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

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

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

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

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

Comparison to Other Corpora

• Unlike LDC's LORELEI corpora (Tracey and Strassel, 2020), 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


