively. At the same time, the group disparities are exacerbated as shown in Table 4.

Table 5 complements results on correlation of the alignment of group responses. It shows Kendall's τ coefficient. Conclusions remain almost the same as studied with Spearman's coefficient, albeit non-native subgroups in Spanish are more correlated in mBERT.

B t-SNE

To give a brief overview of the semantic multilinguality encoded in the pretrained models, we run

mBERT

τ	EN	ES	DE	FR
MN	0.27 (p=0.00)	0.19 (p=0.00)	-0.09 (p=0.15)	0.09 (p=0.16)
FN	0.23 (p=0.00)	0.07 (p=0.24)	0.01 (p=0.89)	0.13 (p=0.04)
MNN	0.25 (p=0.00)	0.15 (p=0.01)	-0.06 (p=0.32)	0.07 (p=0.28)
FNN	0.29 (p=0.00)	0.14 (p=0.01)	0.03 (p=0.57)	0.06 (p=0.27)
Avg.	0.26 (p=0.00)	0.14 (p=0.00)	-0.03 (p=0.41)	0.09 (p=0.01)

XLM-R

τ	EN	ES	DE	FR
MN	0.40 (p=0.00)	0.43 (p=0.00)	0.32 (p=0.00)	0.45 (p=0.00)
FN	0.26 (p=0.00)	0.33 (p=0.00)	0.43 (p=0.00)	0.31 (p=0.00)
MNN	0.26 (p=0.00)	0.35 (p=0.00)	0.20 (p=0.01)	0.29 (p=0.00)
FNN	0.35 (p=0.00)	0.45 (p=0.00)	0.10 (p=0.15)	0.34 (p=0.00)
Avg.	0.32 (p=0.00)	0.39 (p=0.00)	0.25 (p=0.00)	0.34 (p=0.00)

mT5

τ	EN	ES	DE	FR
MN	0.02 (p=0.79)	0.13 (p=0.06)	0.13 (p=0.06)	0.21 (p=0.00)
FN	-0.09 (p=0.16)	0.11 (p=0.11)	0.00 (p=0.98)	0.12 (p=0.08)
MNN	-0.08 (p=0.21)	0.10 (p=0.10)	0.03 (p=0.69)	0.10 (p=0.17)
FNN	-0.04 (p=0.51)	0.25 (p=0.00)	0.03 (p=0.61)	0.10 (p=0.15)
Avg.	-0.07 (p=0.07)	0.15 (p=0.00)	0.05 (p=0.18)	0.13 (p=0.00)

Table 5: Correlation between groups of annotators (MN, FN, MNN, FNN) and models' predictions, classified by language. The degree of correlation is measured with Kendall's τ coefficient ($\tau \in [-1, 1]$). Cells highlighted in red fail to reject the null hypothesis, meaning that their difference is statistically significant (p>0.05). Groups with coloured background show a stronger correlation compared to the average in each language.

several representations with t-SNE. Figure 2 and Figure 3 represent the top-1000 predictions in a t-SNE plot for mBERT and XLM-R respectively. The same sentence is queried to the model in four languages and, accordingly, to annotators:

en	We want to [MASK] innovation .
es	Queremos [MASK] la innovación .
de	Wir wollen zur Innovation [MASK] .
fr	Nous voulons [MASK] l'innovation .

Highest scored predictions are highlighted with a ($\star$). Annotator's answers that fell into the top-1000 predictions are denoted with a black edge. In line with results in (Choenni and Shutova, 2020), we appreciate in both models that languages are projected in separate sub-spaces instead of yielding a neutral representation, even though they share a common space (vocabulary).

Similarly, Singh et al. (2019) shown a trend towards dissimilarity between representations for semantically similar inputs in different languages, in deeper layers of an uncased mBERT. Serve Figure 4 as an example, where the same word 'gases' was answered in different languages but is represented in different subspaces. Figure 5 shows a similar behaviour in XLM-R. The sentences queried are:

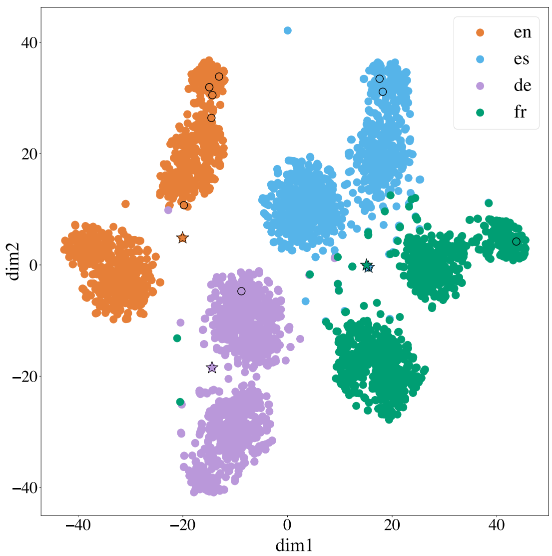

Figure 2: t-SNE representation from the last layer of mBERT for the top-1000 predictions for the parallel sentences in the list above ('We want to [MASK] innovation .' in English). Highest scored prediction is starred; annotator's answers are denoted by a dot with black edge. Legend shows language-color mapping.

en	[MASK] that deplete the ozone layer
es	[MASK] que agotan la capa de ozono
de	[MASK], die zum Abbau der Ozonschicht führen
fr	[MASK] appauvrissant la couche d'ozone

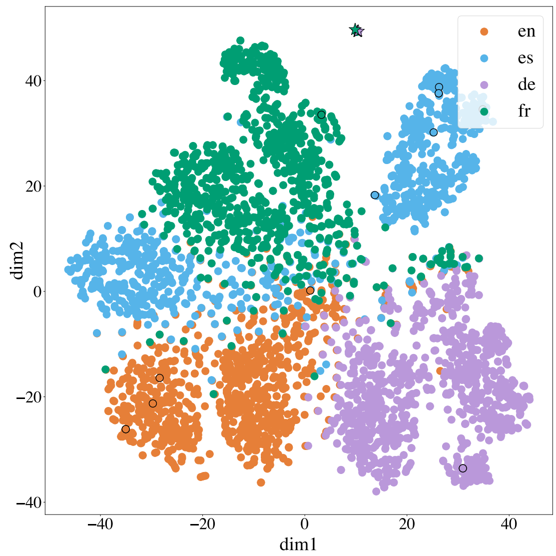

Figure 3: t-SNE representation from the last layer of XLM-R for the top-1000 predictions for the parallel sentences in the list above ('We want to [MASK] innovation .' in English). Highest scored prediction is starred; annotator's answers are denoted by a dot with black edge. Legend shows language-color mapping.

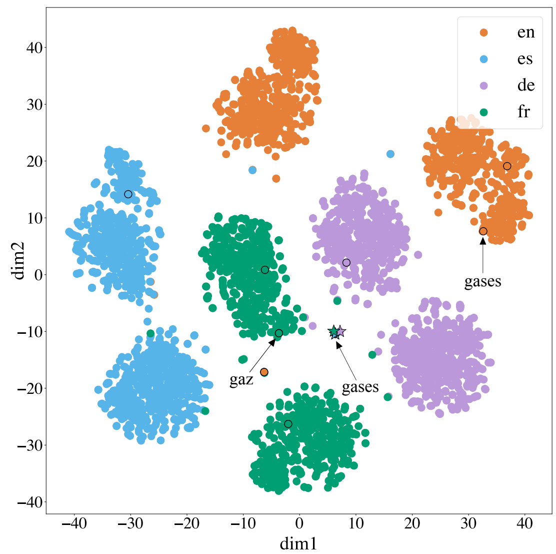

Figure 4: t-SNE representation from the last layer of mBERT for the top-1000 predictions for the parallel sentences in the list above ('[MASK] that deplete the ozone layer' in English). The word 'gases' is pointed out in each language (en: gases, es: gases, fr:gaz), as it was a recurrent answer from different annotators. Highest scored prediction is starred; annotator's answers are denoted by a dot with black edge. Legend shows language-color mapping.

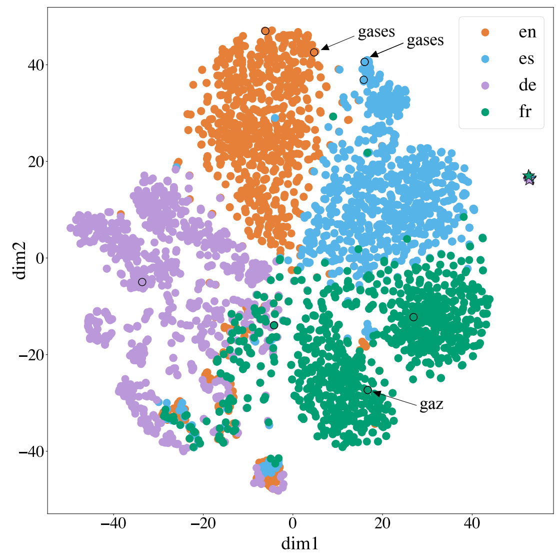

Figure 5: t-SNE representation from the last layer of XLM-R for the top-1000 predictions for the parallel sentences in the list above ('[MASK] that deplete the ozone layer' in English). The word 'gases' is pointed out in each language (en: gases, es: gases, fr:gaz), as it was a recurrent answer from different annotators. Highest scored prediction is starred; annotator's answers are denoted by a dot with black edge. Legend shows language-color mapping.