Multilingual Open Text: Public Domain News in 44 Languages

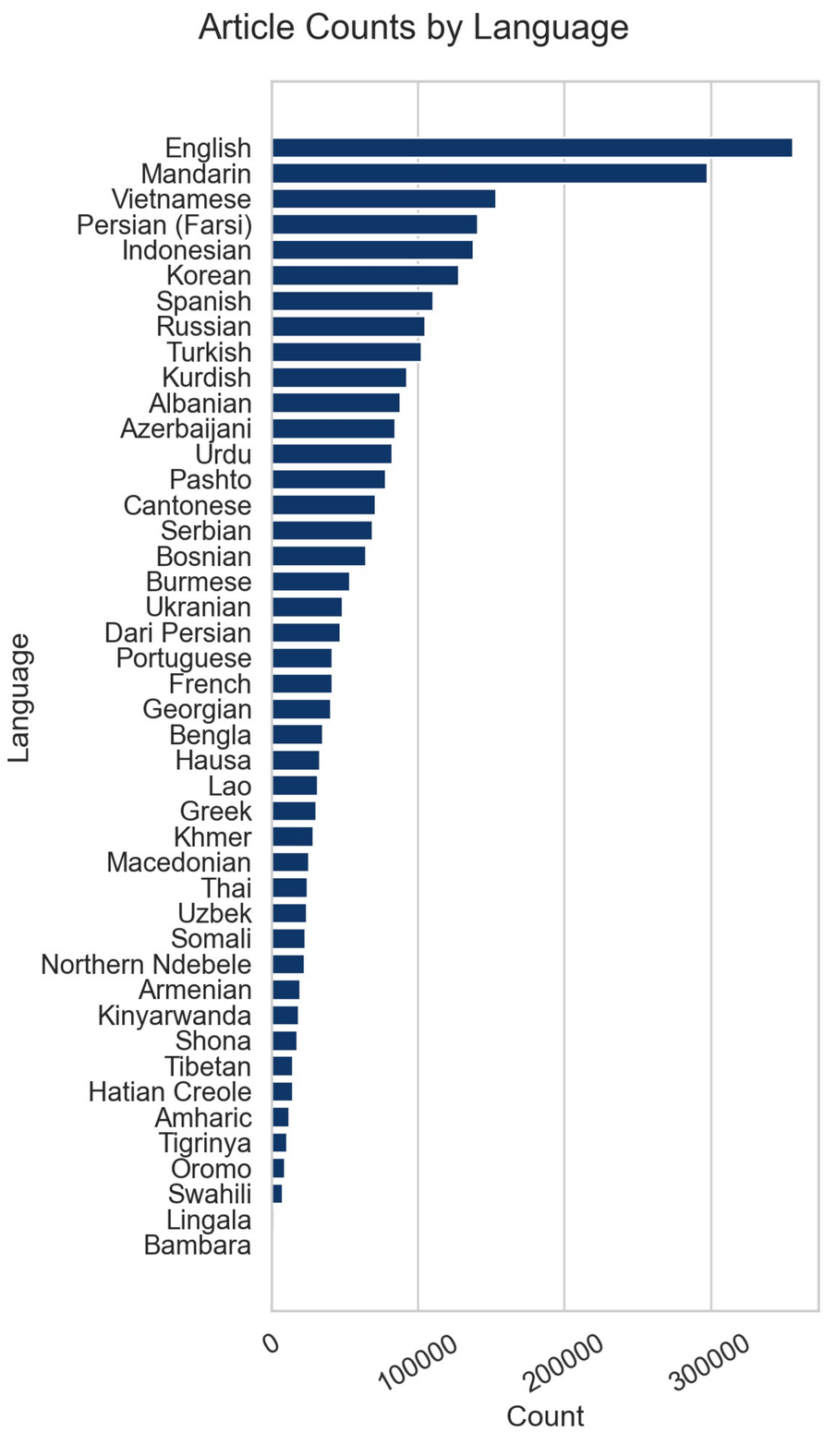
Chester Palen-Michel, June Kim, Constantine Lignos
Brandeis University

Introduction

- New corpus in 44 languages, many lower-resourced
- Public domain news text from Voice of America
  - Independent journalism (Voice of America, 2021)
- 2.7 million news articles
- 1 million short snippets
  - Photo captions, video descriptions
- Collected from Voice of America 2001-2022
- Source material in public domain, collection is CC BY 4.0
- Regularly updated as new documents are published
- Code and data releases available on GitHub: https://github.com/bltlab/mot

Corpus Content

- Metadata includes:
  - Title, authors, time published, time retrieved, time modified, URL, keywords, and section
- Organized by content type:
  - News articles
  - Captions for audio, photo, and video
- Higher resource languages in MOT are frequently region-specific. For example:
- Zimbabwe-focused English
- Mozambique-focused Portuguese
- Africa-focused French



Coverage of LRLs

- MOT provides data for many lower-resourced languages
- MOT has higher character counts than Wikipedia for 13 languages
- No Wikipedia for Northern Ndebele or Dari Persian

Lang.	Wikipedia	MOT
hau	37,141,190	38,341,381
khm	34,048,132	93,948,921
kin	3,822,464	23,881,242
lao	5,999,270	61,419,714
lin	1,502,089	1,744,378
nde	0	31,600,251
orm	2,257,827	10,469,043
prs	0	67,421,867
pus	37,683,579	127,570,695
sna	6,606,352	29,132,817
som	12,018,769	18,244,956
sqi	157,576,602	181,583,961
tir	152,456	7,645,809

Table 1: Character counts for Wikipedia and MOT for languages where MOT provides more characters

Comparison to Other Corpora

- Unlike LDC's LORELEI corpora (Tracey and Strassel, 202), no cost
- More representative of modern texts than Bible
- Although smaller, MOT includes languages that OSCAR (Ortiz Suárez et al., 2021) does not have (Cantonese, Dari, Hausa, Kinyarwanda, Lingala, Northern Ndebele, Oromo, Shona, and Tigrinya)
- JW300 (Agič and Vulič, 2019) also scraped one site and contained lower-resourced languages, but data is unavailable due to copyright
- Caswell et al. (2021) found issues with quality control of large web-scraped corpora

Acknowledgments

We thank the early adopters who used preliminary versions of this corpus and offered feedback. This work was funded by a Brandeis University Provost Research Grant.

References

Agič, Z. and Vulič, I. (2019). JW300: A widecoverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy, July. Association for Computational Linguistics.

Caswell, I., Kreutzer, J., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., et al. (2021). Quality at a glance: An audit of web-crawled multilingual datasets. arXiv preprint arXiv:2103.12028.

Ortiz Suárez, P. J., Romary, L., and Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1703–1714, Online, July. Association for Computational Linguistics.

Tracey, J. and Strassel, S. (2020). Basic language resources for 31 languages (plus English): The LORELEI representative and incident language packs. In Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing
for Under-Resourced Languages (CCURL), pages 277–284, Marseille, France, May. European
Language Resources association.

Voice of America (2021). VOA and the firewall — Law for more than 40 years. https://docs.voanews.eu/en-USINSIDE/2019/07/02/a2cdade1-ffb3-41b5-a086-2a09861ae452.pdf