Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2080–2089
Marseille, 20-25 June 2022

Multilingual Open Text Release 1: Public Domain News in 44 Languages

Chester Palen-Michel*, **June Kim***, **Constantine Lignos**
Michtom School of Computer Science
Brandeis University
{cpalenmichel, junekim, lignos}@brandeis.edu

Abstract

We present a Multilingual Open Text (MOT), a new multilingual corpus containing text in 44 languages, many of which have limited existing text resources for natural language processing. The first release of the corpus contains over 2.8 million news articles and an additional 1 million short snippets (photo captions, video descriptions, etc.) published between 2001–2022 and collected from Voice of America's news websites. We describe our process for collecting, filtering, and processing the data. The source material is in the public domain, our collection is licensed using a creative commons license (CC BY 4.0), and all software used to create the corpus is released under the MIT License. The corpus will be regularly updated as additional documents are published.

Keywords: multilingual corpora, text data, low resource NLP, open access text

1. Introduction

This work describes the first release of Multilingual Open Text (MOT), a collection of permissively licensed texts created with a goal of improving the amount of high-quality text available for lower-resourced languages.

MOT Release 1 consists of data collected from Voice of America (VOA) news websites. Our broader goal is a corpus of open access multilingual text, and we plan to include data from other sources in future releases. As part of the development of this corpus, we created infrastructure to continue to scrape new documents as they are published in order to provide subsequent releases with newly published and updated documents. We have been using this infrastructure for several months to expand the corpus. The corpus contains documents in many different languages, many of which are lower-resourced.

In this paper, we explain our process for collecting, filtering, and processing the data from VOA news websites in multiple languages and describe the resulting corpus. In Section 2, we motivate the need for this corpus and compare with similar lower-resourced language dataset creation efforts. In Section 3, we describe the content of MOT. In Section 4, we detail our process for creating the corpus. Finally, in Section 5, we discuss limitations and future directions. The corpus is available via GitHub.[1]

2. Related Work

A multilingual collection of unlabeled text can be useful for many tasks, especially for lower-resourced languages with limited freely-available text. An unlabeled non-parallel corpus is typically the starting point for further annotation and dataset creation work. Much of modern NLP relies on either pre-trained static or contextual word embeddings; in either case, these methods rely on large quantities of text data, which lower-resourced languages lack.

Even with the existence of multilingual Transformer models, like multilingual BERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020), unlabeled data from lower-resourced languages can be useful for adaptation of these models (Adelani et al., 2021; Pfeiffer et al., 2020). It is also possible to train a multilingual Transformer model without relying heavily on higher-resourced languages (Ogueji et al., 2021).

There have been plenty of other works which have scraped news data for lower-resourced languages (Adelani et al., 2021; Niyongabo et al., 2020). Adelani et al. (2021) that also include partial scrapes of sections of VOA news sites. Gezmu et al. (2021) used random samples of VOA news sites to create a spelling correction corpus for Amharic. Unlike these data collection efforts, MOT intends to include a complete collection of VOA's documents rather than just enough data to meet the goals of a specific annotation effort. Our resulting corpus also preserves metadata for each document which was discarded by other datasets.

There are a number of other existing resources that can be used as unlabeled data for lower-resourced languages. The DARPA LORELEI program (Strassel and Tracey, 2016; Tracey et al., 2019; Tracey and Strassel, 2020) produced datasets for a number of lower-resourced languages. However, these datasets require payment or an LDC subscription which can be prohibitively expensive for speakers of those languages to access. At the time of publication—over six years after the start of the program—many of the datasets planned for publication have not yet been released.

Many text collections for lower-resourced languages focus on parallel text for the purposes of machine translation. The OPUS website hosts a number of paral-

*Equal contribution.
[1] `https://github.com/bltlab/mot/`

lel text datasets and related tools (Tiedemann, 2012). These parallel text datasets can also be treated as unlabeled monolingual text.

Among its many sources, OPUS contains data from the Christian Bible. While the Christian Bible has been translated into more than 1,000 languages, it covers a very narrow domain that is not representative of most modern texts, is often translated into more archaic forms of each language, and reflects the perspective of its religious content.

JW300 (Agić and Vulić, 2019) is a corpus containing data in 300 different languages. It was extracted from jw.org, the website of the Jehovah's Witnesses (Watch Tower Bible and Tract Society of Pennsylvania). While JW300 has been a useful resource for lower-resourced NLP, at the time of writing, it is not currently available due to it being distributed without permission from the copyright holders. While we began work on MOT before JW300 became unavailable, the challenges of working with restrictively licensed source materials were one of the many factors that motivated us to create MOT.

There are also a number of multilingual corpora created from web-crawls such as Paracrawl (Esplà et al., 2019; Bañón et al., 2020), CC-aligned (El-Kishky et al., 2020), WikiMatrix (Schwenk et al., 2021), and OSCAR (Ortiz Suárez et al., 2020).

These web-crawled datasets tend to have a larger number of languages and larger numbers of documents. While OSCAR, for example, contains more documents and a higher number of languages, MOT contains data for some languages that OSCAR does not cover such as Cantonese, Dari, Hausa, Kinyarwanda, Lingala, Northern Ndebele, Oromo, Shona, and Tigrinya. Multilingual Open Text does not intend to compete with the size of these web-scraped corpora. Instead, MOT aims to be a reliable scrape for particular established, edited, and permissively licensed data sources. Web-scraped corpora can have issues with quality control as described in Caswell et al. (2021). While MOT covers fewer languages than many of these web-crawled corpora, it is more carefully curated and aims to avoid many of the pitfalls present in these larger-scaled corpora.

MOT can also be used to build better language identification models to help create or improve larger scale corpora.

3. Dataset Description

3.1. Source: Voice of America Online News

Background. VOA was founded in 1942 and produces content for digital, television, and radio platforms in more than 40 languages. It is the largest U.S. international broadcaster and has a weekly audience of an estimated 300 million people (Voice of America, 2021a). Because VOA's content is produced by employees of the United States government, it is in the public domain under U.S. federal law (17 U.S.C. § 105). VOA's copyright statement in their terms of use also explicitly states that all content produced by VOA is in the public domain (Voice of America, 2016).

All documents not in the public domain were filtered out of this corpus. The VOA copyright statement specifies that VOA has a license with the Associated Press (AP) to use AP content which is not in the public domain. Although the VOA copyright statement does not explicitly mention them, we identified content written by Agence France-Presse (AFP) and Reuters appearing on VOA news websites. We used automated methods to ensure that we did not include any articles from AP, AFP, and Reuters in our corpus.

Independent Journalism. Because VOA is funded by a government, it is worth discussing its independence as a news source and accordingly, the ethical considerations of using it in a corpus. VOA maintains independence from U.S. political influences through the 1994 U.S. International Broadcasting Act, which prohibits any U.S. government official from interference in the objective reporting of news (Voice of America, 2021b). The VOA's journalistic code also requires accuracy, balance, fairness, and context in documents. For example, the code requires all staff who prepare content to not use negative terms to describe persons or organizations unless those individuals use those terms to describe themselves (Voice of America, 2021a). These rules and standards suggest that the VOA operates independently, and thus a corpus derived from VOA content should be similar in its biases to corpora derived from other newswire sources, none of which are free of perspective or bias.

3.2. Corpus Contents

This dataset contains paragraph-segmented data collected from 51 VOA news websites in the following 44 languages: Albanian, Amharic, Armenian, Azerbaijani, Bambara, Bangla, Bosnian, Burmese, Cantonese, Dari, English, French (African), Georgian, Greek, Haitian Creole, Hausa, Indonesian, Khmer, Kinyarwanda, Korean, Kurdish, Lao, Lingala, Macedonian, Mandarin Chinese, Northern Ndebele, Oromo, Pashto, Persian (Farsi), Portuguese (African), Russian, Serbian, Shona, Somali, Spanish, Swahili, Thai, Tibetan, Tigrinya, Turkish, Ukrainian, Urdu, Uzbek, and Vietnamese. As noted, the French and Portuguese data is written primarily for African audiences.

The counts of articles for each language are given in Figure 1. While we have released the Bambara data for completeness, it contains essentially no articles, only short descriptions of other content (for example, photo captions, descriptions of audio stories, etc.). This is largely due to how new the inclusion of Bambara is to VOA.[2] Currently the focus for the Bambara section of VOA is on radio and multimedia, not news articles.

As shown in Figure 2, the corpus at the time of writing is comprised of articles published starting in 2001 up

[2] `https://www.insidevoa.com/a/6241315.html`

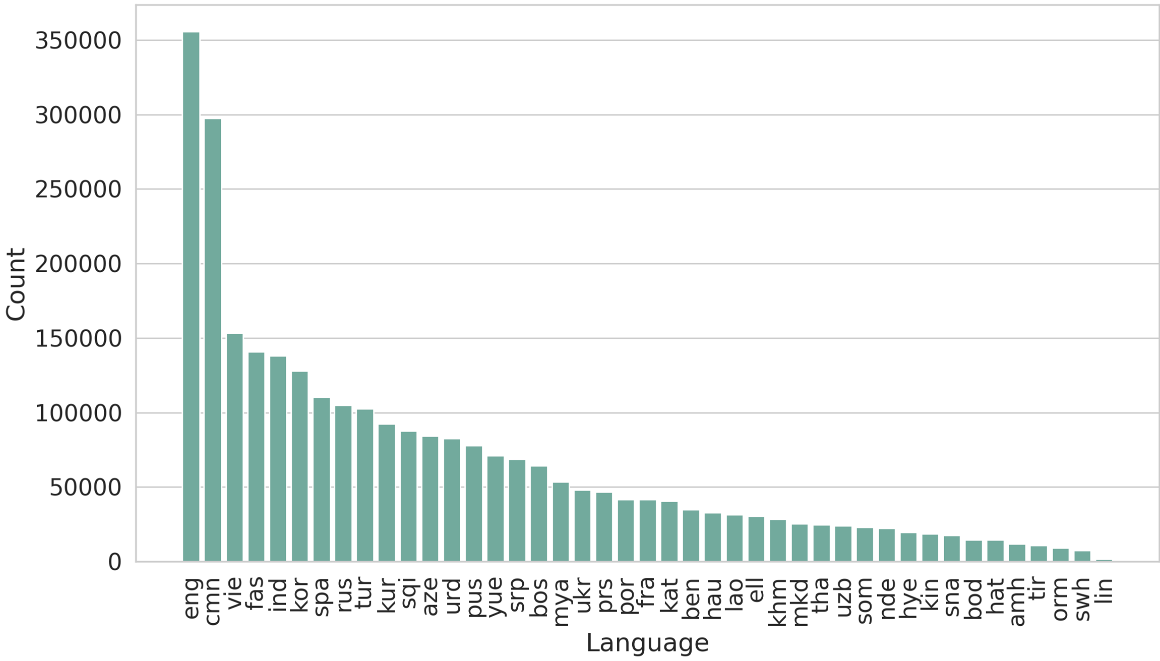

Figure 1: Counts of news articles in MOT by language using ISO 639-3 codes

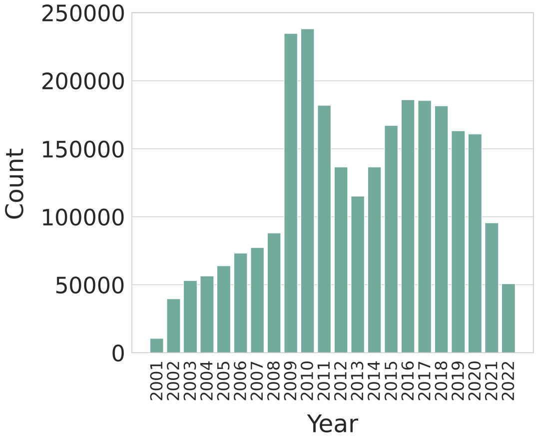

Figure 2: Counts of news articles in MOT by year

until May 1, 2022.[3] We do not know why there is such a large quantity of documents from 2010 or relatively few in 2021 compared to 2020; we suspect these abnormalities have more to do with how content may be subdivided into documents than changes in the overall amount of content. As more articles are published, we will continue to increase the size of the corpus.

The corpus is organized by VOA site and further organized by content type. Some languages in VOA are further divided into separate domains. For example, English includes VOA News (global news), VOA Zimbabwe, Editorials, and an English Learning site. Pashto, Kurdish, and Khmer also have more than one domain, where the distinction is typically a differing region or dialect (for example, Sorani and Kurmanji for Kurdish). The content types that we encountered in VOA pages' metadata were as follows: article, audio, video, photo, poll, quiz, index, author, schedule, subscribe, upload, account, and comment.

We focus on extracting data of content type article,[4] which is a typical news article. However, we also include audio, video, and photo pages as they contain some usable text data in the form of titles, short captions, or descriptions. The content types audio[5] and video[6] includes documents associated with audio and video media. The content type photo[7] includes documents that mainly include a series of captioned images. The counts of documents in each content type can be seen in Table 1[8]. Most languages have more content

[3]Documents with a timestamp prior to 2001 in the `time_published` field were removed from the corpus. This includes 4 articles dated as 1899, 1900, 1997, and 1998 whose timestamps we believe to be incorrect.

[4]Example: https://www.voanews.com/a/2020-usa-votes_bidens-cabinet-picks-include-some-firsts/6198990.html

[5]Example: https://www.voanews.com/t/60.html

[6]Example: https://www.voanews.com/a/episode_nuclear-power-cautiously-embraced-bidens-green-goals-4711476/6117084.html

[7]Example: https://www.voanews.com/a/2808902.html

[8]These reflect our best counts of the data, but these change regularly as new data is scraped and data issues are addressed.

Language	Code	Article	Audio	Photo	Video	All
Albanian	sqi	87,854	4,986	230	16,326	109,396
Amharic	amh	11,990	9,429	220	1,818	23,457
Armenian	hye	19,671	0	63	6,938	26,672
Azerbaijani	aze	84,459	1,658	1,138	11,553	98,808
Bambara	bam	1	8,560	13	1,519	10,093
Bengla	ben	34,940	3,515	19	775	39,249
Bosnian	bos	64,267	7	457	10,192	74,923
Burmese	mya	53,456	9,540	467	18,309	81,772
Cantonese	yue	71,162	19,433	438	16,378	107,411
Dari	prs	46,885	18,423	239	6,334	71,881
English	eng	355,821	188,143	1,561	8,594	554,119
French (African)	fra	41,570	38,491	491	10,796	91,348
Georgian	kat	40,572	5,905	294	5,762	52,533
Greek	ell	30,444	134	26	64	30,668
Haitian Creole	hat	14,478	8,812	300	7,012	30,602
Hausa	hau	32,957	15,873	1,167	2,647	52,644
Indonesian	ind	138,121	83,092	1,400	17,786	240,399
Khmer	khm	28,468	8,916	476	3,161	41,021
Kinyarwanda	kin	18,697	10,322	271	503	29,793
Korean	kor	127,867	1,546	304	1,108	13,0825
Kurdish	kur	92,335	15,640	1,614	7,316	116,905
Lao	lao	31,619	3,519	230	943	36,311
Lingala	lin	1,744	3,026	16	1,471	6,257
Macedonian	mkd	25,500	2	94	4,775	30,371
Mandarin	cmn	297,587	37,060	1,269	16,977	352,893
Northern Ndebele	nde	22,516	5,530	43	3,379	31,468
Oromo	orm	9,225	324	82	513	10,144
Pashto	pus	77,769	46,526	313	16,685	141,293
Persian	fas	140,724	0	1	0	140,725
Russian	rus	104,817	631	456	12,507	118,411
Portuguese (African)	por	41,620	4,266	458	6,170	52,514
Serbian	srp	68,805	173	164	6,476	75,618
Shona	sna	17,594	7,589	10	2,858	28,051
Somali	som	23,192	14,788	194	202	38,376
Spanish	spa	110,428	3,880	44	2,090	11,6442
Swahili	swh	7,388	10,361	458	5,697	23,904
Thai	tha	24,732	7,930	133	1,278	34,073
Tibetan	bod	14,715	22,964	4	7,719	45,402
Tigrinya	tir	10,774	2,359	182	1,094	14,409
Turkish	tur	102,560	168	745	17,560	121,033
Ukranian	ukr	48,297	25	639	16,963	65,924
Urdu	urd	82,642	3,540	2,459	12,724	101,365
Uzbek	uzb	24,129	6,676	2,736	10,083	43,624
Vietnamese	vie	153,287	7,594	674	20,811	182,366
Total		2,837,679	641,356	22,592	323,866	3,825,493

Table 1: Counts of documents by content type and ISO 639-3 codes for each language included in MOT.

of type article, yet some languages, like Swahili, may have more of a focus on radio, and thus contain more audio files.

For some languages, we have few or no documents for certain content types like audio or video. This is typically not because there is no audio or video in that language, but because the audio and video in that language did not contain captions from which to extract text data, the captions were so short that they were unlikely to represent meaningful content, or the captions were in an unexpected format that caused our extraction to miss it. Pages of content type poll, quiz, index, author, schedule, subscribe, upload, account, and comment are not included in our final data release. These typically contained little or no data, were more complicated to extract from, or in the case of index, duplicated descriptions from other pages where we were able to perform more complete extractions.

All content provided in this corpus is text, so for media like photos and videos, the data is the text description or a caption; it is not extracted from the media itself. Paragraph breaks from the original HTML are preserved and documents are represented as lists of paragraphs, which contain lists of sentences, which contain lists of tokens.

File names sometimes contain abbreviated headlines, but occasionally the headline used for the file name is in a different language than the actual headline and text appearing in the document. This is likely the result of editorial errors and may reflect that the document was adapted or translated from a document in another language. Each VOA domain is provided as a separate .tgz file in our release with subdirectories for different content types like article, audio, video, etc.

All languages are identified using ISO 639-3 codes. Each file contains the following fields:

- `filename`: the name of the file derived from the URL
- `url`: the URL from which the document was retrieved
- `url_origin`: the sitemap from which the URL was retrieved
- `content_type`: the type of content (e.g., article, audio, photo, video) of the document
- `site_language`: the language of the VOA site
- `time_published`: the timestamp for when the document was published
- `time_modified`: the timestamp for when the document was last modified
- `time_retrieved`: the timestamp for when the document was retrieved from the sitemap
- `title`: the title of the document
- `authors`: the author(s) of the document
- `paragraphs`: the text extracted from the document
- `n_paragraphs`: the number of paragraphs in the document
- `n_chars`: the number of characters in the document
- `cld3_detected_languages`: the language(s) identified by CLD3 from the full extracted text of the document (see Section 4.3)
  - `language`: the language outputted by CLD3
  - `probability`: the probability that the language identified is correct (passed directly from CLD3)
  - `is_reliable`: if probability is above 0.7 (passed directly from CLD3)
  - `proportion`: the proportion of the text identified as the language (passed directly from CLD3)
- `predicted_language`: the language that we predict that the document is in, based on rules that take into account the site, the CLD3 predictions, and whether the site language is supported by CLD3
- `keywords`: the terms relating to the text content of the document
- `section`: the subpage the document falls under

These additional fields are included only for subset of languages:

- `sentences`: the text extracted from the document segmented into sentences
- `n_sentences`: the number of sentences in the document
- `tokens`: the text extracted from the document segmented into tokens
- `n_tokens`: the number of tokens in the document
- `parallel_english_article`: the URL for the English document from which the current document was translated from into the site language (this currently only appears in Lao articles)

3.3. How Low Resourced?

While there is no single way of classifying lower-resourced languages due to the large number of intersecting factors that contribute to such a designation, Joshi et al. (2020) created a taxonomy of resource availability for languages based on the amount of labeled and unlabeled data. The scale goes from 0 (lowest resources) to 5 (highest resources). Although the taxonomy is an oversimplification of the state of resources for a language since there are many more dimensions (domain, task, medium, register, etc.) by which data can be categorized, it can still provide some sense of low-resourced-ness.

Of the 44 languages included MOT Release 1, only 4 are considered "winners" at level 5. 16 of the languages are classified as level 1, "scraping-bys," which is described as having essentially no labeled data and very little existing unlabeled data. MOT also includes 3 languages classified as level 0, "left-behinds," Haitian Creole, Northern Ndebele, and Dari.

Another way of evaluating the low-resourced-ness of MOT is to compare with Wikipedia. Because Wikipedia is a commonly used resource for multilingual text, languages that have poor representation in Wikipedia could be considered more lower-resourced. We compare MOT articles to Wikipedia articles by counts of characters in each dataset in Table 2. We use character counts since tokens are dependent on the quality of the tokenizer and lower-resourced languages may not have adequate tokenization. As seen in Table 2, MOT contains more data than Wikipedia in 13 languages, demonstrating MOT's potential value in providing more unlabeled text data for lower-resourced languages.

While it is true that the highest-resourced languages such as English or French contained in MOT initially do not appear to be much of a contribution when plenty

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 2: Counts of characters in Wikipedia and MOT for lower-resourced languages where MOT provides a higher count

of resources exist for these languages, we include them for completeness and because much of the text in the VOA documents has a regional focus that may not be present in existing datasets.

For example, portions of the English data focus on news in Zimbabwe while a portion of the Portuguese data is centered around Mozambique. This can matter for annotation projects that may wish to use monolingual data that is region-specific.[9] While there is existing Mozambique-focused Portuguese data available from Davies and Ferreira (2006), we are not aware of any usable data for Zimbabwe-focused English. We were able to identify one corpus focusing on Zimbabwe English textbooks, but as it was stored on magnetic tapes, we were not able to locate a copy (Louw and Jordán, 1993).

4. Data Collection and Processing

4.1. Scraping VOA

While this work is not the first to scrape text data from Voice of America, it is to the best of our knowledge the most thorough and complete scraping effort of the text contained on the Voice of America collection of websites. The data collection process starts with manually creating a list of all the different VOA domains along with their ISO 639-3 language codes. We then use the list of VOA domains to automatically get all the URLs from each site's sitemap. The current release includes documents retrieved from sitemaps between June 16, 2021 and May 1, 2022.

When scraping a page, we extract the title, description, keywords, and author(s) from the HTML meta tags. We also attempt to collect the canonical link, publication date, modification date, and content type for each page. In addition to the sitemaps, we used the Internet Archive's Wayback CDX Server API[10] to collect URLs for each domain. Of the URLs we retrieved using the Internet Archive, the vast majority were duplicates. In the case where the pages were of content type article, only 5 Thai pages and only 3 French pages were not already retrieved through the sitemaps. While this process of using the Internet Archive in addition to the sitemaps did not produce meaningful gains in content, it did help us to verify that we are not missing any easily retrievable content from the sitemaps.

The scraped pages are maintained in a database and we compare against existing pages' URLs and canonical links in order to de-duplicate and use the most recent version of a page. Our collection effort of VOA data differs from other efforts in that we regularly do an updated scrape. We have scraped periodically[11] since beginning our collection effort in summer 2021. The gains in numbers of previously unseen URLs in roughly a month's time varies from a few hundred to about 2,000 for languages other than English. The Greek section of VOA is no longer being updated, so there are never new URLs for that section. We also notice some URLs are no longer found in the sitemaps between our scraping efforts; however, the number of URLs lost is quite small.

For example, only 720 URLs went missing in Persian between December 1, 2021 and January 1, 2022, which is relatively small compared to the 141,060 documents we extracted. For the same time period, 25 languages had not lost any URLs in the sitemaps. We can also report anecdotally that many of these lost URLs are either video clips with little or no caption content or are sites that were updated and have a newer URL, which we attempt to de-duplicate if a canonical link was present.

4.2. Extracting Text from HTML Documents

We now turn to the process of extracting text data from the raw HTML scraped from VOA. All relevant text content from each document is extracted and paragraph breaks from the HTML are maintained in the output. However, not all data that is extracted from paragraph tags or the usual div tags is actually part of the document content. We remove repetitive and meaningless content, such as user comments and sentences that consist of the equivalent of "login." If the page contains no valid text, it is not included in the output.

The `filename` we create is derived from the URL and includes everything following the top-level domain. If the name of the file is too long, the `filename` is shortened to only the last 100 characters.

[9] As an example, one early adopter of our corpus wished to translate news data focused on Mozambique from Portuguese into eMakhuwa to create a parallel corpus.

[10] `https://github.com/internetarchive/wayback/blob/master/wayback-cdx-server/README.md`

[11] Re-scraping occurs roughly once a month.

4.3. Language Identification and Filtering

Not all of the documents in VOA are consistently in one language. While code switching exists, most of the mixed language use that we observed in the corpus were sentences that were translations of other content in the document rather than instances of natural code switching. Unfortunately, these translated portions of such documents did not appear to be systematic enough to extract parallel text in most cases. In some cases, this is because the document is a translation, but the captions remain in the reported language. In other cases, the document may contain the English translation or may be a part of VOA's language learning site that was miscategorized. We attempt to filter out heavily multilingual text along with documents that erroneously contain mostly English despite claiming to be written in another language.

CLD3. We use CLD3 for our language ID in the filtering process. Compact Language Detector version 3 (CLD3) (Salcianu et al., 2016) is a neural network model for language identification that supports over 100 languages and scripts. The model outputs the languages identified as BCP-47-style language codes, along with its corresponding probability, reliability, and proportion (see Section 3.2 for more information about these fields). CLD3 does not support the following languages in MOT: Azerbaijani, Bambara, Cantonese, Dari, Kinyarwanda, Lingala, Northern Ndebele, Oromo, Tibetan, and Tigrinya. Because these languages are unsupported, we do not use the language ID predictions for our `predicted_language` field and instead rely on VOA's reported language based on which domain the site is from. We do include the main CLD3 prediction information, but end users should take note that certain languages are likely to be misrecognized. For example, Tigrinya is regularly classified as Amharic by CLD3 since it is not supported.

Filtering Process. CLD3 was used to identify the language present in the extracted text with a maximum of 5 languages. This was used to determine the predicted language of the document. We filter at the paragraph level and at the document level. At the paragraph level, we filter only for confidently English paragraphs in non-English sections of VOA. If the probability is greater than 0.7 and the proportion of the paragraph is more than 0.25, the English paragraph is discarded. Because URLs in text tend to get identified as English by CLD3, this also helps to filter out URLs. This paragraph level filtering is useful as there are some documents that will be almost entirely in one language with just one or a few paragraphs in English. Typically, these paragraphs in English are also redundant with the main language of the document.[12] It is also common for the English contamination to be a translation of just a few quotes in the document.[13]

At the document level, we also run language ID on the original text before paragraph level filtering. If CLD3 is confident in one language, the predicted language is assumed to be either the original sitemap language or English as CLD3 does not predict all of the languages encountered in the corpus. If CLD3 is confident that the majority of the document is either English in a non-English section, or non-English in an English section, the document is filtered out. If CLD3 has identified multiple languages with a probability above 0.9 and a proportion above 0.05, the predicted language is listed as "mul." All documents include a prediction of the language expected from the output of CLD3. Every document is predicted to be written in the site language unless CLD3 has identified more than one language from the text ("mul") or CLD3 has identified only English present in the document ("eng"), in which case the document is not included.

4.4. Sentence Segmentation and Tokenization

Segmentation. We primarily use Ersatz (Wicks and Post, 2021) for sentence segmentation; however, off-the-shelf monolingual models provided for Ersatz do not cover all of the languages in MOT. We attempted to use the multilingual model provided by Ersatz, but it had unsatisfactory performance in some languages. In Swahili, it failed to segment the abbreviation for *doctor*, *Dkt.* correctly. We also noticed some instances of periods after first initials being treated as sentence boundaries in Greek, likely because Ersatz was not trained on any language using the Greek alphabet. It also did not contain any Ge'ez script punctuation as candidates for sentence splits and was therefore unusable for Amharic or Tigrinya. Thai and Lao, which do not have sentence ending punctuation, also created challenges. Because the multilingual segmentation model had sub-optimal performance for languages it was not trained on, we have chosen only to release sentence breaks and tokenization for those languages where we could provide more reliable segmentation.

We used PyThaiNLP (Phatthiyaphaibun et al., 2016) for Thai and `amseg` (Yimam et al., 2021) for Amharic and Tigrinya. `amseg` is a rule-based Amharic segmenter, but as it is based on whitespace and Ge'ez script punctuation, we used it for Tigrinya in addition to Amharic. Parsivar (Mohtaj et al., 2018) was used for Persian, `khmer-nltk` for Khmer, LaoNLP[14] for Lao, and `razdel` for Russian. We also use Stanza (Qi et al., 2020) for Armenian, Burmese, Greek, Indonesian, Korean, Portuguese, Serbian, Ukrainian, Urdu, and Viet-

[12] https://www.voaswahili.com/a/netanyahu-aipongeza-marekani-kwa-usimamizi-wa-kurejesha-mahusiano-kati-ya-israeli-na-sudan/5634218.html

[13] https://www.voaswahili.com/a/ndege-ya-ethiopian-airlines-imeanguka-na-juhudi-za-kuitafuta-zaendelea/4822036.html

[14] https://github.com/wannaphong/LaoNLP

Language	Documents	Sentences	Tokens
amh	23,457	91,051	1,960,739
aze	98,808	644,156	N/A
bos	74,923	577,114	N/A
cmn	352,893	2,177,530	130,214,418
ell	30,668	155,284	6,090,066
eng	554,119	4,537,686	193,129,912
fas	140,725	871,887	35,929,609
fra	91,348	507,058	19,966,932
hat	30,602	100,558	N/A
hau	52,644	244,043	N/A
hye	26,672	150,086	5,064,730
ind	240,399	1,245,778	38,383,509
khm	41,021	392,724	19,027,222
kin	29,793	119,298	N/A
kor	130,825	1,516,790	41,548,365
lao	36,311	532,944	12,058,686
lin	6,257	18,757	N/A
mkd	30,371	245,127	N/A
mya	81,772	657,459	36,006,802
nde	31,468	211,156	N/A
orm	10,144	57,187	N/A
por	52,514	427,612	13,864,438
prs	71,881	461,203	14,633,719
pus	141,293	838,726	N/A
rus	118,411	1,051,201	51,451,892
sna	28,051	189,093	N/A
som	38,376	131,501	N/A
spa	116,442	911,685	33,352,028
sqi	109,396	793,622	N/A
srp	75,618	618,884	26,544,508
swh	23,904	63,761	N/A
tha	34,073	262,953	9,428,506
tir	14,409	76,283	1,784,820
tur	121,033	861,882	31,419,370
ukr	65,924	363,540	17,232,122
urd	101,365	986,220	40,805,126
uzb	43,624	314,141	N/A
vie	182,366	1,138,882	59,843,930
yue	107,411	70,1411	34,730,065
Total	3,561,311	25,246,273	874,471,514

Table 3: Counts of documents, sentences, and tokens for languages with sentence segmentation and tokenization

namese. We trained custom Ersatz models using paragraph breaks from MOT for the remaining languages.

As Wicks and Post (2021) point out, there tends to be a lack of reliable test sets for sentence segmentation, so we have not yet independently vetted the performance of these segmenters. For languages in which we do not yet have satisfactory sentence segmentation, we do not provide sentence breaks. In Table 3, we provide counts of sentences and tokens for the languages where we are able to provide segmentation and tokenization.

Tokenization. We used spaCy (Honnibal et al., 2020) for tokenization in English, Cantonese, French, Mandarin Chinese, Russian, Spanish, and Turkish. PyThaiNLP (Phatthiyaphaibun et al., 2016) is used to tokenize Thai, and `amseg` (Yimam et al., 2021) to tokenize Amharic and Tigrinya. `khmer-nltk` (Hoang, 2020) was used for Khmer tokenization. Stanza (Qi et al., 2020) is also used for tokenization in the same languages it is used for sentence segmentation. We hope to provide more robust tokenization and segmentation in future releases.

5. Limitations and Conclusion

Extracting text from HTML from a complex network of sites like VOA is non-trivial, and although we have done our best to ensure complete, clean extractions, we expect users of this resource will discover issues.

There are still a number of languages where we do not have reliable sentence segmentation and tokenization. We would like to improve language identification to better identify documents with multiple languages, as CLD3 does not cover all of the languages in MOT. We plan to continue to increase the size of the corpus as VOA publishes more documents, and we plan to expand MOT by adding other permissively-licensed texts to expand our coverage of lower-resourced languages.

There are many ways in which MOT could be used in future work. For lower-resourced languages, MOT provides a valuable source of high-quality unlabeled text, and it could be used with minimal annotation effort to train language identification, sentence segmentation, and tokenization systems. Sections of MOT could also be used for annotation projects to create labeled data for tasks like document classification, named entity recognition, and syntactic or semantic parsing.

Because MOT includes publication time metadata, it may be possible to use MOT to create semi-parallel text. While we do not include audio or images as part of our release, others may want to make use of the included source URL and employ the captions on the photo content type for image captioning in lower-resourced languages.

We have presented a new corpus containing unlabeled text data in 44 languages, many of them lower-resourced languages for which this represents a substantial increase in the amount of available text data. The data in this corpus is in the public domain, and the corpus is positioned to grow in future releases as new documents are published. We look forward to the opportunity to further refine the extraction and increase the usefulness of MOT as speakers of the languages contained in it begin to make use of it.

6. Acknowledgments

We thank the early adopters of our resource—many of whom are members of the Masakhane community—who used preliminary releases and offered feedback. This work was supported by a 2021 Brandeis University Provost Research Grant.

7. Bibliographical References

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S. M., Gwadabe, T. R., Ezeani, I., Niyongabo, R. A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E. P., Oyerinde, S., Siro, C., Bateesa, T. S., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., MBOUP, M., Gebreyohannes, D., Tilaye, H., Nwaike, K., Wolde, D., Faye, A., Sibanda, B., Ahia, O., Dossou, B. F. P., Ogueji, K., DIOP, T. I., Diallo, A., Akinfaderin, A., Marengereke, T., and Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage 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.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, 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*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Davies, M. and Ferreira, M. (2006). Corpus do Português: 45 million words, 1300s–1900s. `https://www.corpusdoportugues.org/`.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (2020). CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online, November. Association for Computational Linguistics.

Esplà, M., Forcada, M., Ramírez-Sánchez, G., and Hoang, H. (2019). ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland, August. European Association for Machine Translation.

Gezmu, A. M., Lema, T. T., Seyoum, B. E., and Nürnberger, A. (2021). Manually annotated spelling error corpus for Amharic. Accepted to the 2nd Workshop on African Natural Language Processing (AfricaNLP 2021). `https://arxiv.org/abs/2106.13521`.

Hoang, P. V. (2020). Khmer natural language processing toolkit. `https://github.com/VietHoang1710/khmer-nltk`.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. `https://spacy.io/`.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July. Association for Computational Linguistics.

Louw, W. and Jordán, J. (1993). Corpus of Zimbabwean English at the University of Zimbabwe computer centre. *Zambezia*, 20:131–138.

Mohtaj, S., Roshanfekr, B., Zafarian, A., and Asghari, H. (2018). Parsivar: A language processing toolkit for Persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Niyongabo, R. A., Hong, Q., Kreutzer, J., and Huang, L. (2020). KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Ogueji, K., Zhu, Y., and Lin, J. (2021). Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

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.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2020). MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, November. Association for Computational Linguistics.

Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., and Chormai, P. (2016). PyThaiNLP: Thai natural language processing in Python, June. `http://doi.org/10.5281/zenodo.3519354`.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Salcianu, A., Golding, A., Bakalov, A., Alberti, C., Andor, D., Weiss, D., Pitler, E., Coppola, G., Riesa, J., Ganchev, K., Ringgaard, M., Hua, N., McDonald, R., Petrov, S., Istrate, S., and Koo, T. (2016). Compact Language Detector v3 (CLD3), October. `https://github.com/google/cld3`.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021). WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online, April. Association for Computational Linguistics.

Strassel, S. and Tracey, J. (2016). LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

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.

Tracey, J., Strassel, S., Bies, A., Song, Z., Arrigo, M., Griffitt, K., Delgado, D., Graff, D., Kulick, S., Mott, J., and Kuster, N. (2019). Corpus building for low resource languages in the DARPA LORELEI program. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 48–55, Dublin, Ireland, August. European Association for Machine Translation.

Voice of America. (2016). Terms of use and privacy notice. `https://learningenglish.voanews.com/p/6021.html`.

Voice of America. (2021a). Mission and values. `https://www.insidevoa.com/p/5831.html`.

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

Wicks, R. and Post, M. (2021). A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, Online, August. Association for Computational Linguistics.

Yimam, S. M., Ayele, A. A., Venkatesh, G., Gashaw, I., and Biemann, C. (2021). Introducing various semantic models for Amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, 13(11).