Data in Brief 54 (2024) 110325

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset for Siswati: Parallel textual data for English and Siswati and monolingual textual data for Siswati

Tanja Gaustad*, Cindy A. McKellar, Martin J. Puttkammer

Centre for Text Technology, North-West University, South Africa

ARTICLE INFO

Article history:
Received 5 December 2023
Revised 11 March 2024
Accepted 11 March 2024
Available online 16 March 2024

Dataset link: Monolingual Siswati Corpus (Original data)
Dataset link: Bilingual English-Siswati Corpus (Original data)

Keywords:
Natural Language Processing
Human Language Technology
Machine translation
Language corpora
Under-resourced languages
South African languages

ABSTRACT

This data article presents a dataset for Siswati, a Bantu language of the Nguni group that is one of the eleven official South African languages and the official language of Eswatini (together with English). The dataset contains parallel textual data between English and Siswati as well as monolingual data for Siswati and was developed for use as training data for machine translation systems, specifically the Autshumato machine translation project. Both corpora can also be used for development and evaluation of Natural Language Processing (NLP) core technologies for Siswati. In addition, the data lends itself for corpus linguistic studies. The article describes how the data was collected, what type of texts it contains and what clean-up was done. It also provides an overview of the number of words contained in the datasets.

* Corresponding author.
E-mail address: Tanja.Gaustad@nwu.ac.za (T. Gaustad).

https://doi.org/10.1016/j.dib.2024.110325

Specifications Table

Subject	Computer Science
Specific subject area	Natural Language Processing, Human Language Technology, Linguistic corpus
Data format	.txt raw UTF-8 encoded monolingual Siswati text .txt aligned UTF-8 encoded Siswati translated text (parallel data) .txt aligned UTF-8 encoded English translated text (parallel data) .txt aligned UTF-8 encoded Siswati crawled text (parallel data) .txt aligned UTF-8 encoded English crawled text (parallel data) .txt ReadMe Monolingual data .txt ReadMe Bilingual data
Type of data	Parallel textual data for English and Siswati and monolingual textual data for Siswati.
Data collection	For the translated data, English texts were selected and sent out for human translation, followed by a stringent quality control. The crawled data was collected from South African governmental web domains, then converting relevant documents to plain text format and identifying parallel and monolingual data. The parallel data was further processed whereas data obtained by crawling that did not align at document-level was gathered for the monolingual Siswati corpus.
Data source location	Institution: Centre for Text Technology, North-West University City/Town/Region: Potchefstroom Country: South Africa
Data accessibility	Repository name: SADiLaR https://repo.sadilar.org/ Data identification number: NA Direct URL to data: Parallel data: https://hdl.handle.net/20.500.12185/560 Monolingual data: https://hdl.handle.net/20.500.12185/559

1. Value of the Data

- This dataset is useful for training Machine Translation systems for English-Siswati and/or Siswati-English.
- Researchers working on Human Language Technology (HLT) applications for Siswati will be able to use the dataset for research, development and evaluation purposes.
- Many different researchers can profit from this data: It is interesting for linguists to study the use of Siswati from a big corpus; for translation specialists it can provide insights into translations used for particular terms; machine learning engineers, computational linguists and other researchers working with linguistic data can use it as input to their tasks.

2. Data Description

The dataset described in this article contains approximately two million words of parallel text for English and Siswati (measured on the English portion) as well as an additional 1,5 million words of monolingual Siswati text. The data is split into five separate text files, all in UTF-8 encoding: One file for the monolingual data, two files for the translated data in English and Siswati respectively, and two files for the crawled English and Siswati data. In the files containing translated data and crawled data, each line in the Siswati corpus has an equivalent in the English corpus on the same line, making this a parallel data set. The format of each file is the same, namely one segment per line. For the purpose of this resource a segment is defined as any combination of one or more words, including complete and partial sentences as well as headings and numbered lists. Also, the data is tokenized, i.e. punctuation is separated from a word by a space. This format works well as input to a potential Machine Translation system. The dataset is distributed under the Creative Commons Attribution 4.0 International license.[1]

Table 1 gives an overview of the exact number of sentences and words contained in the five files.

[1] CC BY 4.0 - https://creativecommons.org/licenses/by/4.0/.

Table 1
Statistics on the English-Siswati data set.

	Segments	English words	Siswati words
Monolingual dataset	*138,651*		*1,536,356*
Translated dataset	53,382	1,000,531	705,060
Crawled dataset	61,457	1,001,762	718,354
Entire parallel dataset	*114,839*	*2,002,293*	*1,423,414*

3. Experimental Design, Materials and Methods

The parallel corpus described in this article contains data acquired in two ways:

- By human translations from English to Siswati; and
- By crawling various South African governmental web domains and extracting parallel data for English and Siswati.

The monolingual corpus contains Siswati data obtained during crawling that could not be aligned with an English counterpart at document-level.

For the translated data, suitable documents that only occurred in English were selected from the crawled data. Preference was given to pdf documents as the html pages often contained few usable sentences once the menus were removed. Sentences were ranked to remove very similar sentences and to ensure no duplication occurred. Also, all sentences with less than 80 % words recognized by the Microsoft Word compatible spelling checker for English were removed from the data to be translated. After this cleaning step, the English data was proofread to only include well-formed sentences and bundled into batches for translation. Each batch was sent out to human translators together with instructions to follow for the translations into Siswati. The brief specified for the translator to use the official orthography and diacritics, to be consistent in translating recurring terms, not to add or remove information when translating and not to change the meaning of the original sentence.

The finished translations went through a quality control script checking for spelling errors, completeness of the translations as well as consistency between source and target sentences. After a feedback loop with the original translator to fix any issues found during quality control, 5 % of the final translated data was checked by an independent translator for overall quality and accuracy of translation. In total, 10 translators worked on this project, all contracted through a translation agency run by a South African Translators Institute (SATI)[2] accredited senior translator.

To obtain the crawled data, various South African governmental web domains were crawled using HTTrack[3]. The subsequent data processing and clean-up process included the following steps:

- Conversion of pdf, doc and html documents to plain text format, encoded in UTF-8, using publicly available modules;
- Language classification to identify English and Siswati documents using a language identification tool for South African languages [1,2][4];
- Anonymisation of data to protect the sources if required[5];
- Alignment on document-level using document names, internal document structure as well as website structure;
- Alignment on sentence-level using the HunAlign algorithm [3];

[2] https://www.translators.org.za/.
[3] https://www.httrack.com/.
[4] The Language Identifier used is a slightly updated version of the one described in the references, but as it is not yet publicly released these is no newer publication to cite.
[5] For the Siswati data, no anonymisation was required.

- Removal of poorly aligned data;
- Sorting of data into monolingual and parallel corpora;
- Removal of text segments that do not include useful data (e.g. lines containing only telephone numbers or punctuation, as well as empty lines);
- Removal of text segments that include a large number of spelling errors or diacritic symbols that were not correctly converted;
- Language identification on sentence level and removal of non-English and/or non-Siswati text segments;
- Randomisation of text segments.

All the data obtained by crawling that could not be aligned on document-level was used for the monolingual corpus and put through the same clean-up process as the parallel data.

The final dataset contains approximately 115,000 parallel segments and 140,000 monolingual segments on various topics from the government domain, such as speeches, press releases, health information, etc.

Limitations

The dataset presented can be considered limited in size when compared to mainstream languages such as Spanish or English, but for a resource-scarce language it is considerable and sufficient to build e.g. a machine translation baseline system. The content being from government web domains only can be seen as another limiting factor, although there is still a variety of different types of texts present (speeches, informational pamphlets, educational sources, etc.). However, there is no creative writing content, such as novels, or informal writing for instance contained in the data.

Ethics Statement

The authors have read and understood the ethical guidelines and followed the ethical requirements for publication in Data in Brief. The data set described in this article did not involve human subjects, animal experiments or any data collected from social media platforms.

Data Availability

Monolingual Siswati Corpus (Original data) (SADiLaR).
Bilingual English-Siswati Corpus (Original data) (SADiLaR).

CRediT Author Statement

Tanja Gaustad: Supervision, Validation, Data curation, Writing – original draft, Writing – review & editing, Project administration; **Cindy A. McKellar:** Methodology, Data curation, Validation, Writing – review & editing; **Martin J. Puttkammer:** Conceptualization, Funding acquisition, Data curation, Writing – review & editing.

Acknowledgments

This research was made possible with the support from the South African Centre for Digital Language Resources (SADiLaR). SADiLaR is a research infrastructure established by the Department of Science and Innovation (DSI) of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Hocking, Language identification for South African languages, in: Proceedingsof the Annual Pattern Recognition Association of South Africa and Robotics andMechatronics International Conference (PRASA-RobMech), 2014, p. 307. http://www.prasa.org/proceedings/2014/PRASA2014.pdf.

[2] M.J. Puttkammer, R. Eiselen, J. Hocking, F. Koen, NLP web services for resource-scarce languages, in: Proceedings of the ACL 2018, System demonstrations, 2018, pp. 43–49, doi:10.18653/v1/P18-4008.

[3] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy, Parallel corpora for medium density languages, in: Proceedings of the RANLP 2005, 2005, pp. 590–596. http://mokk.bme.hu/en/resources/hunalign/.