Proceedings of the 1st Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL) @LREC2022, pages 36–43
Marseille, 24 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Crawling Under-Resourced Languages – A Portal for Community-Contributed Corpus Collection

Erik Körner[1,2], Felix Helfer[2], Christopher Schröder[1], Thomas Eckart[2], Dirk Goldhahn[2]
[1]Leipzig University, Leipzig, Germany,
[2]Saxon Academy of Sciences and Humanities, Leipzig, Germany,
erik.koerner@uni-leipzig.de

Abstract

The "Web as corpus" paradigm opens opportunities for enhancing the current state of language resources for endangered and under-resourced languages. However, standard crawling strategies tend to overlook available resources of these languages in favor of already well-documented ones. Since 2016, the *Crawling Under-Resourced Languages portal* (CURL) has been contributing to bridging the gap between established crawling techniques and knowledge about relevant Web resources that is only available in the specific language communities. The aim of the CURL portal is to enlarge the amount of available text material for under-resourced languages thereby developing available datasets further and to use them as a basis for statistical evaluation and enrichment of already available resources. The application is currently provided and further developed as part of the thematic cluster "Non-Latin scripts and Under-resourced languages" in the German national research consortium Text+. In this context, its focus lies on the extraction of text material and statistical information for the data domain "Lexical resources".

Keywords: CURL, Community-Contributed, Corpus Creation, Dataset Creation, Web Crawling

1. Introduction

Despite various endeavors over the last decades to decrease the gap between well and under-resourced languages, the current situation of language documentation and the availability of language resources for the latter are still unsatisfactory. This is even acknowledged by the UN, which proclaimed the years 2022 – 2032 as the International Decade of Indigenous Languages (IDIL[1]), thereby showing that a global effort and a large variety of stakeholders are necessary for the "preservation, revitalization and promotion" of indigenous languages. Work in this area can only be fruitful and sustainable when local language communities are directly involved in all crucial parts of the process and when all general principles of responsible scientific work are considered. These include policies like the CARE Principles for Indigenous Data Governance[2] but also the general FAIR principles.

Following the "Web as corpus" paradigm (Kilgarriff and Grefenstette, 2003), the aim of the *Crawling Under-Resourced Languages portal* (CURL portal) is to collect Web-based text resources and make them publicly available for everyone. Contrary to explorative Web crawls, the source domains are provided by users via the CURL portal, thereby allowing anyone to collaborate. The resulting web crawls are then processed into datasets which include the pre-processed plain text, i.e. the raw text extracted from HTML that has been cleaned and subsequently segmented into sentences and tokens. The sentences are randomly shuffled to not allow reconstruction of the original documents (see also Appendix A.1. on copyright and license). Moreover, these datasets contain metadata, such as lists of visited Web domains, word co-occurrences and word frequencies. The gathered material is also used to constantly improve applications in Natural Language Processing (NLP), and is offered and employed by projects such as the Leipzig Corpora Collection (LCC) (Goldhahn et al., 2012).

During the last 6 years, the portal was introduced to and discussed with native speakers of different under-resourced and indigenous language communities. Such participation is crucial to obtain Web links containing under-resourced languages. Most of these links are difficult to find using standard Web crawling techniques, and often cannot be handled well by standard NLP components such as language detectors that rely on training data. The former affects – among others – .com domains where relevant material is hard to identify in case of inadequate language detection or domains that are only sparsely linked to and therefore often ignored by popular Web search engines.

The CURL portal allows anyone to contribute to the creation of digital and openly available language resources with minimal effort and a very low barrier of entry. Especially in cases where direct exchange and knowledge transfer "on site" is hard to achieve (e.g. because of organizational or financial reasons) the portal is an easy-to-use alternative.

2. Related Work

The CURL portal is being developed and maintained since 2016 and has been continuously revised and improved since then. Previous publications focused on planning and implementation (Goldhahn et al., 2016) or on presenting first use cases (Goldhahn et al., 2017)

[1] https://en.unesco.org/idil2022-2032
[2] https://www.gida-global.org/care

rather than providing a bigger picture of the portal, its acceptance by language communities and its purpose to foster availability of text datasets and lexical resources, all oh which is discussed in this work. CURL is part of the Leipzig Corpora Collection (Goldhahn et al., 2012) and is built upon its technology such as the processing pipeline for corpora creation or various forms of data access (including different web portals or web services). CURL is part of the LCC's strategy to offer large monolingual corpora for various languages; its results are therefore integrated into the LCC. Furthermore, the portal is part of the German national research consortium Text+ in the dedicated lexical cluster "Non-Latin scripts and Under-resourced languages"[3] which focuses on the creation and maintenance of lexical resources for those languages in a sustainable infrastructure.

Most work concerned with corpus collection and creation for under-resourced languages is invested by individual researchers who prepare a resource for a particular purpose or to answer a specific research question. A significant contribution to the more general collection of corpora for under-resourced languages was made by the An Crúbadán project (Scannell, 2007). Utilizing textual resources from highly multilingual sources and applying a BootCaT like approach (Baroni and Bernardini, 2004), corpora for various languages were created and extended with the help of language experts. Though typically very small, textual samples for a striking number of languages are provided.

Yet other projects are concerned with generic corpus creation without addressing the challenges of under-resourced languages such as LanguageCrawl (Roziewski and Stokowiec, 2016) which builds upon Common Crawl[4].

3. Web Portal & Corpus Creation

The central entry point for contributors is the CURL web page[5] where users can submit new URLs about a language to be processed. Moreover, users can browse a list of 285 languages, showing existing corpora, including statistics about corpora stemming from previous submissions and lists of URLs provided so far, making the whole process as transparent as possible.

New submissions only require selecting a language, identified by its ISO 639-3 code, and providing a list of URLs of web pages with text in the chosen language. In case a specific under-resourced language is not yet listed, the project can be contacted and will add it.

After submission, new jobs are run automatically. The main processing steps as show in Figure 1 include:

1. Crawling the URLs using the open-source web crawler Heritrix (Mohr et al., 2004) for up to 6 hours and being restricted to the provided domains,
2. Extracting all possible text content using jWarcEx[6] from the HTML documents,
3. Detecting the language of the text and filtering out documents not belonging to the target language,
4. Preprocessing documents into sentences with (1) sentence segmentation, (2) rule-based cleaning, (3) language separation on single sentences, and (4) sentence deduplication,
5. Merging with sentences from previous submissions and existing corpora,
6. Corpus creation using word tokenization and cooccurrence computation.

A contact email can be provided to be notified when all steps are completed. Completed corpora will be published on the CURL web page together with basic statistics about the number of word types, tokens, sentences, and sources; searchable lists of URLs and domains are also provided.

Language Detection and Separation

Aside from the problem of missing seed URLs, lack of sufficient language material obstructs language identification since in particular under-resourced languages might not be supported by existing language detectors – and training a new model is only an option if sufficient data is available. Language detection is an essential step that cannot be omitted since (a) the URLs received from the CURL portal may contain text in multiple languages and (b) might also not contain the target language at all (for example if the crawler operates during a temporary downtime of the web site). To circumvent the problem of a missing initial model, we use bible and watchtower texts, which are available to us in about 1,000 languages combined and which have been successfully applied to similar scenarios before (Brown, 2013; Agić et al., 2016). Using a character n-gram based classification model (Brown, 2013), we identify the dominant language of each crawled document, filtering out all text material that is not in the language of interest while preserving text containing foreign or loan words.

After a successful submission to the CURL portal, we can use the resulting text data and create a new language detection model for the corresponding language. As a result, we can then either replace a previous bible and watchtower model or train an improved model on all data that has been crawled so far, thereby iteratively improving the detection of under-resourced languages. For languages with no available model, manual assistance is necessary. After filtering out textual data in other well known languages using the language detection setup described above, project staff checks the remaining texts and its sources using available web documentation and common sense. This ideally results in a first model for the language and eventually in its first dataset to be made available. The same holds for jobs

[3] https://textplus.org/en

[4] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

[5] https://curl.wortschatz-leipzig.de/

[6] https://github.com/Leipzig-Corpora-Collection/jwarcex

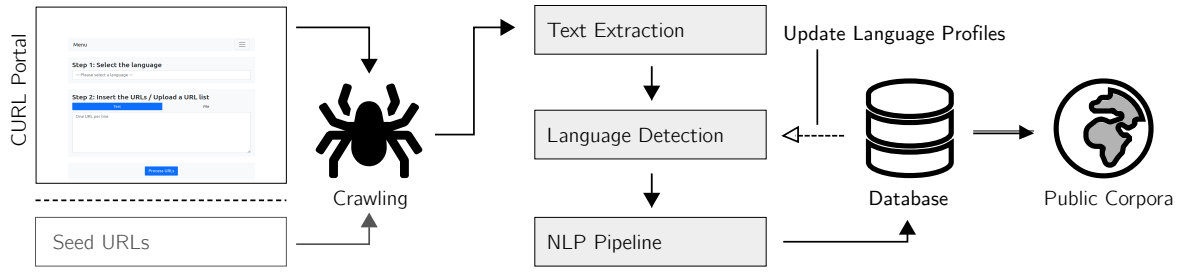

Figure 1: Overview of the full process from crawling to corpora. The lower path from "Seed URLs" to "Public Corpora" is a typical crawling setup, which is, however, infeasible for under-resourced languages with little to no URLs available. The CURL portal helps to alleviate this obstruction in a collaborative fashion. After successful completion of a new submission, the resulting database can be used to train a language detection model for detecting this language in other more general web crawls.

where despite crawling resulting in decent amounts of textual data and having a language model for the respective language, no data is classified to be of the desired language. Only by manually figuring out the reasons can this situation be resolved. Reasons can be, among others, a wrongly assigned language model or websites not actually containing the expected language.

4. Statistics

Between 2016 and 2022, 202 jobs were submitted for 134 languages. Most of these were successful and created new or augmented existing corpora.

4.1. Submissions

Some submissions unfortunately were not able to add new sentences since the corpora already contained those. 42 jobs failed due to various reasons such as crawling errors (page gone, blocked), text extraction (no content found, e.g. pages with JavaScript), or language detection/separation (e.g. page about the target language, not in the target language; e.g. text in English). If possible, it was tried to finish jobs manually, in particular when language detection failed because of insufficient models. In those cases, text segments with high confidence scores for languages such as English were filtered out and the process was continued focusing on the remaining material.

Table 1 shows submissions totaling at least 20 seed URLs and the number of sentences that could be extracted. While the number of URLs does not correlate with the number of extracted sentences after processing due to various reasons, submissions with higher numbers of URLs generally lead to text of the target language with a greater reliability than submissions with only a few (or a single) URL(s).

Submissions in 18 languages (*pes, sqi, nep, mkd, tat, glk, tam, hye, ben, tel, tgl, tgk, war, msa, ceb, uzb, bos, mal*) resulted in more than one million sentences.

Language		Submissions	URLs	Sentences
tsn	*(Tswana)*	4	8,268	28,276
ben	*(Bengali)*	5	4,043	1,200,255
zul	*(Zulu)*	12	1,731	158,644
nso	*(Northern Sotho)*	5	455	9,560
tso	*(Tsonga)*	3	330	10,571
wol	*(Wolof)*	2	311	9,988
ven	*(Venda)*	2	294	9,279
sna	*(Shona)*	2	277	48,339
xho	*(Xhosa)*	8	184	63,387
uig	*(Uyghur)*	4	123	68,736
bam	*(Bamanankan)*	6	61	10,874
ndo	*(Ndonga)*	2	58	13,495
run	*(Rundi)*	6	49	17,361
ckb	*(Central Kurdish)*	2	44	4,978
knn	*(Konkani)*	2	38	14,111
tgk	*(Tajiki)*	3	36	939,144
bcl	*(Central Bikol)*	1	35	15,726
kir	*(Kyrgyz)*	1	32	251,608
nbl	*(Southern Ndebele)*	2	20	318

Table 1: The number of sentences, URLs, and submissions for languages with at least 20 submitted URLs.

4.2. Domains and TLDs

URLs of the domain `wikipedia.org` appeared in submissions of 99 languages. They are in the top-5 based on the amount of extracted sentences for 36 languages and the sole resource for 27 languages.

However, for 37 languages no jobs with Wikipedia URLs were submitted. Nine languages of those only contain a single domain, with the languages *dyu, fon, nyn, tem, tiv* having less than 20 sentences each and *kck, bak, gom*, and *nan* only having 1k, 3k, 40k, and 77k sentences, respectively.

The spread of domains per language varies. Disregarding single-domain languages, we found that the top-5 domains cover almost all the sources of sentences for a language. Exceptions being *ben, hye, kea, kng, knn, ngl,*

Language		Proportion	Sentences
pes	*(Iranian Persian)*	47.8%	3,980,346
hye	*(Armenian)*	27.5%	376,981
sot	*(Southern Sotho)*	34.9%	3,410
knn	*(Konkani)*	15.0%	2,124
kea	*(Kabuverdianu)*	31.8%	82
snk	*(Soninke)*	46.1%	57
ngl	*(Lomwe)*	48.7%	37
kng	*(Koongo)*	46.4%	19

Table 2: Number of sentences for top-5 domains with the proportion of sentences per domain less than 50%.

pes, snk, sot, with the top-5 only amounting to less than 50% (cf. Table 2).

The top TLDs across languages are: **.org** for 106 languages, most occurrences due to wikipedia.org, **.com** (80), **.net** (26), **.edu** (14). Surprisingly often, URLs come from country-code TLDs in which the respective language is not spoken natively. These include **.de** (33), **.pl** (24), **.nl** (18), **.ru** (18), **.jp** (17), **.cn** (16). This demonstrates that language resources are often 'hidden' on international TLDs and require assistance from native speakers to be found.

4.3. Examples

Particularly successful languages were:

Rundi (*run*): After being contacted by language experts, we got 6 submissions with 49 URLs in total. Starting from 3k sentences, we currently have 17,361 sentences.

Zulu (*zul*): After being contacted by language experts, we received 11 submissions with 1730 URLs in total. We currently have 146,216 sentences.

The languages **Bengali (*ben*)** with 1,200,255, **Tswana (*tsn*)** with 28,276, **Tsonga (*tso*)** with 10,571, **Venda (*ven*)** with 9,279, and **Xhosa (*xho*)** with 63,387 sentences can also be counted as successes as we got a variety of new domains (including **.com**) with the proportion of wikipedia being rather low (less than 20%).

Table 3 shows the top-5 domains of those five languages. `wikipedia.org` is highlighted in italics.

Lang.	Domain	Sentences	%
ben	http://www.jugantor.com/	126,841	10.7
	http://www.anandabazar.com/	110,892	9.4
	http://www.prothom-alo.com/	54,647	4.6
	http://bn.wikipedia.org/	34,698	2.9
	http://www.guruchandali.com/	30,308	2.6
tsn	http://www.mmegi.bw/	6,672	23.6
	http://www.dailynews.gov.bw/	4,856	17.2
	http://www.kutlwano.gov.bw/	4,528	16.0
	http://tn.wikipedia.org/	3,578	12.7
	http://www.info.gov.za/	1,063	3.8
tso	http://rivoni.org/	1943	18.4
	http://globalrecordings.net/	1,579	15.0
	http://oldgov.gcis.gov.za/	1,521	14.4
	http://www.info.gov.za/	1,135	10.8
	http://www.nthavela.co.za/	569	5.4
ven	http://www.gov.za/	1,250	13.6
	http://globalrecordings.net/	1,130	12.3
	http://www.saqa.org.za/	1,119	12.2
	http://info.hannasteffens.de/	780	8.5
	http://africanlanguages.com/	518	5.6
xho	http://www.wordpocket.com/	17,669	27.9
	https://builttobrag.com/	10,249	16.2
	http://nalibali.mobi/	5,266	8.3
	https://xh.wikipedia.org/	5,077	8.0
	http://wced.school.za/	3,385	5.3

Table 3: Top-5 domains of sentences with amount and proportion (percentage of the whole corpus).

5. Discussion

As we have shown, the CURL portal is an accessible, uncomplicated tool for the collection and creation of text corpora for under-resourced languages that is actively used and can already provide resources for a significant number of different languages. These high-quality text corpora can in turn be freely used for further research and practical applications. For example, they can function as baseline corpora for a large number of NLP tasks, can serve as a basis for statistical analysis, can help improve language detection tools, and so on. The collected URLs may also be of use as seeds for further crawling processes for the respective language. First tests also show that the collected text material is suitable for enriching existing datasets (Bosch et al., 2018).

A worthwhile future addition to the CURL portal would be the training of word or character embeddings from the collected data. Word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), or contextual string embeddings like the Flair embeddings (Akbik et al., 2018) allow text to be represented in a semantically relevant, machine-interpretable way that has proven to be a valuable tool for a multitude of downstream NLP tasks. Trained models of such embeddings are unfortunately usually only available for a small number of well-resourced languages, making their creation for other languages a worthwhile goal, as this could enable new avenues for further research in the respective communities.

A community-driven endeavor like this is obviously very much dependent on external participation. We therefore also hope to strengthen our visibility to reach even more language communities interested in the contribution to and collaboration with this growing resource collection.

Acknowledgements

This research was partially funded by the Development Bank of Saxony (SAB) under project numbers 100335729 and 100341518. It is being developed further in the German national research consortium (NFDI) Text+ in the lexical cluster "Non-Latin scripts and Under-resourced languages". Text+ is funded by the German National Research Foundation (DFG) under project number 460033370.

6. Bibliographical References

Agić, Ž., Johannsen, A., Plank, B., Martínez Alonso, H., Schluter, N., and Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1313–1316. European Language Resources Association (ELRA).

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bosch, S., Eckart, T., Klimek, B., Goldhahn, D., and Quasthoff, U. (2018). Preparation and usage of Xhosa lexicographical data for a multilingual, federated environment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Brown, R. D. (2013). Selecting and weighting n-grams to identify 1100 languages. In Ivan Habernal et al., editors, *Text, Speech, and Dialogue - 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings*, volume 8082 of *Lecture Notes in Computer Science*, pages 475–483. Springer.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765. European Language Resources Association (ELRA).

Goldhahn, D., Sumalvico, M., and Quasthoff, U. (2016). Corpus Collection for Under-Resourced Languages with More than One Million Speakers. *Proceedings of Collaboration and Computing for UnderResourced Languages: Towards an Alliance for Digital Language Diversity (CCURL)*, pages 67–73.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2017). A Portal for Corpus Collection for Under-Resourced Languages. *Workshop of the African Association for Lexicography (AFRILEX)*, pages 15–17.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–348.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013*, Scottsdale, Arizona, USA.

Mohr, G., Stack, M., Ranitovic, I., Avery, D., and Kimpton, M. (2004). An Introduction to Heritrix – An open source archival quality web crawler. In *In IWAW'04, 4th International Web Archiving Workshop*. Springer Press.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Roziewski, S. and Stokowiec, W. (2016). LanguageCrawl: A generic tool for building language models upon Common-Crawl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2789–2793. European Language Resources Association (ELRA).

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. *Cahiers du Cental*, 5:1.

A. Ethical Considerations and Broader Impact

A.1. Copyright and License

The CURL corpora are automatically collected from public sources contributed by third parties without considering the content of the contained text in detail. No responsibility is taken for the content of the data. In particular, the views and opinions expressed in specific parts of the data remain exclusively with the authors.
Only public web pages are being crawled, and page metadata that restricts crawling, e.g. `robots.txt`, is respected to not affect normal use of the websites and avoid downloading unauthorized content.
In the creation process of the text corpora, only unique sentences with their source reference are kept, but the original order in text is discarded, so no reconstruction of original documents is possible.
All corpora CURL and LCC provide for download are licensed under the Creative Commons license CC BY.[7]

A.2. Impact

We do not expect the CURL corpora to have a significantly higher negative impact in terms of publication of personal data, copyright infringement, or systematic bias, compared to other publicly available web corpora. The web pages were crawled with open source tools (e.g. Heritrix) that allow anyone to download the same texts with consumer-grade computers. The full texts might also be contained and accessible in larger web crawls and collections such as `archive.org` or `commoncrawl.org`. To alleviate the likelihood of copyright and license infringement, the largest related text segments published are sentences.
We seek to lower the entry threshold for prospective new researchers by offering CURL as a free service for anyone to contribute to, thereby creating and making available the cleaned text corpora for academic research. This will allow anyone to quickly start working with and researching statistical properties of the corpora and various natural language-related tasks without first having to acquire raw texts and then build pipelines to prepare the data. The corpora contain monolingual sentences, cleaned and enriched with co-occurrence relations and statistical information, but can be used for further post-processing, including manual cleaning and annotation for various downstream tasks.

B. Languages and TLDs

Table 4 presents the most frequent TLDs for all the sentence resources of the CURL portal combined. TLDs with less than 7 languages are cut off. In particular, the TLD **.com** has the highest number of sentence source domains across all 80 languages. The total number of sentences is about 35 million.

[7] https://wortschatz-leipzig.de/en/usage

C. Tools

C.1. Text cleaning

Text cleaning occurs multiple times in the process:

1. When extracting text from WARCs, filtering out text blocks shorter than a minimum line length (defaults to 20 characters) and documents with too few lines (defaults to 80),
2. when detecting the language of a document, to filter out documents where the majority language does not match the target language,
3. after sentence segmentation by using rule-based filters that employ empirical values for lengths, amounts of characters (e.g. punctuation), and regular expressions for certain undesirable patterns,
4. finally with language detection on the sentence level.

C.2. Sentence Segmentation

For sentence segmentation we consider the following context information:

- List of possible sentence boundaries (or combinations), e.g. punctuation characters,
- tokens that should not occur in front of a sentence boundary (i.e. a list of abbreviations),
- patterns that should not occur in front of a sentence boundary (like typical year dates "[1-2][0-9]{3}"),
- tokens that should not occur immediately after a sentence boundary (like typical names for months),
- patterns that should not occur immediately after a sentence boundary (like a lowercase word).

Each language contains custom abbreviation lists, some of which include custom sentence boundaries or other configurations. For languages without sentence boundaries like Thai, third-party state-of-the-art segmentation tools are used.
Parameters and lists for different languages were refined by language experts and user feedback over time.

C.3. Word Tokenization

Word tokenization for most languages employs a whitespace tokenizer extended with rules about punctuation characters, abbreviations, and multi-word unit lists. Non-whitespace segmented languages are tokenized using third-party language-specific software.

D. Data statement

We document the following dataset information about the CURL corpora collection according to the approach proposed by Bender and Friedman (2018).
Dataset Name: CURL – Crawling Under-Resourced Languages
Dataset License: Creative Commons license CC BY [7]
Link to Dataset: https://curl.wortschatz-leipzig.de/languages

D.1. Curation Rationale

The CURL corpora are a collection of monolingual web text corpora. Seed URLs are community-contributed, with subsequent automated web crawling, text cleaning and corpus creation. The corpora include tokenized sentences, words, word co-occurrences and statistics, sources, and the relations in-between.
The aim of the CURL project is the discovery of digital text resources for under-resourced languages with the help of native speakers. The computed text corpora are made publicly available to support academic research.

D.2. Language Variety

All CURL corpora are monolingual, identifiable by their ISO 639-3 code. The languages covered are listed by Goldhahn et al. (2016), with the current selection available on the CURL web page.[5]

D.3. Speaker Demographic

It was not possible to collect and analyze detailed information about the demographic characteristics of the authors of the collected sentences due to the variety of languages and amount of texts. For the language skills, we assume native speaker proficiency levels and a high number of different authors.

D.4. Annotator Demographic

N/A

D.5. Speech Situation

Contemporary written language on web pages. No detailed information about topics was collected.

D.6. Text Characteristics

Web texts, Wikipedia. From informal to formal. Varying text quality. No restrictions on topic.

D.7. Recording Quality

N/A

D.8. Other

Curators: Non-native speakers, ages between 25–65, female and male. Experts in computational linguistics, natural language processing and corpus creation, with extensive proficiency in German and English.
Contributors (of seed URLs and texts): Anonymous.

D.9. Provenance Appendix

N/A

TLD	Sentences	Domains	Languages
org	6,900,832	755	ace, ach, amh, anw, asm, aym, bam, ban, bcl, ben, bew, bjn, bod, bos, bug, cdo, ceb, che, chv, ckb, diq, ewe, ful, gan, glg, glk, grn, hat, hau, hye, ibo, ilo, jav, kab, kan, kas, kbd, kde, kea, khk, khm, kik, kin, kir, kng, knn, kon, ksw, kur, lao, lin, lug, mal, min, mkd, mkw, mlg, mos, msa, mya, mzn, ndo, nep, nso, nya, oci, ori, orm, pag, pam, pes, pnb, pnt, que, rom, run, sin, skr, sna, snd, snk, som, sot, sqi, sun, tat, tel, tgk, tgl, tha, tir, tsn, tso, tuk, tum, uig, uzb, ven, vls, war, wol, xho, ydd, yor, zha, zul
com	15,548,196	88,885	aar, ach, amh, bam, ban, bcl, bem, ben, bik, bos, ceb, ckb, diq, ewe, fon, glg, glk, gom, hat, hau, hil, hye, ibb, ibo, jav, kan, kck, kde, kea, khk, kin, kng, knn, kur, lao, lgg, lug, mal, mkd, mos, msa, mya, nep, ngl, nor, nso, nya, oci, orm, pag, pes, pnb, prs, run, seh, sin, skr, sna, snk, som, sot, sqi, suk, sun, swa, tam, tat, tel, tgk, tgl, tha, tsn, tso, tuk, ven, wol, xho, ydd, yor, zul
de	39,010	288	ban, bcl, bem, bos, ceb, ckb, diq, emk, fuc, glg, glk, hye, jav, kde, kea, khk, kin, knn, lgg, mos, msa, ngl, nya, oci, pes, seh, snk, som, suk, sus, tgl, tha, ven
net	681,689	50	ben, bod, bos, glg, glk, hye, kan, khk, kur, lao, lin, msa, ndo, nep, pes, run, sna, som, tel, tha, tsn, tso, ven, wol, xho, zul
pl	1,112	143	ach, ban, bcl, bos, ceb, diq, glg, hye, ibb, jav, kde, kea, kng, knn, lgg, msa, ngl, nya, oci, pes, seh, snk, som, tgl
cz	1,562	429	ban, bcl, bem, bos, diq, fuc, glk, hat, ibb, jav, kea, khk, kng, knn, ngl, nya, oci, pag, pam, pes, snk
se	2,496	61	ban, bem, bos, ckb, diq, glk, jav, kea, kng, knn, mos, nya, nyn, oci, pes, prs, snk, som, suk, tha
sk	1,013	175	ban, bcl, bem, bos, glg, ibb, jav, kde, kea, kng, knn, lgg, min, mos, ngl, nya, oci, pam, pes, seh
nl	1,595	54	bos, diq, glk, jav, kea, knn, mos, ngl, nya, oci, pes, snk, som, sus, tgk, tgl, tha, tiv
no	50,661	75	bos, ckb, diq, emk, glk, kea, knn, msa, nbl, nor, oci, pes, pnb, skr, sna, som, tgl, tir
ru	12,205	179	bak, bos, chv, ckb, glk, hye, jav, khk, kir, knn, nya, oci, pes, snk, tat, tem, tgk, tgl
ch	513	74	bos, ckb, diq, glk, jav, kde, kea, kin, kng, knn, msa, oci, pes, suk, sus, tgl, tha
in	23,732	1,208	ace, bem, bos, ceb, glk, jav, knn, lgg, mad, mal, msa, pes, pnb, skr, sun, tgl, tha
jp	313	35	bem, bos, hat, kde, kea, kng, knn, lgg, mos, msa, nya, pes, snk, suk, tgl, tha
cn	26,541	27	bem, bos, ceb, glk, hau, ibb, jav, khk, mos, msa, oci, pes, tgl, tha, uig
ir	2,081,118	5,120	bos, fuc, glk, hau, kir, knn, msa, pes, pnb, run, skr, tgk, tgl, tha, tuk
it	734	85	bcl, bem, bos, glg, hau, ibb, kea, knn, lgg, nya, oci, pes, seh, suk, tgl
dk	923	32	bos, ceb, ckb, diq, glk, ibb, jav, kea, mos, oci, pes, pnb, skr, som
edu	5,607	37	bcl, bik, glg, glk, hat, msa, pes, pnb, tgl, tha, wol, xho, ydd, zul
ca	693	30	bos, glk, hat, hau, hye, mos, msa, pes, pnb, snk, ssw, tgl, tha
co.za	58,163	1,255	bos, kng, nbl, nso, nya, pes, sot, ssw, tsn, tso, ven, xho, zul
fr	4,616	103	bos, glk, hat, hye, kng, knn, msa, oci, pes, run, snk, tha, zul
ro	853	395	ban, bcl, bos, diq, kea, knn, nya, oci, pag, pes, snk, suk, tgl
tr	467	22	bcl, ckb, emk, glk, jav, kde, knn, lgg, ngl, pes, snk, suk, tuk
tw	76,877	25	bos, hau, ibb, jav, kea, khk, kng, knn, mos, msa, nan, pes, tha
ee	38	19	bos, chv, glg, kir, lgg, mad, mos, ngl, nya, oci, snk, suk
za	36,572	260	knn, nbl, nso, nya, pes, sot, ssw, tsn, tso, ven, xho, zul
lt	107	22	ban, bos, glk, kde, kea, knn, lgg, mos, oci, pes, tgl
mobi	10,781	10	bos, hau, mos, msa, nso, snk, sun, tgl, ven, xho, zul
au	533	20	bos, glk, hau, kea, msa, oci, pes, run, som, tgl
be	317	19	bos, diq, glg, glk, knn, lin, oci, pes, tha, vls
fi	546	35	ban, bem, bos, kde, knn, ngl, oci, pes, snk, som
hu	131	43	bem, bos, hye, jav, kde, knn, ngl, nya, oci, pes
ac.za	72,867	86	nbl, nso, sot, ssw, tsn, tso, ven, xho, zul
eu	6,355	51	bos, glg, glk, hye, kea, oci, pam, pes, tgl
hr	36,978	2,984	bcl, bos, diq, emk, kde, knn, oci, tgl
my	189,083	729	ban, bcl, jav, min, msa, pes, sun, tgl
si	610	146	bcl, bos, knn, lgg, oci, pes, sus, tha
tk	982	36	bos, glk, kde, kea, mos, msa, nya, pes
ws	1,943	24	bcl, bos, glk, knn, msa, nya, pes, zul
ac.jp	210	7	bcl, glk, hau, khk, kng, nya, tgl
at	722	62	bos, ckb, glk, kea, msa, pes, tgl
cl	62	21	bos, glg, kea, mad, nya, oci, tgl
co.jp	426	14	bem, diq, khk, kng, pes, tgl, zul
co.uk	820,267	5	mkd, msa, nep, nya, pes, run, sqi
es	122,155	287	bos, glg, glk, kea, knn, oci, pes
gov	4,482	11	bcl, hat, kea, pes, som, tgk, tgl
gr	46	10	bos, knn, mos, oci, pag, pes, tha

Table 4: Most common TLDs with list of languages and their combined amount of sentences and domains.