Journal of Open Humanities Data

Updated Morphologically Annotated Corpora for 9 South African Languages

DATA PAPER

TANJA GAUSTAD
CINDY A. MCKELLAR
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The dataset described in this article presents converted and updated corpora for nine of the twelve official South African languages. After a revision of the morphological annotation protocols, the existing National Centre for Human Language Technology (NCHLT) corpora (Eiselen & Puttkammer, 2014) have been converted to updated morphological tags and consequently checked by linguistic experts for correctness. The resulting corpora are uniformly linguistically annotated for morphology across all nine languages, amounting to approximately 70,000 tokens for the five disjunctively written languages and 45,000 tokens for the four conjunctively written languages. The corpora are primarily aimed at the development and evaluation of Natural Language Processing (NLP) core technologies. In addition, the data can be used for language-specific and cross-language comparative corpus linguistic studies as well as corpus-based investigations of morphological phenomena in the included languages.

CORRESPONDING AUTHOR:
Tanja Gaustad
Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa
tanja.gaustad@nwu.ac.za

KEYWORDS:
language corpora; linguistic annotation; South African languages; under-resourced languages; human language technology

TO CITE THIS ARTICLE:
Gaustad, T., & McKellar, C. A. (2024). Updated Morphologically Annotated Corpora for 9 South African Languages. *Journal of Open Humanities Data*, 10: 38, pp. 1–5. DOI: https://doi.org/10.5334/johd.211

(1) OVERVIEW

REPOSITORY LOCATION

SADiLaR repository: https://repo.sadilar.org/handle/20.500.12185/1

Each language has a separate handle:

- Morphologically annotated corpus for isiNdebele: https://hdl.handle.net/20.500.12185/680
- Morphologically annotated corpus for isiXhosa: https://hdl.handle.net/20.500.12185/679
- Morphologically annotated corpus for isiZulu: https://hdl.handle.net/20.500.12185/678
- Morphologically annotated corpus for Siswati: https://hdl.handle.net/20.500.12185/677
- Morphologically annotated corpus for Sesotho: https://hdl.handle.net/20.500.12185/676
- Morphologically annotated corpus for Sepedi: https://hdl.handle.net/20.500.12185/675
- Morphologically annotated corpus for Setswana: https://hdl.handle.net/20.500.12185/674
- Morphologically annotated corpus for Tshivenḓa: https://hdl.handle.net/20.500.12185/673
- Morphologically annotated corpus for Xitsonga: https://hdl.handle.net/20.500.12185/672

CONTEXT

The data presented in this article was produced as part of the South African Centre for Digital Language Resources (SADiLaR) II (extension) project: *Linguistic corpus enrichment for South African languages*. It contains converted and updated morphologically annotated corpora for nine of the twelve official South African languages: data for the four languages with a conjunctive orthography, i.e. isiNdebele, isiXhosa, isiZulu, and Siswati, as well as for the five disjunctively written languages, i.e. Sesotho sa Leboa/Sepedi, Sesotho, Setswana, Tshivenḓa, and Xitsonga.

The (still) widely used annotated National Centre for Human Language Technology (NCHLT) corpora (Eiselen & Puttkammer, 2014) formed the basis of the data.[1] To encourage and enable cross-linguistic studies as well as guarantee compatibility with more recently morphologically annotated data for the four conjunctively written Nguni languages presented in Gaustad & Puttkammer (2022), the existing morphological annotations have been converted to updated morphological tags after a thorough revision of the relevant protocols. The annotations have consequently been checked for correctness by linguistic experts.

The resulting corpora are uniformly linguistically annotated for morphology across all nine languages: approximately 70,000 tokens for the disjunctively written languages and 45,000 tokens for the conjunctively written languages (approximately 100,000 tokens for the conjunctive languages if combined with the previously published data in Gaustad and Puttkammer (2022)). See Table 1 for an overview of the exact counts.

LANGUAGE	NUMBER OF TOKENS
isiNdebele	42,335
isiXhosa	46,465
isiZulu	45,933
Siswati	43,568
Sesotho	73,727
Sesotho sa Leboa/Sepedi	73,031
Setswana	72,609
Tshivenḓa	66,487
Xitsonga	69,584

Table 1 Overview of total token counts for all nine languages included in the dataset.

1 These datasets are available for download at https://repo.sadilar.org/handle/20.500.12185/1 as "NCHLT <LANGUAGE> Annotated Text Corpora".

(2) METHOD

The dataset contains documents originally crawled from various South African web domains (mainly government sites, municipalities, and official publications) using HTTrack[2] and converted to plain text with publicly available modules. For this updated version, only changes to tokenisation as well as correction of spelling mistakes were allowed on the original text.

STEPS

Before re-annotating the data, the tag set needed to be updated. To make the tag set as comparable as possible between the nine languages in our project, the linguistic experts discussed using the same tags for equivalent linguistic phenomena, and agreed on the most uniform tag set possible. Given the linguistic differences, each language nevertheless has a separate annotation protocol detailing the permissible morphological tags as well as containing examples to guide the annotation process. All tags contain a main category such as "NPre" to denote a nominal prefix or "Fut" for a future tense morpheme. Some morphological tags (for nouns, adjectives, various concords, pronouns, etc.) also include class information, for example [NPre10], which substantially increases the number of tags. Table 2 gives an overview of the total number of main morphology tags as well as the total number of distinct tags including class information per language.

Table 2 An overview of unique main morphology tags and total number of distinct tags per language.

	NUMBER OF UNIQUE MAIN MORPHOLOGY TAGS (WITHOUT CLASS INFORMATION)	TOTAL NUMBER OF MORPHOLOGY TAGS
isiNdebele	71	401
isiXhosa	77	370
isiZulu	74	423
Siswati	69	378
Sesotho	74	292
Sesotho sa Leboa/Sepedi	65	319
Setswana	63	313
Tshivenḓa	64	439
Xitsonga	67	290

Based on the revised protocols, the morphological annotations in the data were subsequently updated following these steps:

1. Retrieve a complete list of the old unique morphological tags for all nine languages.
2. Write a mapping script from old tags to new tags based on the revised protocols and apply to the data. For some languages, class information was not included previously. In those cases, the missing class was indicated with "??" in the new tag (e.g. [AbsPron??]) to indicate that class information needed to be supplied.
3. Pre-check token-morphological analysis pairs and add all unambiguous analyses to a dedicated list per language. These tokens are marked as correct in the annotation software LARA II[3] (Puttkammer, 2014).
4. Linguistic experts check all annotations not marked as correct in LARA II. This is done in batches of 2,000 tokens (for conjunctively written languages) to 5,000 tokens (for disjunctively written languages).
5. After a corrected batch is received back, carry out quality control (QC) using a QC script followed by manual checking. Also, add all new unambiguous token-analysis pairs to the dedicated file so they will not need to be checked in the remaining batches of data.
6. Once all batches have been annotated, do a final round of QC.

The final data is given as one text (.txt) file per language, where each line consists of a token and the corresponding morphological analysis separated by a tab character. The data also includes line numbering to mark sentences. Table 3 shows an example of the annotated data for Xitsonga. Each morpheme is separated by "-" and followed by a morphology tag between

2 https://www.httrack.com/.

3 Available at https://hdl.handle.net/20.500.12185/432.

square brackets – for example, vu[NPre14]-tirheli[NRoot], indicating that there are two morphemes in "vutirheli", namely a nominal prefix of class 14 [NPre14] "vu" as well as a noun root [NRoot] "tirheli".

TOKEN	MORPHOLOGICAL ANALYSIS
<LINE 1>	
Xikongomelo	xi[NPre7]-kongomelo[NRoot]
xa	xa[PossConc7]
website	website[Foreign]
ya	ya[PossConc9]
Vutirheli	vu[NPre14]-tirheli[NRoot]
bya	bya[PossConc14]
Afrika	Afrika[ProperName]
Dzonga	ri[NPre5]-dzonga[NRoot]
,	,[Punc]

Table 3 A Xitsonga example of morphologically annotated data.

(3) DATASET DESCRIPTION

OBJECT NAME

Morphologically annotated corpus for isiNdebele

Morphologically annotated corpus for isiXhosa

Morphologically annotated corpus for isiZulu

Morphologically annotated corpus for Siswati

Morphologically annotated corpus for Sesotho

Morphologically annotated corpus for Sepedi

Morphologically annotated corpus for Setswana

Morphologically annotated corpus for Tshivenḓa

Morphologically annotated corpus for Xitsonga

FORMAT NAMES AND VERSIONS

UTF-8 encoded .txt files

CREATION DATES

2022-04-01 – 2023-08-31

DATASET CREATORS

Jaco du Toit (Independent) – Data curation

Sunny Gent (Centre for Text Technology (CTexT), North-West University) – Project and annotation management

Martin Puttkammer (Centre for Text Technology (CTexT), North-West University) – Funding acquisition

LANGUAGE

isiNdebele (NR), isiXhosa (XH), isiZulu (ZU), Siswati (SS), Sesotho sa Leboa/Sepedi (NSO), Sesotho (ST), Setswana (TN), Tshivenda (VE), Xitsonga (TS)

LICENSE

REPOSITORY NAME

SADiLaR Language Resource Repository

PUBLICATION DATE

2024-03-27

(4) REUSE POTENTIAL

The corpora are primarily aimed at the development and evaluation of Natural Language Processing (NLP) core technologies and applications for the represented languages. When building morphological analysers and/or segmenters, the described data can be used to train and test the tools. Morphological information can also be incorporated into spelling checkers to improve the word recognition rate without increasing the size of the lexicon. In addition, the data can form the basis for language-specific as well as cross-language corpus linguistic studies and investigations of morphological phenomena, potentially leading to new insights into word creation and usage, morphological productivity and more. As the corpora are UTF-8 encoded text files, they can be used with a number of generic analysis tools (e.g. R, Python, SPSS).

ACKNOWLEDGEMENTS

We are grateful to our linguistic experts for their diligent work on the annotations.

FUNDING INFORMATION

This research was made possible with 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).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Tanja Gaustad: Conceptualisation; Data curation; Project administration; Quality Control; Writing – original draft; Writing – review & editing.

Cindy A McKellar: Validation; Writing – review & editing.

AUTHOR AFFILIATIONS

Tanja Gaustad orcid.org/0000-0002-1455-1941
Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa
Cindy A. McKellar orcid.org/0000-0001-9916-6139
Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa

REFERENCES

Eiselen, R., & **Puttkammer, M. J.** (2014). Developing Text Resources for Ten South African Languages. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Gaustad, T., & **Puttkammer, M. J.** (2022, April). Linguistically annotated dataset for four official South African languages with a conjunctive orthography: isiNdebele, isiXhosa, isiZulu, and Siswati. *Data in Brief*, *41*. DOI: https://doi.org/10.1016/j.dib.2022.107994

Puttkammer, M. J. (2014). *Efficient development of human language technology resources for resource-scarce languages* (PhD dissertation, North-West University).

TO CITE THIS ARTICLE:
Gaustad, T., & McKellar, C. A. (2024). Updated Morphologically Annotated Corpora for 9 South African Languages. *Journal of Open Humanities Data*, 10: 38, pp. 1–5. DOI: https://doi.org/10.5334/johd.211

Submitted: 02 April 2024
Accepted: 10 May 2024
Published: 11 June 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.