Data Article

Linguistically annotated dataset for four official South African languages with a conjunctive orthography: IsiNdebele, isiXhosa, isiZulu, and Siswati

Tanja Gaustad*, Martin J. Puttkammer

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

A R T I C L E   I N F O

Article history:
Received 30 September 2021
Revised 15 February 2022
Accepted 21 February 2022
Available online 25 February 2022

Keywords:
Natural language processing
Human language technology
Linguistic annotation
Nguni languages

A B S T R A C T

This data article presents a linguistically annotated data set for four official South African languages with a conjunctive orthography, namely isiNdebele, isiXhosa, isiZulu and Siswati. The data set is parallel for all four languages and can be used for language-specific as well as cross-language development and evaluation of Natural Language Processing (NLP) core technologies. In addition, it can be used for corpus linguistic studies. The article describes how the data was collected, what type of texts it contains and it provides some details on the three different types of linguistic annotation added (morphology, part-of-speech and lemmas), including an example.

© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

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

https://doi.org/10.1016/j.dib.2022.107994
2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
Specifications Table

| Subject                           | Computer Science |
|-----------------------------------|------------------|
| Specific subject area             | Natural Language Processing, Human Language Technology, Linguistic corpus |
| Type of data                      | Textual data with 3 different types of linguistic annotations (linguistically annotated corpus) for 4 Nguni languages |
| How data were acquired            | The data is a cleaned-up subset of a collection of documents originating from the South African government domain websites (*.gov.za) that contains parallel data for isiZulu, isiXhosa, isiNdebele and Siswati. |
| Data format                       | Raw (ASCII) text, analysed (annotated), parallel |
| Parameters for data collection    | All data included in the dataset had to be present in all four languages. Also, data containing only contact information, repetition of whole sentences, lists and headings was removed. |
| Description of data collection    | The South African government domain websites (*.gov.za) were crawled. The retrieved documents in pdf and Microsoft Word format were converted to text format, language identification was performed and the files were aligned on paragraph-level across the relevant languages. After clean-up, the content was split into paragraphs, sentences and tokens before the linguistic annotations were added. |
| Data source location              | Institution: Centre for Text Technology, North-West University |
|                                   | City/Town/Region: Potchefstroom |
|                                   | Country: South Africa |
| Data accessibility                | Repository name: SADILaR |
|                                   | Data identification number: NA |
|                                   | Direct URL to data: https://repo.sadilar.org/handle/20.500.12185/546 |

Value of the Data

• This parallel data set can be used for language-specific and cross-language development and evaluation of Natural Language Processing (NLP) core technologies and applications for isiZulu, isiXhosa, isiNdebele and Siswati.

• Researchers working on Human Language Technology (HLT) applications for the four Nguni languages (isiZulu, isiXhosa, isiNdebele and Siswati) will be able to use the dataset for research, development and evaluation purposes.

• The morphologically annotated data allow for better investigations of morphological phenomena that will in turn lead to new knowledge and insights into the processes involved in word creation in these languages.

• Part-of-speech annotated data is a common linguistic investigation attribute that is widely used in syntactic analysis, as well as more complex syntactic investigations.

• Lemmatisation is viewed as an indispensable linguistic information source and can also be utilised in various text processing applications such as spelling checkers, information retrieval systems, text mining and machine translation systems.

• Also, researchers in the fields of Linguistics and Digital Humanities will benefit from this parallel dataset for comparative corpus studies of Nguni languages.

1. Data Description

The dataset contains eight separate text files in UTF8 encoding, a train and a test set for four Nguni languages, namely isiZulu, isiXhosa, isiNdebele and Siswati. The unannotated English data is also provided as a reference. The start of each original paragraph is marked by a line marker with a counter. In total, the dataset contains 1431 paragraphs for each of the languages, all split randomly into a 90% training and 10% testing part. Every token-annotation combination can be found on a new line. Tokens and all three annotation types, specifically morphology, lemma and part-of-speech (POS), are separated by a single tab character resulting in a four-column data representation (see the example in Table 1). The annotations are described in separate protocols.
Table 1
Example of annotated data in isiXhosa.

| Token         | Morphological analysis                                                                 | Lemma | POS  |
|---------------|----------------------------------------------------------------------------------------|-------|------|
| <LINE# 0002> | i[NPrePre9]-GEP[Abbr]                                                                  | GEP   | N09  |
| l-GEP         | i[NPrePre9]-na[CopPre]-ma[BPre6]-ncedo[NStem]                                         | ncedo | N09  |
| inamancedo    | a[RelConc6]-i[BPre5]-qela[NStem]                                                      | qela  | REL  |
| aliqela       | a[PossConc9]-u[NPrePre15]-ku[BPre15]-xhasa[VRoot]-a[VerbTerm]                        | xhasa | POSS06|
| okuXhasa      | u[NPrePre15]-ku[BPre15]-cwangcisa[VRoot]-a[VerbTerm]                                 | cwangcisa | V    |
| ukuCwangcisa  | a[RelConc6]-wa[PossConc6]-i[NPrePre15]-ku[BPre15]-ku[OC2ps]-nced[VRoot]-a[VerbTerm] | nceda | REL  |
| avokukunceda  | u[NPrePre15]-ku[BPre15]-qulunqa[VRoot]-a[VerbTerm]                                  | qulunqa | V    |
| ukuqulunqa    | i[NPrePre8]-zi[BPre8]-khokelo[NSlem]                                                | khokelo | N08  |
| isizikhokelo  | a[RelConc6]-u[NPrePre15]-ku[BPre15]-lawula[VRoot]-a[VerbTerm]                       | lawula | POSS08|
of the documents containing only contact information, repetition of whole sentences, lists and headings.

The text files consist of approximately 50,000 tokens total per language, parallel on paragraph-level between all four languages as well as English. English is used as the basis for the token counts during data selection due to the agglutinative nature of the four Nguni languages [1]. The final dataset contains 1431 paragraphs (for each of the languages) on different topics from the government domain, and includes speeches, press releases, health information and other information about government departments and services.

Each language corpus has then been annotated for three different types of linguistic information: morphology, part-of-speech (POS) and lemmas Table 1. contains an example for isiXhosa annotated data.

Before the annotation, we developed protocols and tag sets for all three levels of annotation which detail the procedure and standards used as well as the permissible tags. The tag sets and protocols are included in the data release. A total of 380 linguistically permissible morphology tags were defined, i.e. [AdjPref10] was used for the adjective prefix from class 10 or [NStem] to indicate a noun stem. These were combined during annotation to yield full morphological analyses of the tokens present in the data, e.g. izinto (‘things’) was analysed as i[NPprePre10]-zin[BPpre10]-to[NStem].

The POS tagset consists of 20 main word classes, e.g. V for verb or N for noun, with some classes including additional information on class numbers, e.g. N08 (noun class 8) or POSS10 (possessive class 10). A total of 107 unique POS tags was available to the annotators, most of which also occurred in the data (see Table 2).

For lemmatization, the aim was to identify the stem lemma for each token. In isiXhosa for example, the noun izinto will be analysed as the nominal stem to. This stem lemma in combination with the POS tag N10 (Noun class 10) encodes the essential syntactic information for the word izinto.

Table 2 below summarizes some statistics for the annotations contained in the data.

For each annotation task, the data was automatically pre-annotated and subsequently presented to linguistic experts for further annotation and correction in the CTeXT Lara 2 annotation tool.3 The data was first annotated for morphology. Given the complex nature of agglutinative conjunctively written languages [2], the morphological information allowed for easier and more accurate pre-annotation of POS and lemma.

Once all levels of annotation had been completed, rigorous quality control (QC) was carried out to ensure that there is a high agreement between the morphology, POS and lemmatization annotations on token-level. The aim was to reach a 99% agreement on token-morph-lemma-POS combination. This was verified by using a rule-based generator to extrapolate the POS and lemma from the verified morphological analysis. The generated POS and lemma were then compared to the manually verified annotations received from the linguistic experts. Any discrepancies between the generated and QC’ed data was sent back to the linguistic experts to identify the error(s) and make corrections. Cross-language comparisons were also done to ensure the protocols were consistently applied across the four languages.

The final dataset was split into 90% train and 10% test sets. These datasets were used during the development of lemmatisers, POS taggers and morphological analysers for these languages, as described in [3] as well as for the South African Conference for Artificial Intelligence Research (SACAIR) and the Digital Humanities Association of Southern Africa (DHASA) shared task, the 2021 Nguni languages POS tagging challenge.4

3 https://hdl.handle.net/20.500.12185/432.
4 https://2021.sacair.org.za/shared-task/ and https://dh2021.digitalhumanities.org.za/shared-task/.
Ethics Statement

NA.

Data Availability

Linguistically enriched corpora for conjunctively written South African languages (Original data).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

CRediT Author Statement

Tanja Gaustad: Data curation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing; Martin J. Puttkammer: Conceptualization, Funding acquisition, Methodology, 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).

References

[1] C.M. Doke, Bantu languages, inflexional with a tendency towards agglutination, Afr. Stud. 9 (1) (1950) 1–19, doi:10.1080/00020185008706801.
[2] F. Kambarami et al., Computational modeling of agglutinative languages: the challenge for southern bantu languages, arusha working papers in African linguistics, 3:1 (2021) 52–81. ISSN: 2576-8638. http://arushalinguistics.org/publications/Kambarami_et_al_AWPAL_2021.pdf.
[3] J.S. du Toit, M.J. Puttkammer, Developing core technologies for resource-scarce nguni languages, Information 12 (2021) 520, doi:10.3390/info12120520.