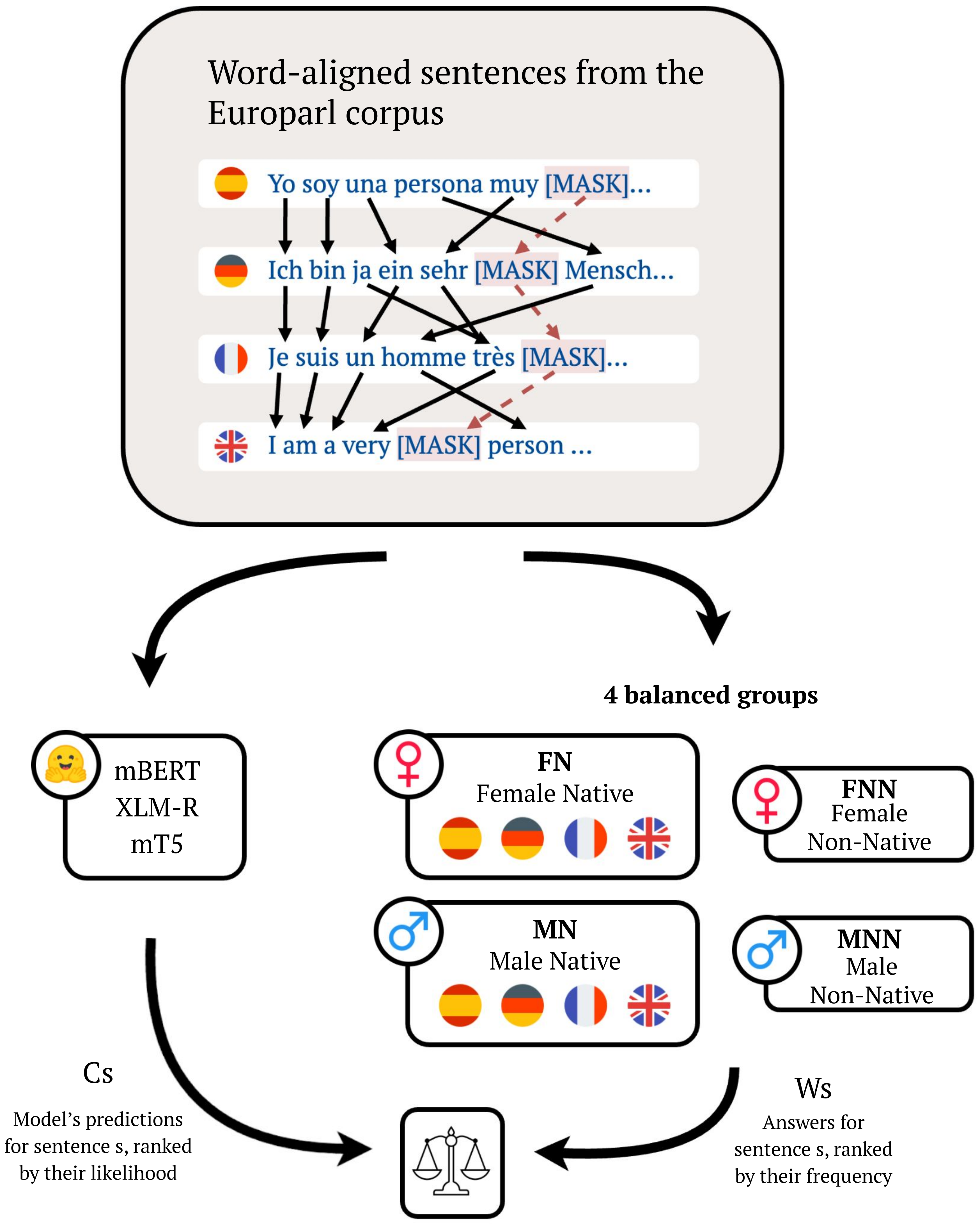
Are Pretrained Multilingual Models Equally Fair Across Languages?

Laura Cabello Piqueras, Anders Søgaard
University of Copenhagen
lcp@di.ku.dk

Main contribution

MozArt: A new multilingual dataset annotated with balanced demographics. github.com/coastalcph/mozart

Method


Word-aligned sentences from the Europarl corpus
Yo soy una persona muy [MASK]...
Ich bin ja ein sehr [MASK] Mensch...
Je suis un homme très [MASK]...
I am a very [MASK] person ...
mBERT
XLM-R
mT5
4 balanced groups
FN
Female Native
FNN
Female
Non-Native
MN
Male Native
MNN
Male
Non-Native
Cs
Model's predictions for sentence s, ranked by their likelihood
Ws
Answers for sentence s, ranked by their frequency

Metrics

Fairness metric: group disparity
(group-level performance differences)

$$\sigma_{\text{gd}} = \sqrt{\frac{\sum_{j=1}^{G} (\text{P@k}_j - \overline{\text{P@k}})^2}{\text{G}}}$$

$$\text{P@k} = \mathbb{1}[c_i \in W_s]$$

G = number of groups

Mean Reciprocal Rank

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

|S| = 100 sentences

Rank Correlations

Spearman's rho, Kendall's tau (see paper for results)

MozArt details

	EN	ES	DE	FR
WordPiece (avg. #tokens)	19.7	22.0	23.6	23.1
SentencePiece (avg. #tokens)	22.3	22.9	24.9	25.3
#Sentences	100	100	100	100
#Annotations	600	600	600	600
#Annotators	60	60	60	60
Demographics	id_u, id_s, gender, age, nationality, first language, fluent languages, current country of residence, country of birth, time taken			

Experimental Results

Precision + std. dev.

mBERT

P@1	**EN**	**ES**	**DE**	**FR**	
MN	13.3	12.7	11.3	10.7	12.0 (1.0)
FN	13.3	12.0	15.3	8.0	12.2 (2.7)
MNN	12.7	12.4	11.4	3.6	10.0 **(3.8)**
FNN	13.3	10.0	5.6	6.9	9.0 (3.0)
	13.2 (0.3)	11.8 (1.1)	10.8 **(3.5)**	7.3 (2.5)	$\overline{\text{P@1}}(\sigma_{gd})$

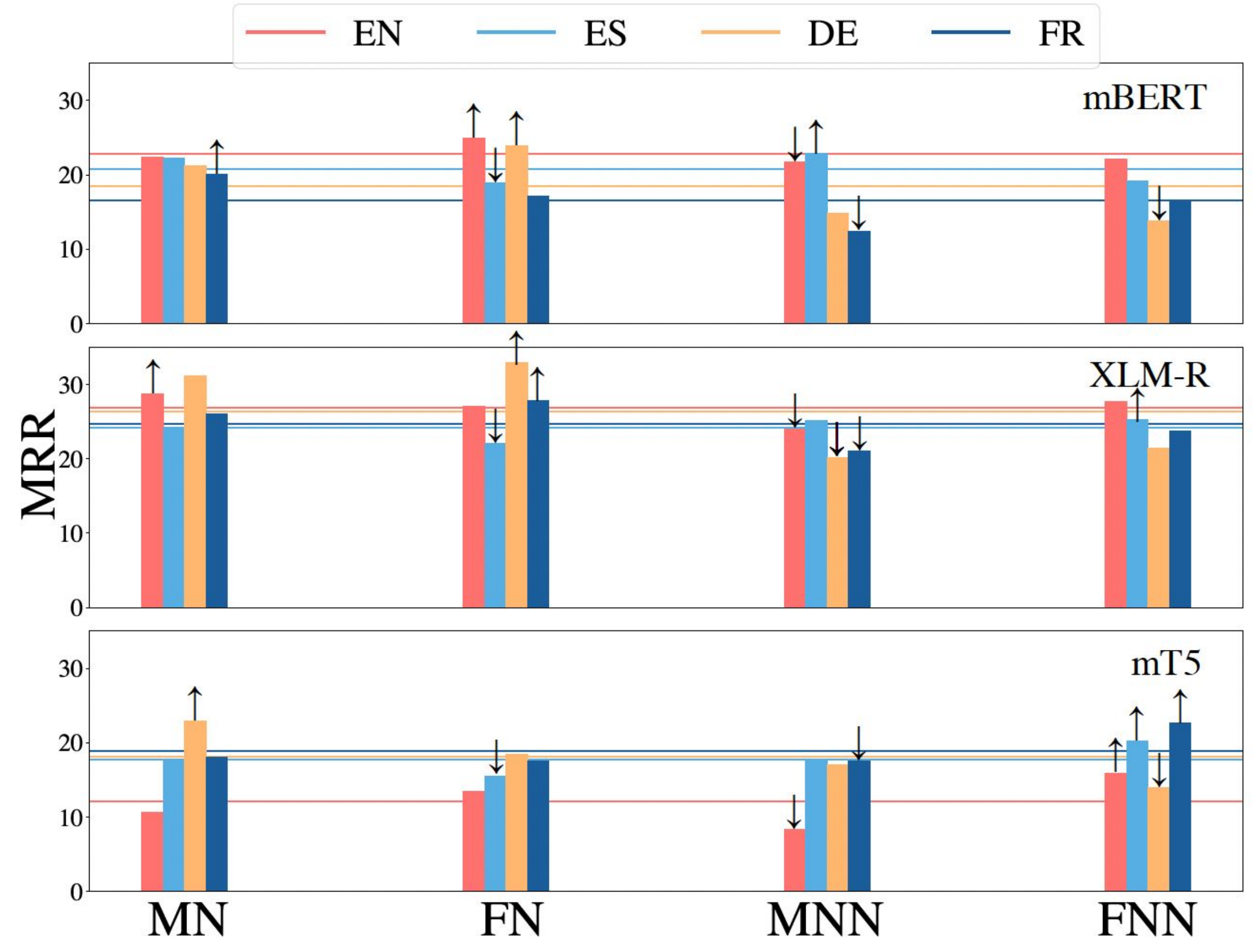
XLM-R

P@1	**EN**	**ES**	**DE**	**FR**	
MN	16.7	13.3	20.7	16.7	16.9 (2.6)
FN	16.0	15.3	24.0	17.3	18.2 **(3.5)**
MNN	15.3	13.5	15.0	11.4	13.8 (1.5)
FNN	20.0	14.7	13.1	12.7	15.1 (3.0)
	17.0 (1.8)	14.2 (0.8)	18.2 **(4.4)**	14.5 (2.6)	$\overline{\text{P@1}}(\sigma_{gd})$

mT5

P@1	**EN**	**ES**	**DE**	**FR**	
MN	2.0	4.7	8.7	5.3	5.2 **(2.4)**
FN	4.0	3.3	6.7	3.3	4.3 (1.4)
MNN	2.0	4.7	6.4	4.3	4.4 (1.6)
FNN	3.3	6.7	1.9	6.2	4.5 (2.0)
	2.8 (0.9)	4.8 (1.2)	5.8 **(2.5)**	4.8 (1.1)	$\overline{\text{P@1}}(\sigma_{gd})$

Mean Reciprocal Rank



Conclusion

MozArt! A new multilingual dataset of parallel cloze examples with demographic information from annotators. We show that **multilingual PLM are not equally fair** across languages; neither across groups of users.