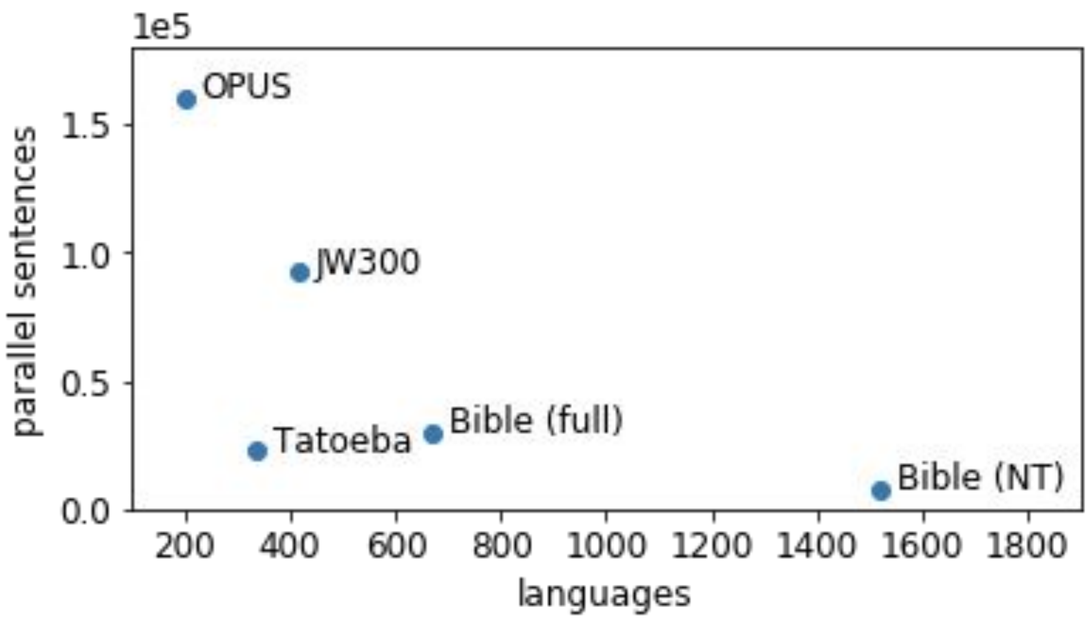
JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages

Željko Agić <za@corti.ai> and **Ivan Vulić** <ivan@poly-ai.com>

PolyAI
find your voice

The Proxy Fallacy

- resource-rich languages posing as low-resource languages in experiments
- bias creeps in
- only real solution = **more resources?**



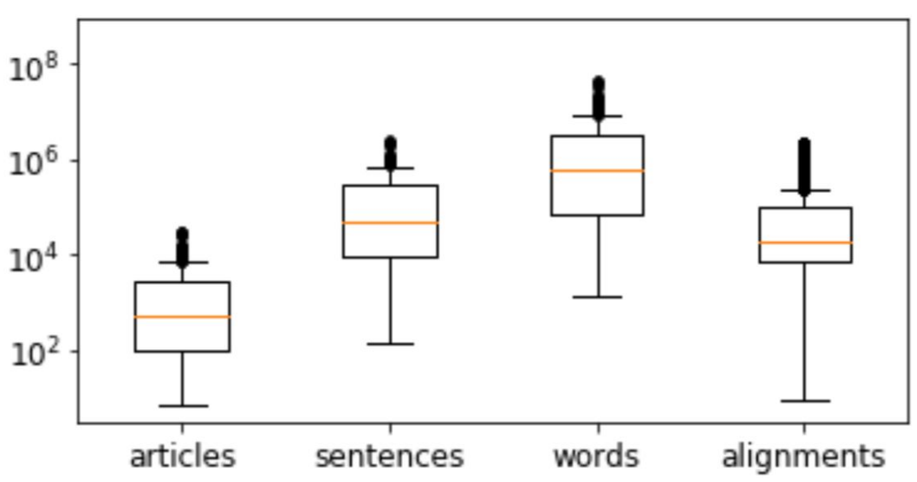
JW300 Dataset

- crawl of jw.org, mainly *Watchtower* and *Awake!*
- plenty of cleanup done, one sentence per line, sentence alignments for all pairs
- particularly good at covering low-resource pairs

- License: **CC-BY-NC-SA**-ish

languages covered	343
language datasets	417
aligned pairs of languages	54,376

	μ	σ
articles	3,202.34	± 5,946.68
sentences	261,573.37	± 464,343.05
tokens	3,544,039.82	± 7,472,321.78
alignments	92,111.61	± 176,563.25



Totals

300+ languages 1.33 M articles
109 M sentences 1.48 B tokens

Informing Bias Studies?

- What should we (not) learn from this dataset?

Experiments

- cross-lingual word embeddings

	EN	ET	HR	MR	MT
EN	–	0.280	0.254	0.0	0.001
ET	**0.314**	–	0.302	0.001	0.0
HR	**0.269**	**0.334**	–	0.002	0.0
MR	**0.094**	**0.144**	**0.112**	–	0.001
MT	**0.131**	**0.206**	**0.164**	**0.141**	–

- part-of-speech projection