Development of linguistically annotated parallel language resources for four South African languages

Tanja Gaustad
Martin J. Puttkammer
Centre for Text Technology (CTexT®)
North-West University, South Africa
Tanja.Gaustad@nwu.ac.za
Martin.Puttkammer@nwu.ac.za

Abstract

For this project, we collected and annotated data to develop language resources for the four official South African Nguni languages written with a conjunctive orthography. The data for these four languages is parallel to allow for comparative (computational) linguistic studies. The corpora have been annotated for three types of linguistic information (morphology, part-of-speech and lemma). The article focuses on the annotation procedure, design choices that were made along the way as well as the quality control steps used. Hopefully this description will give some guidance for similar projects on under-resourced languages in the future.

Keywords: Resource development, Nguni languages, linguistically annotated corpora, parallel corpora, annotation procedure

1 Introduction

The aim of the project described here was twofold: firstly to build annotated corpora for four South African languages and secondly to develop core technologies based on the annotated data, namely stand-alone morphological analysers, part-of-speech taggers and lemmatisers. We will only focus on the corpus development in this paper. For details on the development and evaluation of the core technologies, see (du Toit & Puttkammer 2021).

We will first briefly describe the linguistic background in section 2 followed by an overview of the necessary components for the successful development of the proposed resources (section 3). Each of these components will then be presented in more detail: the data (section 4), the linguistic annotation prerequisites (section 5) as well as the annotation process itself (section 6). Section 7 presents an example of the final data. Conclusions and future work are discussed in section 8.

2 Background

South Africa has eleven official languages comprising nine Bantu languages and two Germanic languages (English and Afrikaans). The South African Bantu languages can generally be categorised into three language family groups: Four conjunctively written Nguni languages (isiNdebele, isiXhosa, isiZulu, and Siswati); five disjunctively written languages including four Sotho languages (Sepedi, Sesotho, Setswana, and Tshivenḓa) and one Tswa-Ronga language (Xitsonga). At least nine of these eleven languages are considered resource-scarce.

Bantu languages have a few interesting linguistic characteristics that make it complex to deal with computationally: They are tone languages, have an elaborate system of noun classification (up to 21 classes), and the verbal morphology is complex and highly agglutinative. Especially the agglutinative nature of Bantu languages accounts for their complexity (Doke 1950): Words are formed by combining morphemes (distinct meaning bearing units), usually a root (for verbs) or a stem (for other word classes), with one or more affixes. These affixes are bound morphemes with a singular function within the word. Especially verbs are very productive through i.e. derivational morphology, resulting in a large vocabulary.

Another factor to note is the writing system used for the different Bantu languages. A distinction is made between linguistic words and orthographic words as these two entities do not always coincide. For *disjunctively* written Bantu languages, several orthographic words can correspond to one linguistic word (Louwrens & Poulos 2006), whereas

for *conjunctively* written Bantu languages (which we are working with here) generally one orthographic word corresponds to one or more linguistic words.

See the following example taken from (Prinsloo & de Schryver 2002) for an illustration of disjunctive (Sepedi) versus conjunctive writing (isiZulu):

Sepedi	*ke a mo rata*			
	ke	a	mo	rata
	I	[pres.]	him/her	love
	'I love him/her'			
isiZulu	*ngiyamthanda*			
	ngi-	-ya-	-m-	-thanda
	I	[pres.]	him/her	love
	'I love him/her'			

In this article, we will be discussing work done on the four conjunctively written Nguni languages, namely isiNdebele, isiXhosa, isiZulu and Siswati.

3 Components for building annotated language resources

Building linguistically annotated language resources requires a fair bit of preparation, especially if they should be usable for computational linguistic tasks or machine learning (Pustejovsky & Stubbs 2012). We identified the following components necessary to successfully compile linguistically annotated language resources:

- Data: without data no resources;
- Prerequisites for linguistic annotation: tag sets and protocols, linguistic experts, annotation tool(s);
- Annotation process: description of processing steps, incl. quality control.

Each component will now be described in more detail including the design choices we needed to make.

Table 1: Tokens per language

Language	Token count	No paragraph markers
isiNdebele	51,120	49,689
isiXhosa	50,166	48,735
isiZulu	50,528	49,097
Siswati	49,104	47,673
English	67,048	65,617

4 Data

The dataset used for this project has been put together using randomly selected documents from the South African government domain websites (*.gov.za) and includes text on different topics, such as speeches, press releases, health information as well as other information about government departments and services. The usual mode of operation in translation departments is to translate from an English source document to one of the other languages. We therefore collected documents and websites that were available in English as well as the four Nguni languages, resulting in a parallel dataset with English as the source language. The reason for using government material was the relatively easy availability of data in general and parallel data in particular. The parallel nature of the data allows for comparative (computational) linguistic studies of these four Nguni languages.

We aimed for about 50,000 tokens for each language. This choice was based on experience with previous projects on the development of computational linguistics tools. Especially for conjunctively written languages with a large vocabulary enough data to train and test such tools is essential. At the same time, the project also needed to adhere to time and budget constraints.

After the final selection and clean-up, the data was separated into sentences and tokenised. Each paragraph is kept as a unit to be shown as context during annotation. The final token counts can be found in Table 1, with and without the 1,431 paragraph markers. For more detail on the data set, see (Gaustad & Puttkammer 2021).

5 Prerequisites for linguistic annotation

Next to the collection of data, another goal of the project was to annotate each language corpus for three different types of linguistic information: morphology, part-of-speech (POS) and lemmas.

Data annotated for morphology allows researchers to investigate morphological phenomena in real language corpora. The study of word creation processes can lead to a better understanding of these processes or even to new linguistic insights.

Assigning categories to words according to their syntactic function in a sentence is called POS tagging. It is often used as a first step in syntactic analysis and can also be successfully leveraged for e.g. authorship analysis or writing style detection.

The third type of linguistic annotation, lemmatisation, is generally seen as an indispensable source of linguistic information for spelling checkers, dictionaries, information retrieval systems, etc.

5.1 Tag sets and protocols

Before starting the actual annotation of the data, we needed to develop protocols and tag sets for each type of annotation. The protocols explain how to annotate the data and what existing international standards apply (EAGLES 1996). The tag sets contain a list of permissible tags along with a description and examples for each tag. The current tag sets are refined versions of the ones developed for the National Centre for Human Language Technology (NCHLT) project described in (Eiselen & Puttkammer 2014).

For the morphological annotation layer the aim was to provide full morphological annotations labeling each morpheme. To achieve this, a total of 380 linguistically permissible morphology tags were defined, i.e. *[VRoot]* was used to indicate the verbal root or *[SC15]* for a subject concord of class 15. These were combined during annotation to yield full morphological analyses of the tokens present in the data. For example the isiXhosa word *izinto* ('things') was analysed as *i[NPrePre10]-zin[BPre10]-to[NStem]*.

The POS tag set consists of 20 main word classes for all four languages, e.g. *ADJ* for adjectives or *CONJ* for conjunctions. Some tags include additional information on class numbers, e.g. *N09* (noun class 9) or *POSS06* (possessive class 6) resulting in a total of 107 unique POS tags available during annotation.

For lemmatisation, the aim was to identify the stem lemma for each token (Prinsloo 2009). The noun *izinto* in isiXhosa will be annotated with the stem lemma *to*. This stem lemma in combination with the POS tag *N10* (noun class 10) encodes the essential syntactic information for the word *izinto*.

5.2 Linguistic experts

Once the tag sets and protocols had been established, our corpora needed to be marked up with the relevant linguistic information. We considered the following two possibilities on how to accomplish the annotation of the data: linguistic experts or students could be recruited and trained or crowdsourcing could be used (Zaidan 2012).

Especially for the morphological annotation, in-depth linguistic knowledge is needed to correctly annotate the languages covered in this project. For crowdsourcing, we did not expect the general public to have enough background knowledge nor to find enough people speaking the four Nguni languages. Using (linguistics) students would have been a possibility, but again the background knowledge was a reason for concern as well as the volume of data to annotate in the given time-frame. That was the main reason we decided to have linguistic experts perform the annotation of the corpora.

5.3 Tools

The linguistic experts worked in the Lara II annotation tool[1] for all three levels of linguistic annotation. Lara II, developed by the Centre for Text Technology (CTexT®), is domain-specific software for the annotation of tokens, lemmas, POS tags, and

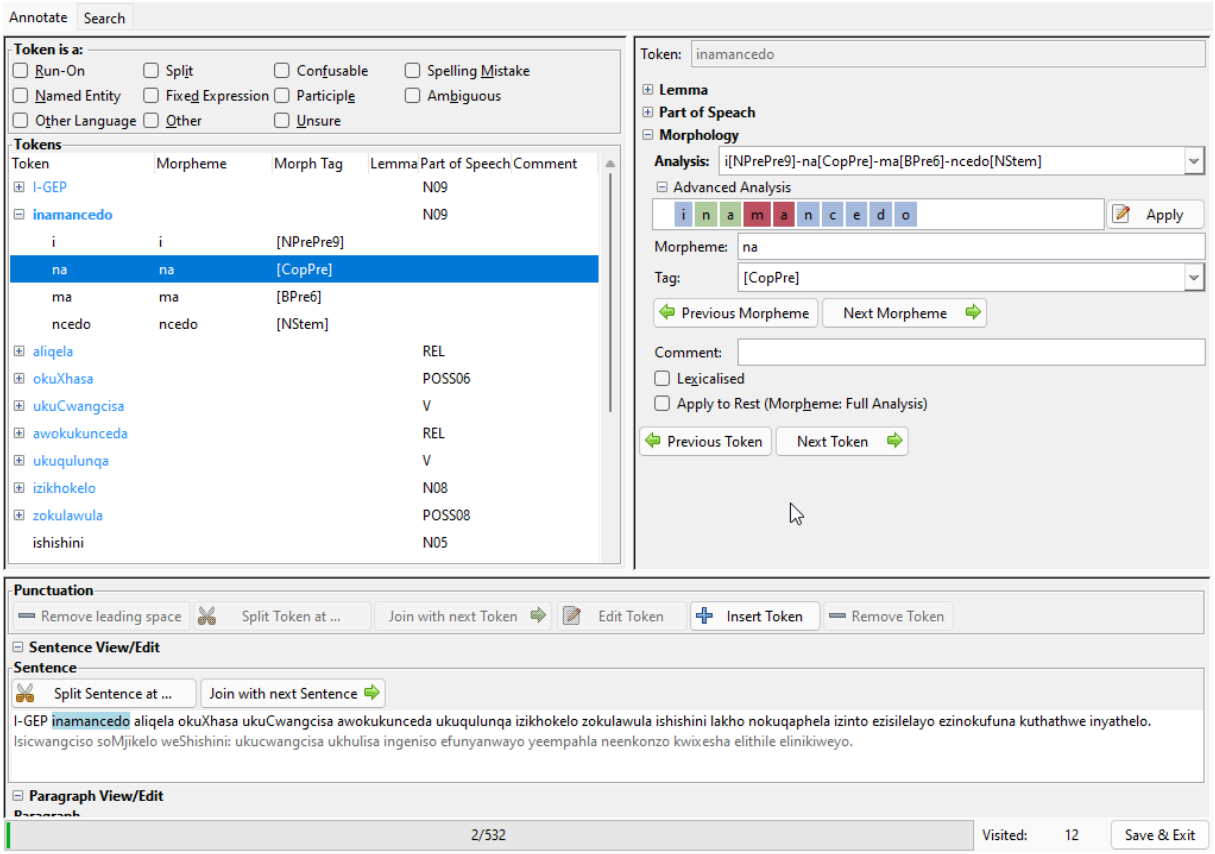

Figure 1: Annotation screenshot from Lara II

morphology. The aim of this tool is to enable users who have limited or basic computer skills to develop annotated, machine-readable corpora. The tool has shown to increase annotation accuracy while at the same time decreasing annotation time (Puttkammer 2014).

The interface contains the token to be annotated (highlighted), the context, an annotation field as well as a comments field. Lara II is easily adaptable to different annotation tasks via a configuration file. In our case this was ideal as each level of linguistic annotation had different requirements, from the allowed tags and the way in which the information is presented to the actions possible. See figure 1 for an example screenshot.

6 Annotation process

An essential part of every annotation project, especially when involving several layers of annotation that build on each other, is mapping out the process to follow. In our case, the following annotation steps were applied:

1. Morphology
   (a) Pre-annotate data for morphology
   (b) Linguistic experts correct pre-annotated morphology
   (c) Quality control for morphology
2. POS
   (a) Pre-annotate data for POS
   (b) Linguistic experts correct pre-annotated POS and if needed morphology
   (c) Quality control for POS

3. Lemmas
    (a) Pre-annotate data for lemmas
    (b) Linguistic experts correct pre-annotated lemmas
    (c) Quality control for lemmas
4. Quality control for all annotations

For both the morphological annotation and the POS annotation, the data was presented sequentially in Lara II. During POS annotation, the linguistic experts were also given the possibility to correct the morphological analysis if needed.

For the lemmatisation task, the linguistic experts were given an alphabetical list where each unique lemma-morphological analysis combination was presented separately together with one sample context paragraph.

6.1 Pre-annotation

As described in the annotation processing steps above, the data was automatically pre-annotated for each annotation task. Pre-annotation has been shown to speed up annotation (Lingren et al. 2014) as well as improve overall quality of the annotation (Rehbein et al. 2012). With morphologically complex languages such as the ones we are working with here, it is imperative to support the manual annotation process as much as possible (Puttkammer 2014).

Pre-annotation for morphology consisted in applying the NCHLT morphological decomposers (Eiselen & Puttkammer 2014) available for our four languages to generate all possible morphological analyses of the identified tokens. The linguistic experts could then choose the most accurate suggested analysis and make changes rather than annotate each token from scratch.

The statistics for the number of morphological analyses generated along with the average and maximum number of morphological analyses per token can be found in table 2. These numbers illustrate well the morphological complexity of the Bantu languages annotated in this project. Usually verbs will have the most generated possible analyses and closed class words like conjunctions will have only one analysis. Overall, the average number of morphological analyses is rather high.

For POS pre-annotation, a rule-based script produced a detailed POS tag per token using the morphological annotation as input. When possible errors in the morphological analysis were found, the token was either tagged as NERR (noun error), VERR (verb error), or ERR (general error). In these cases, the linguistic experts were asked to first correct the faulty morphological analysis and then add the correct POS tag.

Table 3 gives an overview of the generated POS error tags in our data. The statistics show that there were few morphological analyses that triggered error tags. This, however, does not imply that the rest of the generated POS tags were correct.

For lemmatisation, we also used a rule-based script taking the morphological annotation as input, in combination with lookup tables for closed class tokens such as conjunctions, to produce the most likely stem lemma per token. As noted above, for the annotation of lemmas only unique combinations of a lemma and a morphological analysis were presented to the linguistic experts in alphabetical format. Table 4 contains the number of lemmas that had to be checked per language. It also shows the reduction of tokens to annotate that was achieved by applying this strategy.

6.2 Quality control

Rigorous quality control (QC) has been carried out at various stages of the project. During the annotation of a type of linguistic information, QC was carried out to provide feedback to the linguistic experts, gather annotation as well as linguistic questions and resolve these issues in a structured, shared and documented way, for one language and also across languages. This process improved the quality of each type of linguistic annotation and made sure all linguistic experts applied the same rules and standards.

Table 2: Statistics on morphological pre-annotation

Language	Total morph. analyses generated	Average morph. analyses per token	Max. morph. analyses per token
isiNdebele	30,966	2.32	18
isiXhosa	22,298	1.61	30
isiZulu	37,811	2.93	24
Siswati	57,415	4.24	24

Table 3: Statistics on POS pre-annotation

Language	Total POS errors	NERR	VERR	ERR
isiNdebele	80	0	26	54
isiXhosa	317	22	34	261
isiZulu	66	0	27	39
Siswati	193	0	46	147

Table 4: Statistics on lemma pre-annotation

Language	Lemmas to annotate	Reduction
isiNdebele	17,318	65.1%
isiXhosa	17,094	64.9%
isiZulu	18,013	63.3%
Siswati	19,057	60.0%

After every type of annotation was finished, we checked the adherence to the protocols as well as differing annotations for the same/similar tokens. As explained in the overview of section 6, each annotation feeds into the next. To make sure as few mistakes as possible were used as input for the pre-annotation of the next step, QC after completion of each annotation level was crucial.

Once annotation on the data was finished for all linguistic levels, QC was done to ensure that there is a 99% agreement between the morphology, POS and lemma annotations. To be able to quantify the results, a rule-based generator extracted the POS and lemma automatically from the morphological analysis. This generated POS and lemma were then compared to the annotations in our manually verified data. All differences between the two were checked and corrected by the linguistic experts. This process was repeated until the evaluation criteria were met.

Unfortunately, we could not use inter-annotator agreement to measure the accuracy of the annotations (Artstein & Poesio 2008) because only one linguistic expert for each language included in this project was available. We did, however, apply the above described three way comparison in an effort to produce the best quality data with the given human resources.

7 Final data set

After finalizing the annotation and QC steps, the new resources for four Nguni languages were completed. The data can be accessed via the South African Centre for Digital Language Resources (SADiLaR) repository[2] and is distributed under the Creative Commons Attribution 4.0 International licence[3].

Table 5 contains an example of the final data for isiXhosa. The data is in a four-column text format with each column corresponding to a certain type of information: token, morphological analysis, lemma or POS. Each line contains a token-annotation combination. Line markers with a counter show the start of each original paragraph and can be used to align the content of the files.

Table 5: Example of final annotated resource for isiXhosa

Token	Morphological analysis	Lemma	POS
<LINE# 0002>			
I-GEP	i[NPrePre9]-GEP[Abbr]	GEP	No9
inamancedo	i[NPrePre9]-na[CopPre]-ma[BPre6]-ncedo[NStem]	ncedo	No9
aliqela	a[RelConc6]-li[BPre5]-qela[NStem]	qela	REL
okuXhasa	a[PossConc6]-u[NPrePre15]-ku[BPre15]-xhas[VRoot]-a[VerbTerm]	xhasa	POSSo6
ukuCwangcisa	u[NPrePre15]-ku[BPre15]-cwangcis[VRoot]-a[VerbTerm]	cwangcisa	V
awokukunceda	a[RelConc6]-wa[PossConc6]-u[NPrePre15]-ku[BPre15]-ku[OC2ps]-nced[VRoot]-a[VerbTerm]	nceda	REL
ukuqulunqa	u[NPrePre15]-ku[BPre15]-qulunq[VRoot]-a[VerbTerm]	qulunqa	V
izikhokelo	i[NPrePre8]-zi[BPre8]-khokelo[NStem]	khokelo	No8
zokulawula	za[PossConc8]-u[NPrePre15]-ku[BPre15]-lawul[VRoot]-a[VerbTerm]	lawula	POSSo8
ishishini	i[NPrePre5]-(li)[BPre5]-shishini[NStem]	shishini	No5
lakho	la[PossConc5]-kho[PossPron]	kho	POSSo5
nokuqaphela	na[AdvPre]-u[NPrePre15]-ku[BPre15]-qaphel[VRoot]-a[VerbTerm]	qaphela	ADV
izinto	i[NPrePre10]-zin[BPre10]-to[NStem]	to	N10
ezisilelayo	ezi[RelConc10]-silel[VRoot]-a[VerbTerm]-yo[RelSuf]	silela	REL
ezinokufuna	ezi[RelConc10]-na[CopPre]-u[NPrePre15]-ku[BPre15]-fun[VRoot]-a[VerbTerm]	funa	REL
kuthathwe	ku[SC15]-thath[VRoot]-w[PassExt]-e[VerbTerm]	thatha	V
inyathelo	i[NPrePre5]-(li)[BPre5]-nyathelo[NStem]	nyathelo	No5
.	.[Punc]	.	PUNC

8 Conclusions and future work

In the future, it would be interesting to collect data from more diverse sources and not just government text. Linguistically, a wider spread of genres will represent more types of real language use. Also, the core technologies developed on the basis of this data would be more generic.

We also learned that it is important to have quick feedback loops between the corrections done on an annotation layer and the QC carried out on the data. This helps to reach a good quality-level early in the annotation phase and minimizes the need for re-annotation and/or further correction of already annotated data.

One thing to point out is that it is very hard to find linguistic experts for South African languages who are available and qualified to do annotations for a project like ours. In the future, we might need to train new linguistic experts which will definitely influence the time-line as well as overall budgets of annotation projects. A conclusion to draw from this is that South African Bantu languages are not only under-resourced with regards to data, but that the development of human capital is also an important aspect of the development of resources for these languages.

Hopefully these parallel linguistically annotated corpora will prove interesting for researchers from different backgrounds and will help to gain more insight into the workings of these four Nguni languages, be it morphological processes, lemmatisation questions or syntactic structures.

Notes

[1] `https://repo.sadilar.org/handle/20.500.12185/432`

[2] `https://repo.sadilar.org/handle/20.500.12185/546`

[3] http://creativecommons.org/licenses/by/4.0/

Acknowledgements

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

Artstein, R. & Poesio, M. (2008), 'Survey article: Inter-coder agreement for computational linguistics', *Computational Linguistics* **34**(4), 555–596.

Doke, C. (1950), 'Bantu languages, inflexional with a tendency towards agglutination', *African Studies* **9**(1), 1–19.

du Toit, J. S. & Puttkammer, M. J. (2021), 'Developing core technologies for resource-scarce Nguni languages', *Manuscript under review for publication* .

EAGLES (1996), Expert advisory group on languages engineering standards: recommendations for the morphosyntactic annotation of corpora, Technical report, EAGLES, Document EAG-TCWG-MAC/R.

Eiselen, R. & Puttkammer, M. J. (2014), Developing text resources for ten South African languages, *in* 'Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)'.

Gaustad, T. & Puttkammer, M. J. (2021), 'Linguistically annotated dataset for four official South African languages with a conjunctive orthography: isiNdebele, isiXhosa, isiZulu, and Siswati', *Manuscript under review for publication* .

Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q. & Solti, I. (2014), 'Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements', *Journal of the American Medical Informatics Association* **21:3**, 406–413.

Louwrens, L. J. & Poulos, G. (2006), 'The status of the word in selected conventional writing systems - the case of disjunctive writing', *Southern African Linguistics and Applied Language Studies* **24**(3), 389–401.

Prinsloo, D. (2009), 'Current lexicography practice in Bantu with specific reference to the Oxford Northern Sotho school dictionary', *International Journal of Lexicography* **22**(2), 151–178.

Prinsloo, D. J. & de Schryver, G.-M. (2002), 'Towards an 11x11 array for the degree of conjunctivism / disjunctivism of the South African languages', *Nordic Journal of African Studies* **11**(2), 249–265.

Pustejovsky, J. & Stubbs, A. (2012), *Natural Language Annotation for Machine Learning*, O'Reilly Media, Inc.

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

Rehbein, I., Ruppenhofer, J. & Sporleder, C. (2012), 'Is it worth the effort? assessing the benefits of partial automatic pre-labeling for frame-semantic annotation', *Language Resources and Evaluation* **46:1**, 1–23.

Zaidan, O. F. (2012), Crowdsourcing annotation for Machine Learning in Natural Language Processing tasks, PhD thesis, Johns Hopkins University.