Proceedings of the 2nd workshop on Resources for African Indigenous Language (RAIL) at DHASA 2021

Virtual Conference

29th of November - 3rd of December 2021

Workshop Day

29th of November 2021

With special thanks to our sponsors

Peta sponsor

NATIONAL INSTITUTE
FOR THE HUMANITIES
AND SOCIAL SCIENCES

Tera sponsor and best paper award sponsor

GALE
A Cengage Company

Videolectures.net and audio-visual sponsor

SADiLaR
South African Centre for Digital Language Resources

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*.

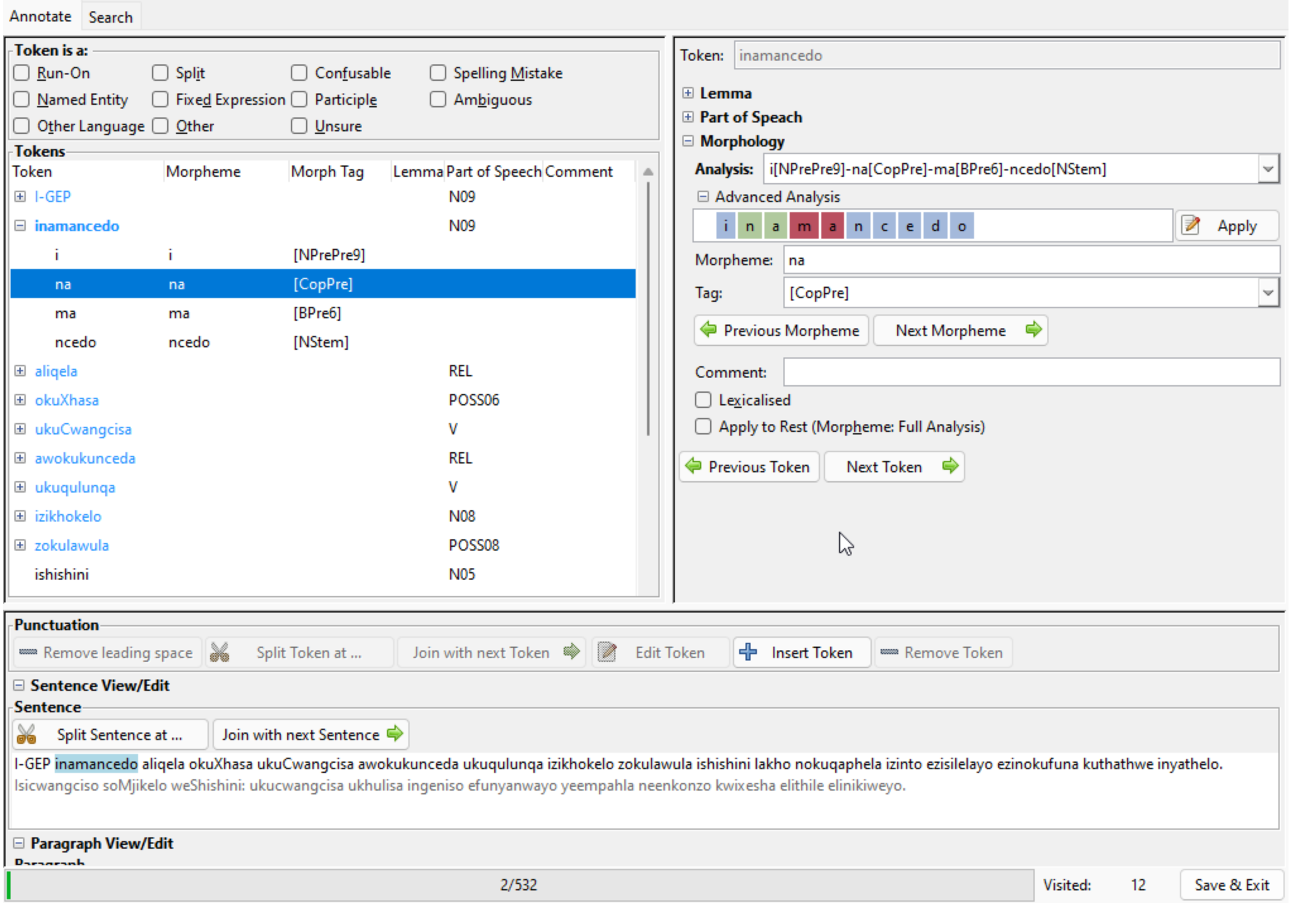
5.2 Linguistic experts

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



Token	Morpheme	Morph Tag	Lemma	Part of Speech	Comment
⊞ I-GEP				N09	
⊟ inamancedo				N09	
i	i	[NPrePre9]			
na	na	[CopPre]			
ma	ma	[BPre6]			
ncedo	ncedo	[NStem]			
⊞ aliqela				REL	
⊞ okuXhasa				POSS06	
⊞ ukuCwangcisa				V	
⊞ awokukunceda				REL	
⊞ ukuqulunqa				V	
⊞ izikhokelo				N08	
⊞ zokulawula				POSS08	
ishishini				N05	

Token: inamancedo

Analysis: i[NPrePre9]-na[CopPre]-ma[BPre6]-ncedo[NStem]

i n a m a n c e d o

Morpheme: na

Tag: [CopPre]

Comment:

I-GEP inamancedo aliqela okuXhasa ukuCwangcisa awokukunceda ukuqulunqa izikhokelo zokulawula ishishini lakho nokuqaphela izinto ezisilelayo ezinokufuna kuthathwe inyathelo.

Isicwangciso soMjikelo weShishini: ukucwangcisa ukhulisa ingeniso efunyanwayo yeempahla neenkonzo kwixesha elithile elinikiweyo.

2/532 Visited: 12

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	N09
inamancedo	i[NPrePre9]-na[CopPre]-ma[BPre6]-ncedo[NStem]	ncedo	N09
aliqela	a[RelConc6]-li[BPre5]-qela[NStem]	qela	REL
okuXhasa	a[PossConc6]-u[NPrePre15]-ku[BPre15]-xhas[VRoot]-a[VerbTerm]	xhasa	POSS06
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	N08
zokulawula	za[PossConc8]-u[NPrePre15]-ku[BPre15]-lawul[VRoot]-a[VerbTerm]	lawula	POSS08
ishishini	i[NPrePre5]-(li)[BPre5]-shishini[NStem]	shishini	N05
lakho	la[PossConc5]-kho[PossPron]	kho	POSS05
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	N05
.	.[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.

Digitising Afrikaans: Establishing a protocol for digitalizing historical sources for Early Afrikaans (1675-1925) as a possible template for indigenous South African languages

Wierenga, Roné
Virtual Institute for Afrikaans, North-West University
Carstens, Wannie
North-West University

Abstract

Afrikaans is one of South Africa's 11 official languages. With the continued rise of English as language of prestige in education, the economy and public sphere a movement towards a monolingual South Africa seems to be underway. This movement results in a loss of interest in the development and use of South Africa's indigenous languages. Another contributing factor is the unavailability or inaccessibility of resources in indigenous South African languages.

This paper reports on the protocol and process of establishing online resources to encourage research in and on Afrikaans.[i] The digitalisation process encompasses establishing a) digital bibliographies for Afrikaans literature and linguistics, where researchers can find links to online resources and pdf documents; b) an online archive where endangered (and often otherwise unobtainable) resources are stored in digital format; c) an online library where digitalized machine readable texts are stored; and d) an online historical corpus for Early Afrikaans (1675-1925). These resources are all angled towards linguistic research, specifically corpus linguistic research. VivA also offers a number of online resources to make research in Afrikaans more accessible – these include a general and school grammar for Afrikaans, a corpus portal for corpus linguists, and a speech atlas to create awareness of linguistic variety in Afrikaans.

The digitisation project is largely based on the Digital Library for Dutch Literature (Dutch: *Digitale Bibliotheek voor de Nederlandse Letteren*) (DBNL)[ii] and is still ongoing with various phases being undertaken simultaneously. *It is our goal to present this project as a template to demonstrate how other indigenous South African languages can go about establishing digital resources and to encourage synergy amongst linguists and linguistic institutions.*

Keywords: digitisation process, digital archive, digital bibliography, historical corpus, Early Afrikaans 1675-1925

1 Background

This paper reports on a project that first began to take shape in 1992. The notion of establishing an online database, similar to the predecessor of the current *Digitale Bibliotheek voor de Nederlandse Letteren* (DBNL), for Afrikaans was first verbalised in 1992. This idea stemmed from an interest amongst Dutch linguists to compare Afrikaans linguistic phenomena with those identified in Dutch. This form of comparative research was hindered by the lack of Afrikaans resources available to international researchers. No centralised digital platform existed to assist researchers in finding the appropriate resources.

Prof Wannie Carstens of the North-West University was tasked with digitising Afrikaans. In its infancy the process was primarily one of cataloguing all available linguistic publications on Afrikaans linguistics. In 2004 this catalogue was published on the North-West University's website as a digital bibliography for Afrikaans, the *Digitale bibliografie vir Afrikaans* (DBAT).[iii]

Initially DBAT was only a bibliography reference source to inform researchers of the available publications and where they are housed. It has since evolved into a digital repository where researchers can gain access to digital copies of language resources (an overview of DBAT is provided in section 2 of this paper). By 2015 a digital bibliography for Afrikaans literature, DBAL, was also established as a mirror image of its linguistic counterpart.

In 2018 the project began to take on another form. VivA undertook to become the home of and driving force behind the digitisation of Afrikaans, launching two legs of the project, namely a digital archive/library for Afrikaans and a historical corpus for Early Afrikaans (1675-1925). By 2019 a workshop was hosted and key institutions like VivA, SADiLAR, CTexT, the North-West University, University of the Free State and Nelson Mandela University came together to contemplate digitisation and the future of linguistics in South Africa.

Since 2019 immense progress has been made towards establishing a centralized database where researchers can obtain digital resources in and on Afrikaans. Other institutions, like the *Afrikaanse Taalraad*, Afrikaans language council (ATR) and private persons have also shown support for the project and been involved in various aspects of it.

2 The current state of affairs

The project has five branches that each function both as a whole and together as a unit to achieve the same goal, namely establishing a digital library for Afrikaans. The current state of affairs for each of the five parts is outlined in this section.

2.1 DBAT

The digital bibliography for Afrikaans linguistics, DBAT, [iv] currently consists of 17 829 publications about Afrikaans linguistics. These publications are, for the most part, academic in nature and include journal articles, books, and chapters within books, but blog posts, online articles and newspaper or magazine articles are also included when their topics relate to Afrikaans linguistics.

Around 60% of the resources listed on DBAT contain either a pdf copy of the publication or a hyperlink to a digital copy housed by another institution. The DBAT database is updated weekly, meaning that the newest publications are readily available. The database also has a Google Form that can be filled out by users to let us know of publications that has not been included. We also urge users to check their own publications on the database in order to insure that their entire publication history is available on DBAT. DBAT is currently working with a number of institutions to digitise the publications that do not have digital copies readily available and no longer have copyright restrictions.

2.2 DBAL

The digital bibliography of Afrikaans literature, DBAL,[v] similar to its linguistic counterpart, DBAT, is updated weekly. DBAL currently contains 17 000 publications of which 30% is available in digital format.

Unlike DBAT, DBAL is not yet current and various older publications are still being added to the database. DBAL is therefore in one of its final developmental stages. This is mostly because DBAL was established 11 years after DBAT, and because Afrikaans has vastly more literary resources than linguistic resources. However, once DBAL is current the focus will naturally shift towards the digitisation of the resources.

2.3 Historical Archive

The archive for historical resources is still in its infancy. The development of the archive began in 2018 and is still, partially, in its cataloguing phase. The first round of cataloguing was done alongside project leaders of the envisaged *Database Geschiedenis Nederlandse Taalkunde (*DAGENTA)[vi] Dutch digitalisation project. This is primarily because the project is based on the Dutch model and, due to Afrikaans and Modern Dutch's shared heritage, many Early Afrikaans texts were already known to the DAGENTA team who were interested in digitising older Dutch grammars. The idea was to join forces and share resources in order to digitise both Early Afrikaans and 17th century Dutch simultaneously, but unfortunately this goal has not yet been achieved.

The second round of cataloguing was started in late 2019 in South Africa. This round aimed to catalogue historical texts in and on Afrikaans that are a) no longer in use at South African universities, b) brittle or valuable due to their age, and c) in danger of being lost or damaged.

This process was unfortunately halted by the pandemic and travel restrictions, however a line of communication was established with the relevant universities and in many cases catalogues were provided by the universities which allowed us to pinpoint the text that are of interest to the project. In many cases private persons or institutions contacted us and made historical texts available to us for digitisation. For this reason a small collection of texts where digitised during 2020 and 2021.

The aim of this leg of the project is to preserve historical texts in a digital format and to make them available for research even if they have not yet been digitalized for corpus linguistic purposes.

2.4 Historical corpus of Early Afrikaans

Early Afrikaans 1675-1925 was a language form that originated from 17th century Dutch but was adapted for use by the Cape Malayan people, various European immigrants and strongly influenced by the Khoi-San community. The result is that this language variety differs from the Dutch used in Europe and shows notable signs of language influence.

Establishing a historical corpus of Early Afrikaans[vii] will allow researchers to explore the developmental phases of Afrikaans, but also to gain a greater understanding of language influence during this phase of South Africa's history. Kirsten (2016) outlines the development of a historical corpus for Afrikaans and list a variety of language changes that would likely not have been observed without a historical corpus. These changes include: a) a decrease in the use of passive constructions; b) the replacement of the preterit forms "had" and "wis" with "het gehad" and "het geweet"; c) the grammaticalisation of "gaan" to a future reference marker; and d) the replacement of the Dutch genitives with the Afrikaans variants "se" and "van".

Many of the texts to be included in the corpus has been catalogued, and the next phase would be to begin the digitisation process. The digitisation process is slower for the development of the corpus than for any other legs of the project. This is primarily because the text that are digitised are valuable, delicate or already partially damaged and therefore require greater care.

The quality of the digital copies also needs to be higher for these text in order for the digitalisation process to be effective and efficient. Low quality digital copies cannot always be digitalized (that is to say made machine readable), and would therefore need to be transcribed manually. Transcriptions can take a long time to complete, they also require a person to devote their time to the transcription, which can be expensive. Both manual transcriptions and automatic digitalisations need to be moderated in order to ensure that no errors occur.

A protocol for the digitisation of historical sources has been established and will be put to the test in 2022. This protocol outlines how to:

- establish networks with libraries and other institutions that house historical resources;
- establish synergy with other digitisation projects (like the Tracing History Trust corpus, [viii] Nuuseum project and DAGENTA project);
- develop an inventory that reflects the quality and scope of the project;
- establish a time line for the project;
- perform the digitisation process in a way that ensure the highest possible quality scans without damaging fragile texts; and
- implement quality control.

2.5 A book on historical sources in Afrikaans

In 1991 Prof Edith Raidt published a book about the origins of Afrikaans, *Afrikaans en sy Europese verlede* ("Afrikaans and its European history"). This publication outlines the landscape in which Afrikaans originated as well as the early development of the language before its standardization in 1925.

In her publication Prof Raidt provides a chapter wherein she names vital sources about the origins of Afrikaans. The idea to write a book that delves deeper into the value of these sources and the insight they give about the origins of Afrikaans. This book will complement the archive for historical Afrikaans resources and the Historical corpus of Early Afrikaans and act as guide for linguists and historians who use these digital resources. The layout and content of the book has already been established and the writing process will commence in 2022.

3 Afrikaans digitisation as a template

Due to commonalities between Afrikaans and Dutch, Dutch language technologies are often easy to adapt for Afrikaans and the needs of the Afrikaans linguistic communities. This does not, however, mean that the adaptation process does not pose any challenges. Often, as is the case in this project, only the basic outline for the DBNL could be used for Afrikaans. The institutional networks, database software, housing and creation, as well as the cataloguing, digitisation and digitalisation processes all needed to be established for Afrikaans. As is often the case, many challenges were only uncovered once a process was already underway and needed to be resolved as they arose.

This is the main benefit of this project for other indigenous languages. The Afrikaans version of the project has already navigated many pitfalls and is now capable of assisting other languages to navigate this terrain.

This project was undertaken with the goal of a centralized platform for digital Afrikaans resources in mind – eventually a comprehensive digital library for Afrikaans – but in a multilingual society and world, a centralized digital library for Afrikaans is of little value if it is not alongside digital libraries for all of South Africa's indigenous languages. Synergy is the key factor needed to undertake the digitisation of an entire language's resources successfully. Synergy on an institutional level, but also on a personal level. In the case of Afrikaans, it is a single person who became the driving force behind the project, motivating and enlisting others, who ultimately become the support structure and current home of the project.

A lot of footwork has been done in order to establish a digital resources for Afrikaans, and this footwork can ultimately benefit a similar project for another South African language.

Acknowledgements

A word of gratitude to everyone who has been involved with, or supported this project in a personal or institutional capacity. Your input,

whether man hours, insight, or willingness to provide access is indispensable.

References

Breed, C.A., Carstens, W.A.M. & Olivier, J. 2016. Die DBAT: 'n Onbekende digitale taalkundemuseum. *Tydskrif vir Geesteswetenskappe,* 56(2-1): 392-409.

Kirsten, J. 2016. Grammstikale verandering in Afrikaans van 1911-2010. Vanderbijlpark: North-West University.

Liebenberg, H. 2018. Die Wes-Kaapse Argie fen die begin van Afrikaans. *Tydskrif vir Geesteswetenskappe*, 58(2):204-236.

Raidt, E. 1991. Afrikaans en sy Europese verlede.

[i] This is an ongoing project.

[ii] This link can be used to access DBNL: https://www.dbnl.org/

[iii] For a detailed overview of the development and functionalities of DBAT refer to Breed *et al.,* (2016).

[iv] This link can be used to access DBAT: https://collections.nwu.ac.za/dbtw-wpd/textbases/bibliografie-afrikaans/dbat.html

[v] This link can be used to access DBAL: https://collections.nwu.ac.za/dbtw-wpd/textbases/bibliografie-afrikaans/dbal.html

[vi] This link can be used to access the DAGENTA repository: https://cls.ru.nl/dagenta/

[vii] Although the historical corpus for Early Afrikaans (1675-1925) has not yet been completed, examples of digital corpora for Afrikaans can be found on VivA's corpus portal: https://viva-afrikaans.org/

[viii] For a detailed overview of the Tracing History Trust corpus and the digitisation projects of the Western Cape Archives refer to Liebenberg (2018). The Tracing History Trust's digitisation products can be accessed at: http://www.tracinghistorytrust.co.za/products.htm

A novel method for redefining language ecology and endangerment in Nigeria – towards a geospatial solution

Udoh, Imelda
University of Uyo
imeldaudoh@uniuyo.edu.ng
Ekpenyong, Moses
University of Uyo
mosesekpenyong@uniuyo.edu.ng
Urua, Eno-Abasi
University of Uyo
eno-abasiurua@uniuyo.edu.ng
Adeniyi, Harrison
Lagos State University
harrison.adeniyi@lasu.edu.ng
Obiamalu, Greg
Nnamdi Azikiwe University
go.obiamalu@unizik.edu.ng
Yusuff, Ayo
University of Lagos
yoyussuf@yahoo.co.uk
Anyanwu, Ogbonna
University of Uyo
ogbonnaanyanwu@uniuyo.edu.ng
Obikudo, Ebitare
University of Port Harcourt
ebitare.obikudo@uniport.edu.ng

Abstract

Being a multilingual and multicultural nation, Nigeria is blessed with over 525 languages (Blench, 2014) from four different language families. The sheer number of indigenous languages makes an interesting tapestry! Unfortunately, not much attention has been paid to the study of our indigenous languages, with all the abundant prospects. This paper is a work in progress and a recently funded research project by the Tertiary Education Trust Fund (TETFund) of Nigeria, to rescue language endangerment in Nigeria by redefining the ecology of languages through geospatial technologies; concentrating on those languages located in the southern part of Nigeria, and scalable to other regions of Nigeria and beyond. The main objective of the project is to implement a location-aware infrastructure or framework that functions proactively in real-time, for enhanced language ecology, and precise visualization of the language's vitality status for each language spoken in the region under study. The project draws a multidisciplinary team consisting of linguists and language experts, computational scientists, and geographical information system (GIS) specialists, to examine our language ecology through practical fieldwork and integrate same into a geospatial framework for the purpose of revitalizing our indigenous languages and making them readily available. This research project is therefore significant as it will not only provide a cooperative solution for advancing our understanding of the spatial pattern that describes language ecology, but also provide an early appraisal of the degree of vitality and endangerment in relation to factors such as social and cultural changes. The immediate impact is the prompt revitalization of our indigenous languages and the development of an effective language policy to strengthen our local heritage.

Keywords: Nigerian languages, endangerment status, geo-linguistic database, geospatial framework, language ecology app, location-based system.

1 Introduction

Language endangerment can be described as a process whereby the vitality of a language (i.e., the extent at which the language is applied as a means of communication in various social contexts and use in specific domains, or purpose) is no longer being pursued by the 'owners' of the language or language community, causing a threat of disuse or abandonment of the language. Endangerment can be caused by natural factors (wars, earthquakes, etc.) or social factors (e.g., abandonment). Today several Nigerian languages are gradually going into extinction because many speakers of the language have either abandoned their languages for other languages or have neglected the core responsibility of developing their languages to support intergenerational transfer (i.e., inability to transfer the language from one generation to another),

The endangerment status of a language can be classified into five levels as follows (Udoh and Urua, 2015 after Connell 1994):

- Moribund – no longer in use, nor transmitted, or are threatened.
- Retreating – dying in an area but flourishing in another area: inter-country boundaries.
- Under-developed – without orthographies, written literature, meta-language.
- Developing – with fairly-developed orthographies, and which literature tradition and meta-language are in the process of development.
- Infiltrated or pidgin (Blench, Spriggs and Connell 1999) – mostly used in informal conversation.

The integrity of a language is further enhanced when its status significantly permeates new domains, develops into a language of the educated class, and adopts new societal values (Urua and Ekpenyong, 2018).

Various indices permit the determination of the endangerment status of a language. The most dominant ones include: the rate of population growth; dominance by a more powerful language or a language with greater economic influence; and/or, the lack of adequate descriptive evidence of the language, hence limiting the development or effective use that language (cf. Bamgbose 1976; Fakuade 1999). Whereas the degree of language use may influence the determination of language endangerment, the various variables work in varying ways to improve the vitality of the language (Adewale and Oshodi, 2013).

Interests in pursuing a spatial solution to language infrastructure development have greatly declined. While most studies concentrate on the use of GIS for literary studies (Kretzschmar, 2013), only three studies to the best of our knowledge are situated within the context of our study and are discussed as follows.

In 2009, the GIS-based Linguistic Geography of Thailand Project was initiated under the sponsorship of Chulalongkorn University with its key aim to promote the use of Geographic Information System (GIS) in linguistics. The project enabled scholars from different fields of knowledge to work together, in this case, geographers and linguists. A series of research work has been conducted since then. The first in the series was the research work of the Word Geography Maps of Thailand project, producing a geographic database of 170 Thai dialect vocabularies based on the data collection in 2002-2003 (Teerarojanarat and Tingsabadh, 2008). The second was the extension work of the first project – the creation of the boundary map of Central and Non-Central Thai Dialects by overlaying 170 map layers (Teerarojanarat and Tingsabadh, 2011).

A recent work was the Word Geography Maps of the Northeastern Thai Dialect. In this work, Thai dialect vocabularies in the northeastern region of Thailand based on data collection in 1979 was converted and transformed under a GIS environment to be available in a digital map.

In Podobnikar et al. (2009), a detailed determination of the local speech areas which are parts of the Slovenian Linguistic Atlas (SLA) was performed. The SLA is a geo-linguistic project designed in the 1930s and published in 2010. One of the goals of this project was to analyze the impact areas and to geographically allocate local speech data for a more precise determination of the isoglosses. Various evidence that influenced formation of the speech were also analyzed with GIS, resulting in intra-linguistic and extra-linguistic indicators (i.e., geographic, and historic).

While the foregoing related works concentrate on aspects of map production and analysis, our project introduces a novel integration of location-aware information plus details of data that determine the vitality and endangerment status of languages, to enable real-time spatial information storage, updates, and retrieval. The purpose of introducing a spatial information system is to enable prompt visualization of stored language attributes, and enhance policy decisions, instead of relying on belated information.

1.1 The Problem

Nigerian languages have not been adequately harnessed. Although a few of them are developing, several are still at risk of endangerment and death. The indigenous languages archive the rich cultural values of the people and valuable knowledge systems, which are important for development. The death of the languages leads to the loss of the rich cultural heritage of the people. Many of the languages are non-vital because they are used in rather restricted domains, and many have even lost the privilege of intergenerational transfer, a serious threat to the languages, their speakers and humanity. Besides, we lack correct information on the number of languages, not to talk of updates on them. There is no absolute one-to-one mapping between languages and the increasing amount of data on dominant languages on the Web, as there appears to be an unbalanced relationship between languages, as the level of endangerment further widens without notice. Also, no defined criteria/models exist to give immediate and precise picture (in real-time) of how these languages fare and when endangerment happens.

Traditional archiving methods (e.g., use of static media – texts, non-animated images, maps) could not offer real-time processing and visualization of resources through location-based (or spatial) services, as they are limited by the opportunities for language development and internalization. Furthermore, not giving back the products of research to the language community has caused loss of interest in preserving and propagating them into the future. Hence, the urgent need to characterize existing linguistic knowledge and resources, to crucially influence a possible revived use or abandonment of linguistic structures and varieties is pursued in this project.

1.2 Project Objectives

The objectives of the project include:

- To investigate the vitality and level of endangerment of languages in the 3 Zones of Southern Nigeria using the UNESCO 2003 LVE (language vitality and endangerment) Scale, for efficient knowledge extraction.
- To investigate the relation between the languages and their speakers and ground truth these details to the specific geolocations.
- To develop a spatial database system of the languages spoken in the 17 States in the 3 Geopolitical Zones of Southern Nigeria, with descriptive information about their location, point clouds, and high-resolution satellite images.

- To integrate extracted information on the spatial system, embedding or mapping important language attributes, for real-time visualization of the language's vitality and endangerment status.

2 Conceptual Framework

The proposed conceptual framework as presented in Figure 1, explains the key concepts or variables of the study and their relationships. It begins with the physical environment and connects technology, for precise information system development. Within the physical environment, the main variable of our study is the spoken language where sociolinguistic, ecological and location parameters are discoverable for the purpose of building:

- spatial linguistic database (for location-based queries and analysis)
- natural language processing (NLP)
- tool development (for linguistic and model development) and
- GIS tool and linguistic database (to establish natural patterns and trends that document and process available spatial datasets as well as model relationships that exists between them).

The outcomes include precise knowledge of linguistic diversity, endangerment threats projection and informed policy decision.

3 Methods

This project will be conducted in the 3 Zones of Southern Nigeria made up of 17 States. The project will be multidisciplinary and shall involve a tripartite collaboration between the academia (staff and students from various disciplines), professional association (Linguistic Association of Nigeria) and the speaking community. To allow for efficient monitoring of the project, supervisors and coordinators will be appointed to supervise/coordinate the primary data collection and training in the respective geopolitical zones.

The Research Population: The research population covers the number of people in a community that will be visited by the field assistants in each LGA, as provided by the 2006 census.

Sample and Sampling Procedure: a multi-stage sampling will be adopted. A systematic sampling method will be used to select houses, where every tenth house on the right side of the street will be selected. A stratified random sampling will then be used to select from a range of recipients within the houses – elders, youths, men, women, children from households in the community.

The Primary Data: The primary data are the languages spoken by the communities in the 355 local government areas (LGAs).

The Data Collection Instrument: The 2003 UNESCO LVE instrument will be adapted for use to collect the primary data. The questionnaire has three main components: the meta data, the real data, and a reliability index. A 5-point scale will be designed to get responses from the communities in the following areas:

- Inter-generational transmission from parents to children
- The absolute number of speakers per language.
- The proportion of speakers within the total population.
- Loss of existing language domains.
- Response to new domains and media.
- Existence materials for language education and literacy.

- The policies of government and institutions concerning the respective languages.
- The attitude of community members towards their own languages
- The amount and quality of documentation the languages have enjoyed.

The reliability index on each of these issues will cover:

- Evidence from field work and direct observation.
- Evidence from other reliable sources.
- Very little evidence – a 'best guess'.
- No data available.

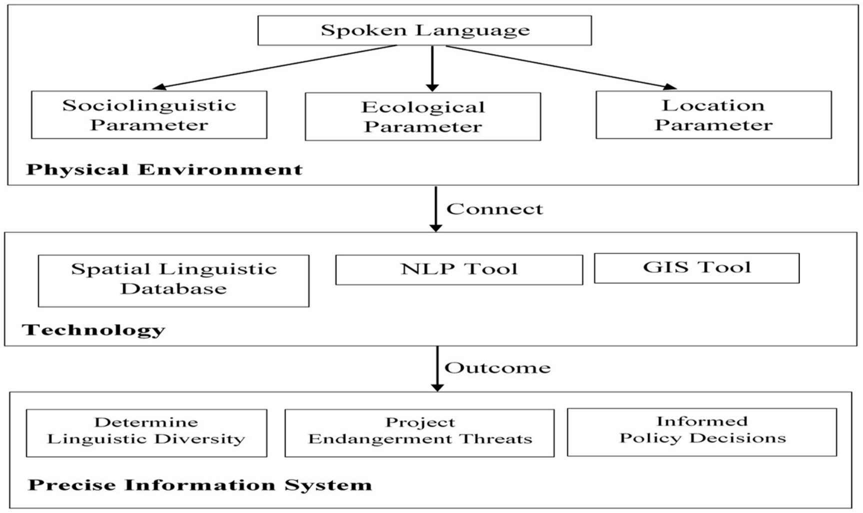

Spoken Language
Sociolinguistic Parameter
Ecological Parameter
Location Parameter
Physical Environment
Connect
Spatial Linguistic Database
NLP Tool
GIS Tool
Technology
Outcome
Determine Linguistic Diversity
Project Endangerment Threats
Informed Policy Decisions
Precise Information System

Figure 1. Proposed conceptual framework

Primary Data Analysis: The responses will be used to code the items in the questions in the questionnaire. These responses will be used to feed the GIS and statistical assessment of the linguistic ecology.

The methodology is divided into 5 major activities as summarized below:

Training of Coordinators/Research/Field Assistants: Training workshop will be organized for coordinators/research/field assistants and students involved in the project.

Fieldwork: Data for the study will be collected from 355 LGAs in the 17 States through fieldwork. Linguistic and location data including boundary information will be collected, processed, and documented as outlined above.

Spatial Database Design: A spatial linguistic database for storing linguistic and location information will be developed. The steps to accomplish this task include (1) satellite image (base map) acquisition and reconnaissance survey of study area to capture relevant location data and establishment of the start and end boundary, using GPS device; (2) Identification and abstraction of relevant features within the study environment.

Geo-Modelling, GIS Mapping and Test-bed Design: Mapping of the linguistic and location information will be achieved using ArcGIS. The steps to accomplish this process include (1) superimposing the GPS data on the extracted surface after image digitization. (2) geo-database modelling and integration; (3)

prototype test-bed system development for test datasets visualization.

Language Ecology App Design: Using the Python programming language, the user interface for communicating with the spatial database will be developed to enable real-time information storage and retrieval.

4 Expected Results

4.1 Outputs and Outcomes

The research outputs and outcomes of this project include:

- Open-Source Geo-Linguistic Database, for efficient community crowdsource and documentation of language resources in Nigeria.
- Language Data Model, for integration of extracted linguistic features into High-Definition maps.
- Geospatial (language ecology) App, for real-time visualization of available language resource, monitoring of the vitality and endangerment status of languages, efficient policy formulation, quick educational reference tool, and community-wide access.
- High quality publications and patent, for advancing linguistic research on language ecology, documentation, and endangerment

4.2 Dissemination

The project outcomes will be brought to the attention of key stakeholders through several avenues. There is a plan to host an international conference to do this. A project website will also be developed to share the project achievements, publications, and other resources, especially with the research community. An App will be developed for use by all. These efforts will contribute to redefining the linguistic ecology of our communities as well as improve the vitality of these languages.

4.3 Expected Impact

The immediate impact of this project will be a redefinition of approach to preserving and revitalizing Nigerian languages, especially the endangered ones. The proposed approach will strengthen the vitality of our languages, improve their ecology, and stem the tide of their death. Technologically this project will:

(1) enhance the investigation and quantification of language change across speaking groups;

(2) provide detailed analysis and trace of language change and contact;

(3) encourage community contribution of language resources.

To achieve efficient community participation/contribution, the proposed application will be open source, with access rights for manipulating data. Furthermore, resources can be deposited to mapped locations and retrieved from same in real-time.

With a rich repository enabled by community contributions, a spatio-temporal analysis of the language is possible.

The long-term impact of the project is that it will help policy formulation and implementation, and contribute greatly to the language policy, as well as confirm the languages spoken in the LGAs.

5 Conclusion

GIS has found usefulness in documenting the languages of the world, but its application is still evolving. The multidisciplinary approach for redefining language ecology proposed in this project is innovative, as it integrates technology to remodel our perception towards proactive solution to the endangerment problem, which hitherto was achieved via intensive fieldwork, some of which appear inconsistent. The solution this proposal will present will enable community efforts to develop language resources and could be upscaled to the other three zones of Northern Nigeria and beyond. The Ecology app will embed descriptions of culture, polity, and diversity of the language for effective visualization; hence, enabling cost-effective solution to the current language crises in Nigeria.

Acknowledgements

This research project is funded by the Tertiary Education Trustfund (TETFund), Nigeria, 2020 National Research Fund (NRF) Grant.

References

Adewale, R. K, Oshodi, B, 2013. Language Endagerment in Nigeria: Focus on Small Ethnolinguistic Communities in Niger State. Dialectologia, 11 (2013), 17-45.

Bambose, A (ed.), 1976. Mother Tongue Education: The West African Experience, London: Hodder and Stoughton.

Blench, R. M, Spriggs, M, Connell, B (eds.), 1999. The languages of Africa: Macrophyla proposals and implications for archaeological interpretation. Archaeology and Language 4. London: Routledge.

Connell, B., 1994. The Lower Cross Languages: A Pronegomena to the classification of the Cross River languages. Journal of West African Languages. XXIV, 1: 3-46.

Connell, B. 1998. Moribund languages of the Nigeria-Cameroon Borderland. Endangered languages in Africa. (Ed.) Brenzinger, M. Cologne: Koppe, 207-225.

Blench, R. M., 2014. An Atlas of Nigerian Languages. Oxford: Kay Williamson Educational Foundation.

Fakuade, G, 1999. Language Endangerment in the North-Eastern Part of Nigeria: Instances Language Empowerment: Theory and Reality, Aba: National Institute for Nigerian Languages, 50-66.

Haruna, A., 1998. Language death: The case of Bubbure in Southern Bauchi Area, Northern Nigeria.' Endangered languages in Africa. (Ed.) Brenzinger, M. Cologne: Koppe. Pp 227-251.

Kretzschmar, W.A., 2013. GIS for language and literary study. In: Siemens, R. and Price, K. (eds.) Literary Studies in the Digital Age: An Evolving Anthology. Modern Language Association: New York, NY, USA.

Podobnikar, T., Skofic, J. and Horvat, M., 2009. Mapping and analysing the local language areas for Slovenian linguistic atlas. In Cartography in Central and Eastern Europe (pp. 361-382). Springer, Berlin, Heidelberg.

Teerarojanarat, S. and Tingsabadh, K., 2011. Using GIS for linguistic study: a case of dialect change in the northeastern region of Thailand. Procedia-Social and Behavioral Sciences, 21, pp.362-371.

Udoh, I., Urua, E-A., 2015. Sustaining Nigerian Languages Through Use, Technology and Education Education in the 21st Century: A Festschrift in Honour of Professor Mrs. Comfort Ekpo, Urua, E., Udofot, I. M. & Uduk, H., University of Uyo, Uyo, pp. 62-73.

Urua, E-A., Ekpenyong, M, 2018, Strengthening Language Documentation in Africa through Effective Cooperative Research. In: Leonard, J. L. and Rialland, A. (eds). Linguistique Africaine - Perspectives Croisees. Journees scientifiques de la societe de linguistique de paris, 113-142.

Canonical Segmentation and Syntactic Morpheme Tagging of Four Resource-scarce Nguni Languages

Du Toit, Jakobus S.
Puttkammer, Martin J.
CTexT®, North-West University
Jaco.duToit@nwu.ac.za
Martin.Puttkammer@nwu.ac.za

Abstract

Morphological analysis involves investigating the syntactic class of a word but can also extend to the decomposition and syntactic analysis of its underlying morpheme composition. This is especially relevant to languages with an agglutinative writing system where multiple linguistic words are expressed as a single orthographic word. In this paper, we propose a memory-based approach to canonical segmentation using a windowing approach to recover the uncondensed morphemes that differ from the surface form of a word. Additionally, we propose treating the syntactic labelling of morphemes as a sequence labelling task, similar to part of speech tagging. This approach leverages the internal morpheme composition of a word as local context in much the same way that the surrounding sentence of word serves in the disambiguation of its part-of-speech. Both tasks are modelled separately but performed sequentially by cascading the decomposed morphemes of a word into the task of syntactic labelling. When evaluated on four resource-scarce, conjunctively written Nguni languages, the proposed approach achieves an overall accuracy ranging between 82% and 92% which outperforms previously developed rule-based analysers for the same languages.

Keywords: morphological analysis, canonical segmentation, syntactic morpheme tagging

1 Introduction

Access to quality linguistic resources is crucial to any research and development efforts in the field of Natural Language Processing (NLP). Collecting and assembling such resources is generally expensive and time-consuming, especially in the case of resource-scarce languages as is the case for ten of the eleven official languages of South Africa. To address this scarcity and promote research and development efforts in the area of human language technology, the South African government has established several legislative frameworks that promote the use and advancement of these languages and has funded various development projects over the last two decades. One such project was the National Centre for Human Language Technology's Text project (Eiselen and Puttkammer 2014), that collected circa 50,000 tokens, annotated with part of speech (POS), lemma and morphological analysis information for ten of the official South African languages. POS taggers, lemmatisers and rule-based morphological decomposers were also developed. A follow-up project expanded on this work by developing an additional 50,000 annotated tokens and improved technologies for four of these languages, *viz.* isiNdebele (NR), Siswati (SS), isiXhosa (XH) and isiZulu (ZU). In this paper, we focus on the morphological analysers that were developed in the follow-up project.

In the next section we provide a brief overview of the project, followed by a description of morphological decomposition versus analysis. We then (section 4) present our two-tiered approach to morphological analysis and results (section 6). Conclusions and future work are discussed in section 7.

2 Background

The aim of this project was to create corpora annotated with POS, lemma and morphological analysis information for four resource-scarce South African Nguni languages and to develop associated core technologies using these datasets. In this paper the focus is on the development of the morphological analysers. For a detailed description of the development process followed during the annotation of the data, see Gaustad & Puttkammer, 2021. A description of the other core technologies can be found in Du Toit & Puttkammer, 2021.

South Africa is a linguistically diverse country with eleven official languages, nine of which are Southern Bantu languages that follow either a conjunctive or disjunctive writing system. This study only focuses on the conjunctively written Nguni languages namely, isiNdebele, Siswati, isiXhosa, isiZulu. To illustrate the orthographic difference between these writing systems, Prinsloo & De Schryver (2002) use the example phrase "I love him/her". The agglutinative nature of isiZulu produces the phrase as a single word, *ngiyamthanda*, whereas it would be written disjunctively in Sepedi as four separate words, *ke a mo rata*. Although this study is restricted to the four South African Nguni languages, insights gained during development could also be to be applied to languages that share a similar morphology.

3 Morphology

Morphological features are an important aspect in developing language engineering technologies and applications such as machine translation and spelling error correction. Morphological features typically refer to either lexicalized (e.g. lemmas) and non-lexicalized features (e.g. gender, case, number). Most approaches that operate at word-level annotate tokens with lexicalized features using either combined feature sets or by modelling individual features as separate tagging tasks.

A combined feature set expresses multiple aspects as a single composition of non-lexical tags (e.g. Noun+A3sg+Pnon+Nom). This allows for underlying relationships between different features to be modelled explicitly but it does result in a large target space that further increases sparsity for morphologically rich languages. On the other hand, modelling these features separately yields a smaller target space but constrains the capacity of a model to learn any inter-feature dependencies which is crucial for our intended task. When operating at the morpheme-level, the functional role of a morpheme in the Nguni languages is both influenced and constrained by its internal context. Two approaches that utilise morphological features are so called morphological decomposition where morphemes are identified, and morphological analysis, where tags are assigned to each morpheme based on its grammatical function.

Morphological decomposition entails dividing a word into its constituent morphemes, the smallest meaning-bearing units of a word (Ruokolainen *et al.* 2013). However, these morphemes may not be orthographically equal to the corresponding segment of the word in written form when spelling transformations manifest during agglutination. We thus distinguish between two forms of segmentation, surface segmentation and canonical segmentation. The former yields a set of substrings that concatenate to the original word from, whereas the latter yields a sequence of canonical morphemes that are true to the underlying forms of the morphemes but potentially differ in their orthographic representation within the original wordform. Morphological decomposition is beneficial in helping to support further analyses for NLP tasks, especially in resource languages where data sparsity can undermine the quality of a task.

The set of decomposers for conjunctive languages previously developed as part of the NCHLT project were rule-based implementations that follow the work of Bosch *et al.* (2006). These implementations entailed recursively identifying all affixes in a token whereafter the remaining constituent would be verified against a lexicon of roots and stems. Only in instances where a valid stem or root and a valid combination of affixes were confirmed would the decomposition be deemed successful. These rule deductions were based on the collection of 50,000 annotated tokens developed as part of that project. These and the newly developed decomposers split tokens into their constituent morphemes, including each constituent affix, roots in the case of verbs, and stems in other parts of speech. For example, the isiZulu word *ukusebenzisa* ("use") is split into its constituent morphemes as *u-ku-sebenz-is-a*, where each affix boundary is marked in conjunction with the verb root.

Morphological analysis is of particular importance when applied to the Nguni languages since its words are naturally composed of aggregating morphemes that may undergo spelling alterations after their unions. These languages follow a conjunctive writing system which leads to an agglutinative orthography where morphemes are written unseparated. Yet, the meaning of a word is a function of its morphemes, and it is therefore necessary to isolate these individual morphemes for further syntactic analysis. Full morphological analysis entails both the segmentation of morphemes and the analysis of the interactions among the underlying morphemes of a word by determining their syntactic classes (Van den Bosch and Daelemans 1999). For example, the same isiZulu word used in the previous example, *ukusebenzisa* ("use"), is morphologically decomposed and analyzed as u[NPrePre]-ku[BPre]-sebenz[VRoot]-is[CausExt]-a[VerbTerm], where the syntactic class of each morpheme is assigned. For this example, the syntactic classes consist of a secondary noun prefix (NPrePre), a primary noun prefix (BPre), a verbal root (VRoot), a causative verbal suffix (CausExt), and a verb terminative (VerbTerm).

4 Morphological Analyser Design

This section describes the task of modelling morphological analysis and the level of granularity in analyses that existing solutions support. Our initial investigations into a suitable approach to morphological analysis included NLP pipelines that typically accommodate multiple tagging and morphological tasks, as in the case of UDPipe[1] (Straka 2018) and MarMoT[2] (Müller *et al.* 2013). MarMoT is a generic CRF framework capable of both POS tagging and morphological analysis, similarly UDPipe is a trainable pipeline for tokenization, POS tagging, dependency parsing, and morphological analysis that employs contextualized BERT embeddings. However, the capacity for morphological analysis in these solutions only accommodate non-lexicalised features that distinguish lexical and grammatical properties of words. Attempts to adopt this system to annotate Nguni language tokens with syntactic morpheme classes were expectedly unsuccessful, since its capacity for morphological analysis is limited to only the word-level and context is derived from the encompassing sentence rather than the local composition of morphemes.

Taking into account the required depth of morphological analysis and because canonical segmentation and class annotation at the morpheme level are not usually addressed as a single task, we approached morphological analysis as two separate problems. Both morpheme segmentation and morphological analysis are treated as a sequence tagging task, as was the approach followed by Sorokin & Kravtsova's (2018) for Russian, a comparably agglutinative language.

In their approach, morpheme segmentation was represented using the BMES labelling scheme where the classes account for beginning (B), middle (M), and ending (E) as well as single (S) single letter morphemes. Additionally, morphemes were tagged according to their type namely, root, prefix, suffix, ending, postfix, link, and hyphen. Thus, the task of their system was to predict segmentation and type labels for a sequence of letters. However, morpheme classifications constitute a small target space of only 7 labels thereby ensuring better prediction accuracy. Adopting the approach taken by Sorokin & Kravtsova (2018) to our task resulted in a low tagging accuracy given our larger target space and subsequent increased sparsity as our labelling scheme contains 71 tags for ZU, 70 for NR, 68 for SS and 62 tags for XH.

The before-mentioned NLP pipelines and other approaches to morphological analysis typically model morphological features at the word-level and context representation spans the entirety of the encompassing sentence. In contrast, morpheme segmentation is modelled at the character-level and thus the difference in granularity explains why most approaches approach each task separately unless the target space can be kept relatively small as in the case of Sorokin & Kravtsova (2018). In contrast, Van den Bosch & Daelemans (1999) was however successful at jointly modelling morphological boundaries, syntactic classes, and spelling transformations at the character-level as a single

[1] https://ufal.mff.cuni.cz/udpipe/2

[2] http://cistern.cis.lmu.de/marmot/

task which resulted in a large target space but was supported by a large dataset. This approach is compatible with our intended task but applying a similar method to our limited data sets would introduce sparsity and only diminish the quality of predictions. Nonetheless, inspiration was taken from Van den Bosch & Daelemans (1999) in reducing the task of segmentation to a sequence of classification tasks using a windowing method across the characters of a word. To accomplish tagging each morpheme with its syntactic morphological class we opted to treat the task similar to that of POS tagging and representing individual morphemes as words to realise local context from within the internal morpheme composition of a word. This approach is discussed in greater detail in the next section.

The proposed approach follows a two-tier design for sequential segmentation and analysis. Isolating the two tasks allowed for greater accuracy in segmentation, which suffered the most in prior attempts given the complexity of the languages. A pipeline approach is subject to cascading errors, but the significant accuracy of the morphological decomposer as the initial component in the pipeline does little to impact the overall accuracy.

4.1 Tier 1: Morphological decomposition

The first tier of the approach is a memory-based learning system that models morphological decomposition and spelling transformations as a series of classification tasks inspired by the work of Van den Bosch and Daelemans (1999). Memory-based learning is a class of supervised, inductive machine learning algorithms that learn based on examples of a task stored in memory. The Tilburg Memory-Based Learner (TiMBL) facilitates this function as an open-source software package that supports a selection of k-nearest neighbour classification and feature weighting algorithms (Daelemans *et al.* 2004). When new instances of a learnt task are presented to the TiMBL model, computational effort is invested in finding the best-matching instances from memory as determined by a similarity metric. Once the nearest neighbour (or instance) is identified in memory, the associated class is transferred to the new instance. Memory-based approaches have been successfully applied to other natural language processing tasks such as hyphenation and compounding analysis (Pilon *et al.* 2008).

Morphological decomposition can be treated as a context sensitive mapping problem, similar to most linguistic problems (e.g. source to target language translation, text to speech synthesis etc.). As part of this approach, TiMBL is tasked with learning this particular mapping through a windowing method from the surface form of a word to its canonical segmentation. Windowing in this manner transforms a word into multiple instances where each instance is focused on the boundary between letters, or the start and end boundary of the word. The method is illustrated in table 1, where a sliding context window of 6 letters traverses the length of the word with the point of focus positioned 3 letters left and right of the given boundary. The instance class expresses whether the given boundary maps to a point of segmentation in its decomposed form and if any letters within its right 3 letter window should undergo a spelling transformation in obtaining its canonical segmentation. A window size of 6 letters was found to be sufficient for the task since most conversion-type transformation rules range between 1 and 3 characters. In the end, the canonical segmentation is obtained by constructing the surface form of the word according to the predicted transformation rules.

Table 1: Segmentation rules for the word ngokuphathelene

Instance Number		Left Context	Point of Focus		Right Context		Rule Class
1	-	-	-	n	g	o	=
2	-	-	n	g	o	k	=
3	-	n	g	o	k	u	=
4	n	g	o	k	u	p	o>a*u*
5	g	o	k	u	p	h	=
6	o	k	u	p	h	a	*
7	k	u	p	h	a	t	=
8	u	p	h	a	t	h	=
9	p	h	a	t	h	e	=
10	h	a	t	h	e	l	=
11	a	t	h	e	l	e	=
12	t	h	e	l	e	n	=
13	h	e	l	e	n	e	0>an*il*
14	e	l	e	n	e	-	=
15	l	e	n	e	-	-	ne>0
16	e	n	e	-	-	-	=

Per illustration of the approach, Table 1 contains 16 instances that were generated from the isiZulu word *ngokuphathelene* ("in relation to") with their associated morphological transformation and segmentation classes that produce its canonical segmentation. The generated classes can express five different types of transformations, the first is represented in instance 1 with the "=" class. This indicates that no transformation or segmentation takes place at the current point of focus in the original surface form. These classes denote no difference between the surface form and the canonical segmentation for the right context at the point of focus. The second type of class is represented in instance 4 (o>a*u*) which is indicative of a conversion transformation. These classes are assigned to the letter immediately left of the point of focus but can include letters from the right context when a transformation affects multiple letters like in instance 15 (ne>0). Any asterisks contained within the class represent segmentation points in the canonical form. Instance 6 (*) depicts an asterisk as an independent class which denotes a segmentation point in the canonical form of the word at the current point of focus. The fourth type of classification depicts the insertion of characters at the given boundary as in instance 13 (0>an*il*) where the letters and segmentation points "an*il*" are to be inserted between the letters l and e. The fifth and last class denotes the removal of letters like in the case of instance 15 (ne>0) where the trailing letters "ne" are to be omitted in the canonical segmented wordform.

The classes were derived using diminishing longest string matching between the surface and canonical forms to isolate the differences at character level. This yielded an instance base ranging between 447,605 (SS) and 481,153 (NR) that consist of 98 (NR), 122 (SS), 96 (XH), and 124 (ZU) unique classes, but excluding exceptional classes with an occurrence frequency of less than 3. Across all four languages, the most frequent classes are (=) and (*), making up just over 50% of all class instances in each language.

In order to determine which learning algorithm best served the task of canonical segmentation, TiMBL was trained on the generated instances using each of the five k-nearest neighbour algorithms that it accommodates. A parameter search was used as part of this experiment to determine which hyperparameter adjustments could provide the greatest prediction accuracy. Since the two-tier approach relies on the predicted segmentation, it was important to obtain a reliable morphological decomposition to ensure as few errors as possible that may hamper the second tier's capacity to reliably annotate each morpheme with the related morphological analysis. In the end, IB2 was found to provide the greatest accuracy. IB2 operates similarly to other memory-based learning algorithms by keeping instances in memory that contribute to the potential classification of unseen instances during learning and uses a distance metric to determine class association. IB2 however employs an incremental editing strategy where its instance base is seeded with a certain (typically small) number of instances and only adds to instances in memory when it is misclassified by the k-NN classifier. The intention behind this approach is to construct an instance base that naturally establishes boundaries or key instances within memory with deviating or atypical instances to allow for greater generalisation. To further improve generalisation, each windowed instance was also associated with a seventh feature namely, the actual POS tag of the token. This improved the accuracy of predicted segmentation classes by around 1 to 2 percentage points.

4.2 Tier 2: Morphological Tagging

The second tier of the approach entails adopting a POS tagger to model syntactic morpheme classes of canonically segmented tokens. The chosen candidate for this approach is MarMoT, (Müller *et al.* 2013) given its trainable NLP pipeline and its successful application as a POS tagger on the considered Nguni languages in Du Toit & Puttkammer, 2021. Because morphological classes are context dependent, treating the task of syntactic morphological tagging of morphemes similar to that of POS tagging helps to take advantage of its capacity to learn predictions within the context of a sentence. This is achieved by treating each segmented morpheme as a word and sequentially ordering them to resemble a sentence, thereby realising the internal morpheme composition of a word as local context in their tagging predictions.

The segmented morphemes are provided to the tagger as words along with the actual POS tag of the word as an additional feature. This improved the quality of its syntactic morphological class predictions by increasing the tagging accuracy by around 1 to 3 percentage points.

5 Results

The annotated data developed as part of this project is divided according to an approximate 90% training and 10% test split. Without including any punctuation, the following token counts make up each dataset split.

Table 2: Token counts per language dataset

	NR	SS	XH	ZU
Train	39,251	37,223	37,926	38,489
Test	4,441	4,084	4,277	4,345

To evaluate the segmentation competency of TiMBL, the before-mentioned test sets were first transformed into windowed instances of 6 letters and segmentation rules were generated. Each instance was then associated with the transformation and segmentation rule as its intended classification. The instances also included the POS tag of the token as a seventh, additional feature alongside its 6 windowed letters. Finally, the four trained TiMBL models were evaluated by presenting the windowed instances for classification. By comparing the class predictions to the intended segmentation rules, a prediction accuracy ranging between 96% and 98% was achieved for each of the language-specific models. These results are listed in Table 3.

Table 3: Segmentation and transformation rule class prediction accuracy

	NR	SS	XH	ZU
Accuracy (%)	97.02	96.92	98.30	96.55

Similarly, to evaluate the syntactic morphological tagging competency of MarMoT, the test set of tokens were segmented into their intended morphemes and associated morphological tags. The test sets were then presented to MarMoT for tagging along with the POS tag of the token as additional feature. By comparing the predicted morphological tags to the intended tags, a prediction accuracy ranging between 91% and 96% was achieved for each of the language-specific models. These results are listed in Table 4.

Table 4: Prediction accuracy for syntactic morpheme tagging

	NR	SS	XH	ZU
Accuracy (%)	92.89	91.27	96.77	94.64

Finally, the combined accuracy of both tiers was evaluated by cascading the resulting canonical segmentation into MarMoT for syntactic morphological tagging. This was performed by applying the predicted transformation and segmentation rules in the first tier to the tokens in the test set to produce the canonically segmented morphemes. These morphemes were then tagged by MarMoT to obtain their syntactic morphological class. The results of the first tier were evaluated by comparing the number of corresponding morphemes between the intended test set of canonical segmentation and the predicted rule-based transformation of the token. Similarly, the morphological tag and morpheme associations were evaluated by comparing the predicted morpheme and tag pairs to the intended test set pairings for each token. These results are listed in Table 5.

Table 5: Canonical segmentation and morphological tagging prediction accuracy

	NR	SS	XH	ZU
Segmentation	86.71	84.94	94.13	86.87
Tagging	83.63	80.61	92.27	83.46

6 Conclusion

To ensure the continued development of human language technologies, it is important that resources be developed and distributed. We have described one such effort funded by the South African government for four resource-scarce Nguni languages. These resources are available as open-source modules from the SADiLaR resource catalogue[3] and can aid researchers and developers in improving and furthering the reach of language technology. Furthermore, the approaches taken toward morphological analysis at the morpheme level of agglutinative languages may also provide evidence for its viability and applicability to similar languages. The lexical resources provided as part of this project will enable further improvements and alternative approaches in developing related language technologies.

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

Bosch, S, Jones, J, Pretorius, L & Anderson, W 2006, 'Resource development for South African Bantu languages: computational morphological analysers and machine-readable lexicons' In *Proceedings on the Workshop on Networking the Development of Language Resources for African Languages at the 5th International Conference on Language Resources and Evaluation*, pp. 38-43.

Daelemans, W, Zavrel, J, Van Der Sloot, K & Van den Bosch, A 2004, TiMBL: Tilburg memory-based learner, version 6.4: reference guide, Tilburg, Tilburg University.

Du Toit, JS & Puttkammer, MJ 2021, Developing Core Technologies for Resource-scarce Nguni Languages. Manuscript submitted for publication.

Eiselen, R & Puttkammer, MJ 2014, 'Developing Text Resources for Ten South African Languages', In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 3698-3703.

Gaustad, T & Puttkammer, MJ 2021, Linguistically annotated dataset for four official South African languages with a conjunctive orthography: isiNdebele, isiXhosa, isiZulu, and Siswati. Manuscript submitted for publication.

Müller, T, Schmid, H & Schütze, H 2013, 'Efficient higher-order CRFs for morphological tagging.' In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 322-332.

Pilon, S, Puttkammer, MJ & Van Huyssteen, GB, 2008. 'Die ontwikkeling van 'n woordafbreker en kompositumanaliseerder vir Afrikaans', *Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, vol. 29 no. 1, pp. 21-41.

Prinsloo, DJ & De Schryver, GM 2002, 'Towards an 11 x 11 array for the degree of conjunctivism/disjunctivism of the South African languages', Nordic Journal of African Studies, vol. 11 no. 2, pp. 249-265.

Ruokolainen, T, Kohonen, O, Virpioja, S & Kurimo, M 2013, 'Supervised morphological segmentation in a low-resource learning setting using conditional random fields' In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 29-37.

Sorokin, A & Kravtsova, A 2018, 'Deep convolutional networks for supervised morpheme segmentation of Russian language' In *Conference on Artificial Intelligence and Natural Language*, pp. 3-10

Straka, M 2018, 'UDPipe 2.0 prototype at CoNLL 2018 UD shared task' In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 197-207.

Van den Bosch, A & Daelemans, A 1999, 'Memory-based morphological analysis' In *Proceedings of the 37th annual meeting of the association for computational Linguistics,* pp. 285-292.

Zalmout, N & Habash, N 2017, 'Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for Arabic' In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* pp. 704-713.

[3] https://repo.sadilar.org/handle/20.500.12185/7

New uses for old books: Description of digitised corpora-based on the Setswana language collection in the WITS Cullen Africana Collection

Rahlao, Malebogo
The University of the Witwatersrand, Johannesburg
thabonglebo@gmail.com

Lewin, Nina
The University of the Witwatersrand, Johannesburg
Nina.lewin@wits.ac.za

Surtee, Taariq
The University of the Witwatersrand, Johannesburg
Taariq.surtee@wits.ac.za

Abstract

This paper described a Corpus of 104 books. The books were catalogued into a standard library and archival metadata: Dublin core. A subset was digitised and cleaned. The books were then divided into five subsets and compared against each other and the entire Corpus. We speculated that the Corpus as a whole could be used as a general language register. Some examples are also given of the characteristics of the genre subsets. The paper aims to introduce the Corpus to natural language processing (NLP) researchers and offer it for further research.

1 Introduction

> "*The inability to communicate with human beings in their own language may be one of the most significant barriers to adopting information and communication technology in the third world and bridging the digital divide" (Getao & Evans, 2000: 128).*

At the same time, NLP is one of the most complex computational problems that have faced computer scientists. The calls for actions to develop resources for these languages, preferably in a coordinated and systematic manner, has been responded to in this paper. NLP researchers in English have had the advantage of many Corpora for research and testing. In this paper, we present the collection of the African language: Setswana sub-collection, for African Language NLP researchers in the 2nd Workshop on Resources for African Indigenous Languages.

1.1 Background

This project has a long history, from preserving books that would have been thrown out as outdated to creating a Corpus. We see it as unlocking the resources that many generations of archivists, librarians, researchers, authors, and translators have preserved in the belief that the books had value. Although, the individuals involved could not have imagined the computational uses in 1980. The William Cullen library: Africana Collection holds a subcollection of rare African language books dating from the 1800s in a locked area without researcher access. These books were originally the private collections of lecturers in the department. They were added to and carefully chosen by the African languages department as an internal library which was then given the Africana Library)[1]. These books provide a history of the development of various languages and varied orthography.

2 Collection Preparations

The first step was to catalogue the collection. There was an existing catalogue, but it was on physical cards and, we discovered, incomplete. We removed the books from the shelves and did a basic catalogue. We focused on creating a complete catalogue of books in Southern African languages. There is still an extensive collection in a variety of other African languages awaiting further research. The complete collection in Southern African languages is described elsewhere. We then compiled a metadata in Dublin core (and included some archival elements) of the Setswana books containing language varieties from Botswana and South Africa. The original catalogue designated these books as common readers. This paper describes the sub-collection of 104 Setswana books. These include Bibles, New and Old testaments, hymn books, prayer books, children's stories, grammar books, school books and novels.

[1] Margaret Atsango private correspondence 29-30/09/2021

3 Digitisation Methodology

The Avision book edge scanner in the Wits University Library Digitisation Centre converted all the materials into tiffs at high resolution. We used the international standards techniques and workflows adapted by the centre for the scanning of local content. After that, all the collected tiffs were cropped into Microsoft office 2010 documents. All scanned and cropped tiffs were then transferred into Abbyy FineReader 12 software. This software managed to split combined tiffs, skewed tiffs were deskewed, blank pages were removed, the pre-processing was applied, and lastly, they were saved in PDF format. The next step was to use the language recognition embedded in Abbyy FineReader 12 software to all the unrecognised PDFs by applying optical character recognition (OCR). This tool allowed the computer to recognise Setswana texts, moving from physical documents to text interpreted as data. This proves that the ground-breaking work of Otlogetswe in Setswana over many years (Otlogetswe, 2020, 2016, 2015, 2013, 2011a, 2011b, 2010, 2009a, 2009b, 2008; Otlogetswe and Chebanne, 2018; Otlogetswe and Ramaeba, 2014) which powered these recognition features is substantially effective. The recognition was challenging because the text was, like all older print, difficult, containing nonstandard fonts. Following the OCR process, we edited texts from the scanned documents as the OCR was approximately 80% effective in Setswana, with some unrecognised words. Such words were turned into different signs or symbols. Certain letters were misrecognised as other letters. We then cleaned the data by eliminating those signs and letters, replacing them with the correct Setswana letters manually. Pictures, text lines, and artefacts from the scanning process also had to be eliminated. The final product was then put into standard word documents.

4 Description of Corpora

The Corpus was created in Voyant (Sinclair & Rockwell, 2016). Basic descriptive tables as well as a Cirrus word cloud (see the appendices) were created. We created trends, collocations and correlation tables of the entire Corpus. To explore the linguistic varieties of the Corpus, we create five subsets: Bibles, Fairy tales, Grammar, Novels, and Poetry based on Genre.

The entire Corpus has 367149 total words and 23686 tokens. The appendices show a (10-line) sample of each.

Table 1: Example of Corpus Frequencies

Title	Words	Types	Ratio	Words/ Sentence
Bibles	57965	5164	0.089	17.984
Fairy tales	18980	2685	0.141	16.561
Grammar	9218	1774	0.192	10.030
Novels	34207	5799	0.169	19.435
Poetry	11550	3289	0.284	24.732

The average words per sentence in these subsets showcase Poetry as the highest (24.7) followed by Novels (19.4), Bibles (18.0), and Fairy tales (16.6), with Grammar (10.0) as the lowest. There is a case to made that the Corpus average of 19.3 could potentially be used as a language average, but this would require further research.

We also looked at the major terms

Table 2: Example of major terms

Term	**RawFrequency**	**RelativeFrequency**
batho	1305	3554.4153
kgosi	1169	3183.9934
modimo	967	2633.808
motho	764	2080.899
dira	690	1879.346
ja	618	1683.2402
bana	602	1639.6613
monna	579	1577.0165
utlwa	567	1544.3322
letsatsi	558	1519.8188
mosadi	502	1367.2922
bua	499	1359.1212
feta	475	1293.7527
tsaya	450	1225.6604
tloga	446	1214.7657
tsamaya	435	1184.805
ngwana	414	1127.6075
dikgomo	394	1073.1338
morafe	367	999.59424
motse	361	983.2521
lefatshe	354	964.1862

banna	315	857.9623
metsi	308	838.8965
tau	299	814.38324
baiseraela	276	751.7384
mokgwa	276	751.7384
lencwe	267	727.2252
mosimane	251	683.6461
nako	239	650.96185
ntlo	238	648.23816

4.1 Some features of the Corpus

The Corpus is able to support multiple types of research. The following were some initial features of interest that we found on an initial examination. The Corpus is still being annotated.

Orthography

This subset gives us information on how speech sounds form patterns. We were able to track changes because of the date range that our corpora spans. All our Bibles were written in an older orthography of the Serolong & Setlhaping dialect on a phonological level.. Our data also shows that most of our Bibles from the 1800s were written in the Serolong & Setlhaping dialect, known as the early orthographies of Setswana. Thus, early orthographies were based in Serolong & Setlhaping as the missionaries first created stellements among the Barolong & Batlhaping in Kudumane in 1821. In the older orthography, the letter [y] was used with the same pronunciation as the letter [j] used in the current orthography. For example, the words Yosefe (Joseph,), Yehofa (Jehova) and Yoshue (Joshua) are currently written as Josefa Jehofa and Jošua.

In the Serolong & Setlhaping dialect, people pronounce the phonemes / f / as [h] and / tsh / as [tšh]. For example, they pronounce the words 'world' and ‘resemblance’ as lehatshe and tšhobotsi, instead of lefatshe and tshobotsi.

Fairy Tales

Fairy tales are often intended for children, features fanciful and wondrous characters such as elves, goblins, wizards, and fairies. The term "fairy" tale refers more to the fantastic and magical setting or magical influences within a story rather than the presence of the character of a fairy within that story. The language register is simple because the writer speaks to children warning them in a casual and relaxed tone about a cannibal that eats children, for example. The vocabulary of the story is also easy for children to understand. Characters in fairy tales may be fairy folk or even talking animals, believable characters that children will care about such as a good-hearted hero, a scheming villain dimo (cannibal), in our example above), a wise helper, as so forth.

The word dimo (cannibal), is one of the highest frequencies (88) in the genre. Cannibals are often used in Setswana children's stories to scare them and warn them about the scheming villain that eats children. Many animals appear often in the genre. Setswana children's stories frequently consist of speaking animals that live with people. In most cases, an animal like nonyane (bird) is there to protect children from danger by keeping them under their wings while flying. Other examples of talking animals are phuduhudu (deer), phokojwe (jackal), and phiri (wolf). There are also human protagonists or participants in children's stories like Ntswakae, Ntitiagatsana, and Tsetsenyane.

The word ja (eat) appears 43 times in the genre as children are often eaten in these stories.

It is clear that the register of the Corpus is mostly formal. In particular, the register used in the grammatical genre. It is academic in the sense that the grammatical textbook shows the writing style of the Setswana language.

We were keen to work on this Corpus especially because many other language collections depend on written speech such as newspapers, radio, or tweets that are relatively informal in register.

It would be interesting to compare this to (Marivate et al., 2020) newspaper Corpus.

5 Conclusion

This collection sat, from approximately the mid-1970s, on dusty shelves in a closed room. It was not used because the library did not have the language skills to catalogue it, and it was a fragile collection. This collection was nevertheless chosen by linguists because of its high value across a range of African languages. The language resources often available to NLP researchers are newspapers and social media forms that contain informal speech. This collection has well-described genres of formal writing. It is often asserted that there are insufficient African language texts available for research. However, as this work shows, the resources do exist but require work to convert into digital format.

The tables attached are samples to encourage an NLP researcher to explore the Corpus. We believe that our data (comparisons, colocations, and trends) can serve as a resource to pick tones. There is further work to be done because many of the books of the Corpus are translations. This means that the collection can also, with some work, provide line by line translations for test datasets.

Acknowledgements

Katlego Khoetha

eResearch Committee member

The University of the Witwatersrand, Johannesburg

Katlego.Khoetha@wits.ac.za

Katlego helped troubleshoot the original translation and assisted from a computing perspective.

Stéfan Sinclair

We would like to recognise the memory of Stéfan Sinclair, McGill University, who imagined, designed, and programmed much of Voyant Tools until 2020, when he passed away. See obituary.

Peter Duncan

Special Collections Librarian

William Cullen Library (retired)

The University of the Witwatersrand, Johannesburg

Peter curated and advocated for the collection to be catalogued and used throughout his career. The collection came to the library from the private collections of lecturers and departmental collections of the African Languages at WITS.

References and Bibliography

Aitchison, J., 1992. Teach yourself linguistics. Hodder & Stoughton., London.

Alexander, N., 1999. English Unassailable but Unattainable: The Dilemma of Language Policy in South African Education. Individual papers available online at http://www.

Atkins, S., 1991. Corpus design criteria. ICAME Journal 1–31.

Berg, A.S., 2018. Computational syntactic analysis of Setswana (Thesis). North-West University (South Africa), Potchefstroom Campus.

Bock, Z., Mheta, G., 2014. Language, society and communication an introduction.

Brits, K., Pretorius, R., Van Huyssteen, G.B., 2005. Automatic lemmatization in Setswana: towards a prototype. South African Journal of African Languages 25, 37–47.

Butt, M., King, T.H., 2003. Grammar writing, testing, and evaluation, in: Farghaly, A. (Ed.), Handbook for Language Engineers. CSLI, Stanford, California., pp. 129–180.

Evert, S., 2005. The statistics of word cooccurrences : word pairs and collocations (PhD Thesis). Universität Stuttgart, Stuttgart.

Getao, K.W., Miriti, E.K., 2006. Computational Modelling in Bantu Language.

Kubler, S., 2004. Memory-based parsing. J. Benjamins, Amsterdam.

Linguist, n.d. Setswana Scrabble | Sunday Standard [WWW Document]. URL http://www.sundaystandard.info/setswana-scrabble (accessed 7.29.19).

Louwrens, L.J., 1991. Aspects of Northern Sotho grammar. VIA Afrika, Pretoria.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R., Modupe, A., 2020a. Low resource language dataset creation, curation and classification: Setswana and Sepedi -- Extended Abstract. arXiv e-prints arXiv:2004.13842.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R., Modupe, A., 2020b. Low resource language dataset creation, curation and classification: Setswana and Sepedi -- Extended Abstract. arXiv:2004.13842 [cs].

Masoke-Kadenge, E., Kadenge, M., 2013. 'Declaration without implementation': An investigation into the progress made and challenges faced in implementing the Wits language policy. Language Matters 44, 33–50. https://doi.org/10.1080/10228195.2013.837949

Mogapi, K., 1998. The teaching of Setswana. Gaborone.

Molosiwa, A., Ratsoma, N., Tsonope, J., 1996. A comprehensive report on the use of Setswana at all levels of Botswana's education system, in: Cross-Border Languages. Reports and Studies. Regional Workshop on Cross-Border Languages, NIED, Okahandja, Namibia. pp. 99–134.

Nfila, B.I., 2006. Standard in Setswana in Botswana (Dissertation). University of Pretoria.

Nordlinger, R., Bresnan, J., 2011. Lexical-Functional Grammar: Interactions Between Morphology and Syntax 112–140.

Otlogetswe, T., 2020. Beef cuts amongst the Bangwaketse: the case of motlhakanelwa. Anthropology Southern Africa 43, 233–245. https://doi.org/10.1080/23323256.2020.1836667

Otlogetswe, T., 2016. The design of Setswana Scrabble. South African Journal of African Languages 36, 153-161–161.

Otlogetswe, T., 2015. Treatment of Spelling Variants in Setswana Monolingual Dictionaries. LEXI 25. https://doi.org/10.5788/25-1-1299

Otlogetswe, T., 2013. Introducing Tlhalosi ya Medi ya Setswana: The Design and Compilation of a Monolingual Setswana Dictionary. Lex 23. https://doi.org/10.5788/23-1-1228

Otlogetswe, T., 2011a. Challenges to Issues of Balance and Representativeness in African Lexicography. Lex 16. https://doi.org/10.5788/16-0-653

Otlogetswe, T., 2011b. Populating Sub-entries in Dictionaries with Multi-word Unitsfrom Concordance Lines. Lex 19. https://doi.org/10.5788/19-0-449

Otlogetswe, T., 2010. Challenges to issues of balance and representativeness in African lexicography. lex 16. https://doi.org/10.4314/lex.v16i1.51493

Otlogetswe, T., 2009a. Setswana Sports Terms: A Genre Analysis. Marang: Journal of Language and Literature 19. https://doi.org/10.4314/marang.v19i1.42819

Otlogetswe, T., 2009b. Populating sub-entries in dictionaries with multi-word units from concordance lines. lex 19. https://doi.org/10.4314/lex.v19i1.49140

Otlogetswe, T., 2008. Corpus design for Setswana lexicography (Thesis). University of Pretoria.

Otlogetswe, T., Chebanne, A., 2018. Setswana, in: Kamusella, T., Ndhlovu, F. (Eds.), The Social and Political History of Southern Africa's Languages. Palgrave Macmillan UK, London, pp. 187–221. https://doi.org/10.1057/978-1-137-01593-8_12

Otlogetswe, T., Ramaeba, G., 2014. Developing a Campus Slang Dictionary for the University of Botswana. Lex 24. https://doi.org/10.5788/24-1-1267

Pollard, C., Sag, I.A., 1994. Head-Driven Phrase Structure Grammar. University of Chicago Press.

Pravec, N.A., 2002. Survey of learner corpora. ICAME journal 26, 8–14.

Pretorius, R., 2014. The sequence and productivity of Setswana verbal suffixes. Stellenbosch Papers in Linguistics Plus 44, 49–70. https://doi.org/10.5842/44-0-644

Pretorius, R.S., Posthumus, L.C., Kr??ger, C.J.H., Potchefstroom University for Christian Higher Education, 1997. Auxiliary verbs as a subcategory of the verb in Tswana.

Pretorius, R.S., Potchefstroom University for Christian Higher Education, 1997. Auxiliary verbs as a subcategory of the verb in Tswana.

Publications - Namibia Statistics Agency [WWW Document], n.d. URL https://nsa.org.na/page/publications (accessed 11.9.21a).

Publications - Namibia Statistics Agency [WWW Document], n.d. URL https://nsa.org.na/page/publications/ (accessed 11.9.21b).

Speech and Language Processing, 2nd Edition [WWW Document], n.d. URL https://www.pearson.com/content/one-dot-com/one-dot-com/us/en/higher-education/program.html (accessed 11.9.21).

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG, 2003. LANGUAGE POLICY.

Viljoen, B., Pretorius, L., Berg, A., Pretorius, R., 2008. Towards a computational morphological analysis of Setswana compounds. Literator : Journal of Literary Criticism, Comparative Linguistics and Literary Studies 29, 1–20.

Appendices – These are sample tables meant to introduce the types of data created from the Corpus.

Appendix 1 - Trends

Doc	Term	Raw	Relative	Z-Score	Z-Score	TF-	Distributions		
Index		Frequency	Frequency		Ratio	IDF			
0	ba	3647	62917.277	33.157654	-1017.6505	0.0	0.003450358	0.005520573	0.0055723283
4	ba	16906	45898.81	55.502193	-1703.4329	0.0	0.0065837344	0.005714953	0.00538373
2	ba	1471	43002.89	25.90773	-795.1412	0.0	0.00324495	0.0023679363	0.0019294296
0	le	2391	41249.03	21.703148	-666.09717	0.0	0.0059346156	0.0043819547	0.0039506597
1	go	378	41006.727	14.760752	-453.0262	0.0	0.004664786	0.0030375354	0.0042308527
2	go	1393	40722.66	24.528439	-752.80896	0.0	0.0038880932	0.004063496	0.003215716
4	le	14310	38850.816	46.971714	-1441.6216	0.0	0.0044199256	0.003885082	0.0037954887
5	ba	730	38461.54	13.78445	-423.06226	0.0	0.004056902	0.0072181243	0.0052687037
4	go	14079	38223.668	46.212646	-1418.3248	0.0	0.0031248983	0.0028588339	0.003700466
1	ba	344	37318.29	13.414559	-411.70987	0.0	0.004664786	0.0027120851	0.0018442179

Appendix 1 is a sample of Bible trends. There are ten distribution points. Three are presented here.

Appendix 2 – Collocated (example from Novels)

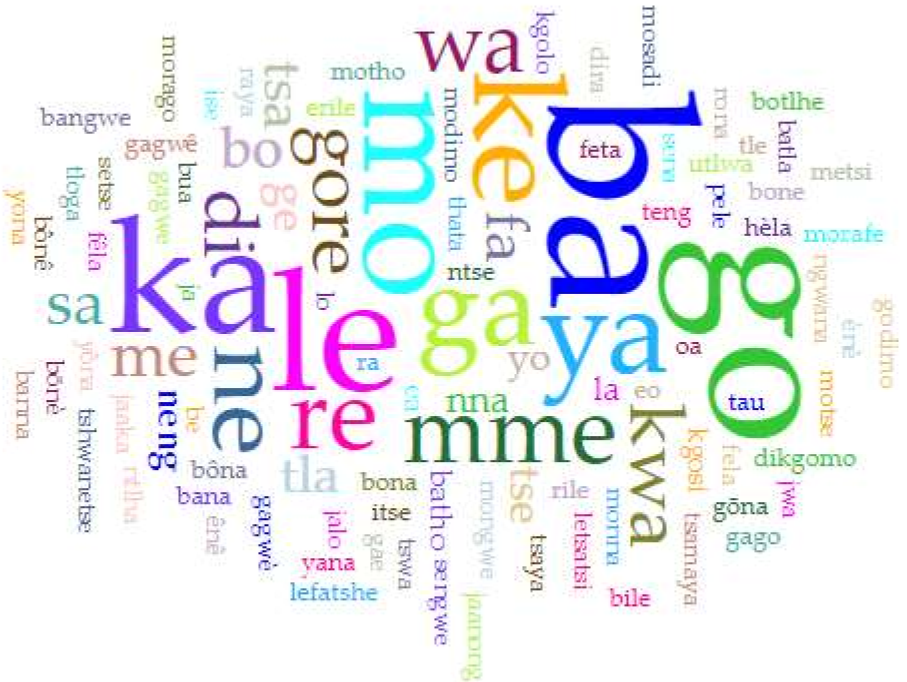
Term	Frequency	Context	Contextual Frequency
ba	1471	ba	1790
go	1393	go	736
le	1262	le	557
le	1262	go	502
go	1393	ba	497
ba	1471	go	486
ba	1471	le	480
go	1393	le	465
le	1262	ba	457
ke	802	ke	373

Appendix 3 – Correlation (example from Poetry)

Term 1	Term 2	Correlation
boÃªla	ngwale	1.0
matsodi	matsoke	1.0
diritibatsi	puiso	1.0
matute	morara	1.0
modikwadikwane	puiso	1.0
fofa	sefofu	1.0
ikaruse	sefofu	1.0
kitso	puiso	1.0
leo	matute	1.0
matute	semane	1.0

Appendix 4 – Cirrus word cloud – including the 105 most frequest words

Table 2 was derived from this list. We found the most frequent keywords.



Investigating the feasibility of harvesting broadcast speech data to develop resources for South African languages

Badenhorst, Jaco
Voice Computing Research Group, CSIR Next Generation Enterprises and Institutions Cluster, Pretoria, South Africa
jacbadenhorst@gmail.com

de Wet, Febe
Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
fdw@sun.ac.za

Abstract

Sufficient target language data remains an important factor in the development of automatic speech recognition (ASR) systems. For instance, the substantial improvement in acoustic modelling that deep architectures have recently achieved for well-resourced languages requires vast amounts of speech data. Moreover, the acoustic models in state-of-the-art ASR systems that generalise well across different domains are usually trained on various corpora, not just one or two. Diverse corpora containing hundreds of hours of speech data are not available for resource limited languages. In this paper, we investigate the feasibility of creating additional speech resources for the official languages of South Africa by employing a semi-automatic data harvesting procedure. Factorised time-delay neural network models were used to generate phone-level transcriptions of speech data harvested from different domains.

Keywords: low-resource languages, data harvesting, automatic speech recognition, TDNN-F, domain adaptation

1 Introduction

At least 10 of South Africa's official languages are currently regarded as resource-constrained. New ways to rapidly expand the resources required for ASR development are needed to unlock the full potential of applying state-of-the-art modelling techniques to these languages. Radio broadcasts represent one such potentially unlimited speech data source, which in South Africa is still largely untapped. Broadcasts typically contain various data types, including speech produced by different speakers in various speaking styles but also many non-speech events. For this reason, a data harvesting project was recently initiated in collaboration with the South African Centre for Digital Language Resources[1] to collect, curate and develop new corpora of speech data from broadcast data.

Speech data harvesting could be automated if automatic transcription systems existed. However, the level of ASR technology development in South Africa still has a long way to go before that would be a feasible option for data harvesting in the country. As a first step toward automatic annotation, initial acoustic models to bootstrap the harvesting work were derived from the existing, though limited, NCHLT Speech corpus (Barnard et al. 2014). The viability of using this data set to train deep neural network-based acoustic models was investigated in a previous study in which baseline systems for all 11 official languages were implemented using factorised time-delay neural networks (TDNN-F) (Badenhorst & de Wet 2019). This paper subsequently reports on the adaptation of these models using harvested speech data to improve the quality of the transcription system. The aim is to enhance the system's ability to automatically transcribe speech from various domains.

In a recent study Szaszáak & Pierucci (2019) investigated several approaches to adapt English TDNN-based acoustic models to Indian accented English. They tested different strategies including attempts to employ transfer learning, simple re-training of the entire network for a few epochs and applying i-vector adaptation in Kaldi (Povey et al. 2011). Apart from the speaking style, other differences between existing training data and newly obtained speech examples are important to consider. Regarding i-vector adaptation, for example, Vaněk et al. (2019)

found that a new Czech telephone model trained on short split utterances resulted in poor recognition of much longer real call centre utterances. Further analysis revealed the estimation of i-vector statistics with the TDNN models to be incorrect, resulting in failed adaptation. Conversely, i-vector adaptation in conjunction with data perturbation techniques have been shown to be an effective approach to improving TDNN-based acoustic models for low-resourced languages (Kumar & Aggarwal 2020).

Section 2 of this paper describes the data collection process and its associated challenges. Section 2.1 explains how new data were selected followed by a description of the test data sets that were defined for system evaluation in Section 3. Two of the 11 official languages were included in our investigation: Afrikaans (Afr) and Tshivenda (Ven). Section 4 introduces the important configurations of refined transcription systems, while Section 5 describes procedures to extract shorter segments of speech from raw data. Section 6 provides results for different system configurations. A discussion on the results (Section 7) and subsequent conclusions (Section 8), concludes the paper.

2 Data collection

Initially, it was anticipated that broadcast data from various radio stations would be streamed to a server. However, during the process of identifying possible data sources, other options also emerged. For example, some radio stations offered to provide physical copies of their content while others were willing to transfer data via a dedicated link. Another possibility identified was to transfer data from a content hosting service. A short description of these three data capturing scenarios are as follows:

1. Streaming data from the cloud to local storage
2. Transferring data from individual contractors
3. Transferring data from a hosting service.

In addition, various broadcast data sets could be obtained from each data capturing scenario. For example, some broadcasts contain accurately scripted news bulletins of high quality audio, but only for a limited number of news readers. Scripted news broadcasts were only available under data capturing scenario 2. There are also much larger quantities of unscripted audio data available. South African hosting services provide numerous radio show podcasts in all languages. Collection of broadcasts from individual contractors were severely impacted by the Covid-19 pandemic. Operations at smaller radio stations were diminished as the stations faced many new challenges. One university-based station even had to withdraw from the project. Agreeing on data sharing agreements was also an extremely slow and time-consuming process.

Fortunately, the collaboration with Iono[2] was successful. Agreements could be reached with a few radio stations producing and distributing podcasts on the Iono online audio platform. As a result, the resources collected for the two languages considered in this study could only be obtained under scenarios 1 and 3.

2.1 Content selection

In terms of the speech resources that are available for South Africa's languages, Afr is in a slightly better position than some of the other languages. It was included in the study to develop the harvesting strategy and to provide a broader perspective on different possibilities, given the large variety of podcasts. Ven was chosen as an example of a severely resource-constrained language. Podcasts were much more limited and no previously compiled broadcast news data were available as in the case of Afr. The material from which data were selected for harvesting included three speaking styles: read speech (e.g. news bulletins), conversational speech (e.g. radio dramas), and studio speeches (single speaker messages delivered in a studio).

Factors that were considered to select broadcasts for harvesting included availability, speaking style and the presence or absence of non-speech events like music, advertisements, etc. We also tried to limit the acoustic mismatch between the data to be harvested and the NCHLT data from which the initial acoustic models were derived. The NCHLT data consisted of clear, read speech prompts with minimal background noise. Therefore, read speech data such as the news or studio speeches were considered first.

For Afr, ample data were available for studio quality speeches. One radio show had more than 85 hours of clear recordings, a single speaker per episode and a sufficiently large number of speakers across all episodes. Speaker meta-data were also available for the episodes. Importantly, the podcasts of this show did not include any start or end jingles, and the duration of an episode was under five minutes, which considerably reduces the need for pre-processing and segmentation. The data represented a new speaking style of clear speech, which the presenter presented in a conversational manner. Data from this radio show was therefore considered suitable for testing NCHLT bootstrapping.

Unfortunately, no show of matching audio quality could be identified for Ven. A set of 1 118 Ven news bulletins that were recorded from streamed radio was therefore chosen as the best available match for the NCHLT data. Each recording included a news bulletin of approximately five minutes, which translates to almost 90 hours of data. The audio also included clips by reporters and news chimes.

3 Test data

Representative test data were required to evaluate the performance of baseline transcription systems as well as the impact of acoustic model adaptation to improve data harvesting for both Afr and Ven. Test sets were therefore selected from the available data, ensuring that the acoustic conditions and speaking styles that occurred in the data selected for harvesting were also represented in the test data. In addition to the existing Afr News test set, four test sets were transcribed manually: Ven News, Afr Messages as well as Ven and Afr Drama. The drama episodes were included because they contain speech produced by multiple speakers, mostly in a conversational style and in various acoustic conditions. The duration in hours and speaker distribution for each test data set is provided in Table 1.

Table 1: Duration and speaker information of test data

Test set	Dur (h)	Spk info
Afr News	7.89	18 male, 10 female
Afr Messages	0.36	4 male, 4 female
Afr Drama	0.82	multiple
Ven News	0.54	3 male, 2 female
Ven Drama	0.54	multiple

4 System refinement

Refined transcription systems were created to enable automatic transcription of the selected content introduced in Section 2.1. In this paper, we report results for the following three types of systems: 1) Baseline, 2) Text-based refined and 3) Acoustically adapted systems. Baseline systems utilised the initial NCHLT TDNN-F models (Badenhorst & de Wet 2019) in combination with a flat ARPA language model consisting of equiprobable phone uni-grams. Text-based refined systems employed language modelling given the limited text corpora that are available in the languages and acoustically adapted systems were built by updating the acoustic models within the above configurations using automatically transcribed data. The Kaldi toolkit was used to create all the acoustic models that were used in this study. It was also used to perform segmentation and acoustic model adaptation.

4.1 Text-based refinement

To configure the text-based refined transcription systems, language models were built from a number of text corpora produced during various previous projects for all 11 languages. Applying these texts to system development ensured that the systems developed for different languages would be comparable, since the same types of texts are available in each language. TTS prompts originating from the CSIR Lwazi projects proved most useful for configuring the baseline transcription systems. Subsequent combinations of the TTS prompts with text from the CSIR NCHLT Speech[3], CTexT NCHLT Text (Puttkammer et al. 2014*a*,*b*), and CTexT Autshumato (McKellar & Puttkammer 2020, Groenewald & du Plooy 2010) projects were also evaluated. Table 2 presents the unique word token counts (N=1) for each text in Afr and Ven. In addition, the counts of word 2-grams (N=2) provides some perspective on word sequences and the total number of tokens (T) in each text report on the size of the corpora.

The values in Table 2 indicate that the Afr Lwazi TTS text corpus contains more unique words (12 447) than the annotations of the NCHLT Speech data. The table also shows that the vocabulary size for the NCHLT Text corpus is almost four times that of the Lwazi TTS text corpus. Although the Afr component of the Autshumato parallel text corpus consists of a similar number of words than the NCHLT text corpus, its vocabulary size is substantially smaller (30 440 unique words).

In contrast, the Ven Lwazi TTS text corpus contained only 3 488 unique words, fewer words than the NCHLT speech data annotations. While the NCHLT Text corpus contained more than seven times the vocabulary of the Lwazi TTS text, compared to the Afr component, the vocabulary size was about half the number of unique words in Afr. The total number of Ven NCHLT Text corpus words was approximately 1 million. However, the larger texts may include more out-of-language words than the small, curated TTS text resources. Proper names and English terminology that occur in the NCHLT Ven text could contribute to the larger vocabulary sizes.

ASR systems with vocabulary derived from limited text cannot predict out-of-vocabulary (OOV) words in the test data. To better understand the impact of OOV words on text-based refined system results, Table 3 summarises the OOV rates for the TTS, NCHLT Speech, NCHLT Text as well as a combination of the NCHLT Text and TTS texts (Txt Corp + TTS). It is clear that predicting news data using the limited TTS text (especially in the case of Ven) would include more errors due to the larger OOV rate of these data sets. The values indicate far fewer OOV words for the larger NCHLT Text corpus. Moreover, the vocabulary of these texts produces similarly low OOV rates for both news and drama test data sets.

4.2 Acoustic refinement

The different speaking styles in broadcast data necessitates refinement of the acoustic models. In this work, the adaptation approach was based on re-taining the initial NCHLT model, but keeping NCHLT i-vectors intact. This meant that i-vectors for the adaptation data were also generated using the NCHLT i-vector extractor.

The approach featured a two-stage TDNN-F model adaptation recipe. For adaptation data, we used the automatic transcriptions of 6-gram text-based refined systems generated for the new data. The first stage of the adaptation recipe reproduced the training setup of the initial model, but instead of finalising the process left the training setup in such a condition that a second stage of training could be applied. The second stage is therefore an adaptation stage, re-training the standard (four-epoch) model for an additional epoch, now using the adaptation data. To enable model training and standard feature extraction, triphone alignments and data perturbation was applied to both the NCHLT and adaptation data. The initial NCHLT models were used to perform triphone alignment.

With re-training the intensity of the adaptive training was controlled, adjusting the learning rate for the training iterations of the last epoch and not the number of iterations (or training for more epochs). In essence, the standard start and end learning rate thresholds of the TDNN-F recipe across the initial four epochs of training were left to the same setting, restricting the training algorithm to these limits. For training of the last epoch, similar thresholds to the lower threshold of training (0.000020 and 0.000015 respectively) were applied. This restricted the algorithm to apply a learning rate close to the lower setting for the remaining iterations of training.

Table 2: Comparison of vocabulary sizes for text data sources

		Lwazi	**NCHLT**		**Autshumato**			
Language	**N**	**TTS**	**Speech**	**Text corpus**	**Words**	**Phrases**	**Eval**	**Parallel text**
Afr	1	12 447	8 565	56 192	11 785	1 226	3 015	30 440
	2	66 652	18 205	424 438	61	1 061	11 126	167 830
	T	**143 958**	**173 128**	**2 357 560**	**11 892**	**2 508**	**38 125**	**2 341 627**
Ven	1	3 488	7 578	24 314	4 435	994	2 877	-
	2	12 134	26 135	198 136	62	1 636	15 132	-
	T	**30 835**	**217 526**	**996 393**	**4 899**	**3 894**	**45 513**	-

Table 3: Out-of-vocabulary word rates (as a percentage) of different texts given the test data

	Lwazi	**NCHLT**		
Test set	**TTS**	**Speech**	**Txt Corp**	**Txt Corp + TTS**
Afr News	18.46	20.69	6.96	6.38
Afr Messages	6.28	17.48	3.83	2.86
Afr Drama	11.23	19.02	6.69	6.11
Ven News	21.66	14.50	6.44	6.44
Ven Drama	17.57	13.62	8.22	8.09

5 Data harvesting

As said in Section 2.1, data harvesting usually requires audio segmentation. ASR systems are trained on relatively short segments of speech, excluding non-speech events such as jingles and music. In this study, two different segmentation techniques were applied. The first, a speaker diarisation-based approach, enabled automatic detection of the start of the news inside the 10 minute Ven recordings. A second Kaldi alignment-based silence detection method was then applied to both the Afr and Ven harvest data. Sufficient segmentation was possible for detected silence labels of 0.1 seconds or longer in duration. The produced segments had durations between 5 and 15 seconds.

5.1 Segmentation

To detect the start of news (after the news chime), we applied a heuristic algorithm. The heuristic utilised an unsupervised implementation of speaker diarisation, as implemented in the Open-Source Python library for audio signal analysis (Giannakopoulos 2015). Applying the speaker diarisation to each recording and setting the predefined number of clusters to three and four speaker labels, respectively, resulted in separable classes. The other chosen parameter values of the heuristic were informed by a short analysis previously conducted on the manually segmented Ven news test set segments. As reported in Table 1, the Ven News reader segments had a total duration of almost 33 minutes and the mean and median duration of the segments was about 38 seconds. This meant that the typical news reader segment may be about 40 seconds or longer. Subsequently, the speaker cluster labels of the new recordings were validated to check if such long segments were detected within the first 120 seconds of each recording. For 44% of the recordings, no detections were made, so the minimum segment duration was lowered to 20 seconds instead. This lowered the non-detection rate further to 12% of the recordings. Given these findings, the following heuristic steps were implemented to select the first news segment at the beginning of the recording:

1. Select only those segments starting at least two seconds after the start of the recording, since inspection revealed that longer segments starting at time zero seconds usually contained music.
2. Select only segments with a minimum duration of 25 seconds.

3. Compute the overlap in time between segments created by the three- and four-speaker label diarisation for the first 180 seconds of audio.
4. For any overlapping segments detected during Step 3, set the start time to that of the first segment with an overlap of 50% or more.
5. Cut the audio files so that each file starts at the updated start time.

Applying the steps above ensured that for 90% of the recordings news start times could be detected. Spot checks confirmed valid news starts.

5.2 Segment selection

The complete Ven bulletin recordings included news clips, some of which contained English or another language. Furthermore, some or part of the news chimes might be included in a segment since the news not only started with, but also ended with the chime and music. Apart from these factors, the first automatic segment transcriptions would include labelling errors. In fact, the previous study that led to the development of the initial Afr acoustic model highlighted the importance of acoustic selection for imperfect speech training data, applying phone-based dynamic programming (PDP) scoring. Therefore, the adaptation experiments performed in this study were conducted by first including those audio segments that showed the best PDP scores given the new broadcast domains. To accomplish this, subset selections of the adaptation data were made. The PDP scoring technique requires two transcriptions of the same segment of audio. One transcription is usually produced by free phone recognition (employing a language model based purely on a string of phone labels). The second transcription comes from a better, more refined recognition system.

To create new speech corpora from the selected data (Section 2.1), the best automatic transcription systems were applied to transcribe each of the audio segments to be harvested. For each segment, a final PDP score was also estimated. In this study, the final refined systems were sometimes based on word recognition (see Figure 2) whenever lower error rates could be obtained this way.

Table 4: Baseline and better text refined systems PERs

Test set	Flat ARPA	6-gram ARPA
Afr News	20.98	16.34
Afr Messages	25.82	22.42
Afr Drama	42.69	39.52
Ven News	27.12	20.84
Ven Drama	45.80	40.35

6 Results

To test transcription performance, each of the test data sets were transcribed by the transcription systems described in Section 4.1. Subsequent phone error rates (PERs) were calculated including only speech phone labels during the estimation. Table 4 shows the performance achieved for two types of systems: 1) flat phone-based ARPA language model transcription and 2) systems applying 6-gram phone-based ARPA language models given the text data. In an iterative strategy, different ARPA-based systems were built for combinations of the text data. The PER values in Table 4 reflect the best 6-gram systems in combination with NCHLT acoustic models. Only the best 6-gram text refined systems in each language was chosen. This meant that the 6-gram language model was derived from the TTS text for the Afr systems and from a combination of the TTS and NCHLT Text corpus in the case of Ven.

Utilising the 6-gram language models, systems produced substantially lower PERs for all Afr and Ven test data sets. Similarly, results for a combination of the TTS and NCHLT Text corpus are presented for Ven. The results show higher PERs (above 40%) for Drama data compared to News data. In each case, the 6-gram ARPA systems deliver an improvement over just using a flat ARPA transcription system.

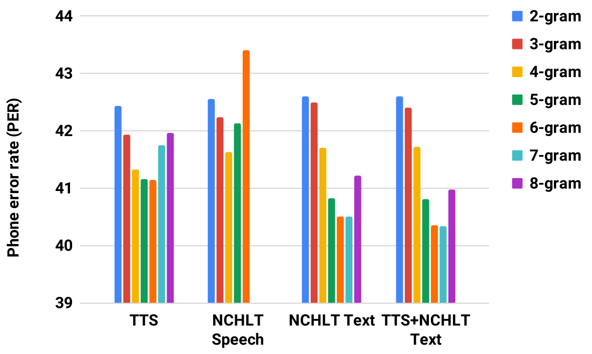


Figure 1: Ven PERs on drama data for various texts and context sizes

The 6-gram context size of the language models were chosen as a point where lower PER rates could be achieved in both languages for TTS and larger sets of text data. Figure 1 presents an example of such an analysis. Lower error rates were measured as the context size of the n-gram language models increased up to a point of about 6 or 7-gram phone contexts. Except for the NCHLT Speech text, this finding generalised well for the various combinations of text.

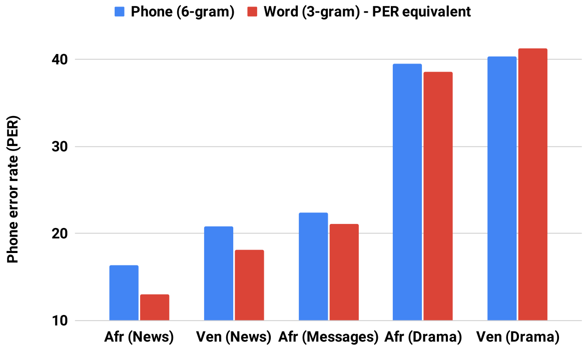


Figure 2: Comparing PER equivalents for phone and word recognition systems

Interestingly, even lower PERs could be obtained for most test sets when applying word recognition first. Using 3-gram language models, the automatic word-level transcriptions were converted to phone labels using a pronunciation dictionary. From the arrangement in Figure 2 it seems that the phone transcriptions for test sets with better transcription rates, such as those for the News and Messages data, benefited most from recognition at the word level. Only for the Ven Drama test data with a PER of over 40% word recognition did not provide a benefit.

Table 5: Comparing the PERs between NCHLT baseline and adaptively trained models on Afr Messages episodes.

	Flat ARPA		6-gram ARPA	
Test set	NCHLT	Adapt	NCHLT	Adapt
Afr News	20.98	21.25	16.34	16.56
Afr Messages	25.28	21.92	22.42	17.71
Afr Drama	42.69	43.70	39.52	39.87

Subsequently, a first iteration of acoustic refinement was applied to evaluate the potential benefit of acoustic model adaptation. Table 5 shows the effect that using the adapted acoustic models in conjunction with the phone language models had on recognition. In this experiment, almost the entire Afr Messages data set (71 hours, excluding the test data) was applied as adaptation data, treating each 3 minute episode as a single audio segment. Clearly, PERs were reduced more for the Afr Messages test data from the same speaking style. These results compared well with what was also produced using about 18 hours of adaptation data, if PDP scoring was applied to select shorter segments of well-transcribed data from the larger set for adaptive training first.

7 Discussion

Language models built from the vocabulary of the large corpora and combinations of texts produced fairly low OOV rates. The values in Table 3 showed that OOV rates of less than 10% across the test data sets representing different speaking styles from different sources (news, studio messages and drama episodes) could be achieved. While this fact may seem promising in terms of vocabulary coverage, PERs applying these language models did not reflect such significant reductions when employing the larger texts. Instead, fairly low PERs could be obtained using phone language models with adequate phone coverage, such as the models based on phonetically balanced TTS prompts alone. Only small improvements were seen for Ven where the size of the TTS prompts text was much smaller than that of the Afr. The analysis in Figure 1, where Ven drama transcriptions were analysed provides an example. It records less than 1% PER difference for employing additional text beyond the TTS text. Therefore, including more vocabulary from additional text corpora (other than those already included) next, might not benefit the construction of phone level systems built with baseline acoustic models as much as first thought. Transcribing the conversational speaking style test sets generated significantly more error than news reading test data. For continued data harvesting, the fairly high error suggests that more acoustic development is required to lower PERs, before building larger vocabulary language models. Another perspective on transcription error comes from analysis of the context size of the phone language models. Larger phone contexts played a significant role to lower the PER.

It was also shown that with larger context phone language models, larger reductions in the PER were achieved for the test sets where lower PERs had been obtained in the first place. This finding supports the idea that larger vocabulary systems require better (well-matched) acoustic models. In essence, the usefulness of including the information of text resources through language modelling in transcription systems increases as the acoustic modelling improves. Initial word recognition tests produced a relatively high transcription error for the test data. As with well-resourced languages, where acoustic models can be developed with sufficient audio data, incorporating larger texts should, in future, make a significant contribution to achieve more accurate recognition.

The adaptation strategy that was employed successfully produced significant PER reductions, adapting to the domain of the new Afr studio messages target data set. The speakers in the selected radio show spoke in a more conversational manner. A similar observation was made for the Ven language, where adapting to the news domain also produced some improvement but, as expected, this adaptation did not really improve Ven drama recognition. These findings proved the ability of the adaptation technique to adapt and develop TDNN-F acoustic models for new acoustic domains.

More surprisingly, these adaptations could also be achieved utilising less than 20 hours of data, choosing the best ranking PDP scored audio segments. While it is expected that the technique will produce a similar outcome when applied to data from other languages, future work should still focus on iterative training using the updated transcriptions that are produced. Given the unbalanced ratio of accurately-transcribed NCHLT training data to the automatically transcribed adaptation data samples, it was also necessary to keep using NCHLT i-vectors created by a speaker-specific transformation that is required for TDNN-F training. Additionally, this strategy has the advantage of being speaker independent and does not require speaker labels for the adaptation data. As more accurately transcribed adaptation data samples become available, it should be adjusted to include appropriate i-vector adaptation.

8 Conclusion

The results of this study indicate that the effectiveness of including text data as a language resource during automatic transcription system development depends on the state of the acoustic models employed for harvesting. While including a well-balanced within-language text (such as a set of TTS prompts) is important, acoustic models that are well-matched to the acoustic domain of the speech data to be harvested, are also required. Including larger quantities of text for language modelling then becomes more effective. While the above approach would create more data, reducing the transcription error would increase the yield of correctly-annotated speech data. It is recommended that new ways to improve the model development should be sought. Better acoustic adaptation should eventually be possible for sets of adaptation data with sufficiently accurate transcriptions.

The outcome of this study proved that the current data harvesting technique would be applicable to all official languages of South Africa. Good generalisation to speech data from a new domain, incorporating a new speaking style, was achieved for Afr. Secondly, the same approach could be duplicated successfully in the much more resource-constrained Ven language, producing PERs on broadcast news data that were similar to the PER obtained for the Afr data from the new speaking style.

Notes

[1] `https://www.sadilar.org/`, SADiLaR is funded by the South African government's Department of Science and Innovation (DSI).

[2] `https://https://iono.fm/` Iono maintains an online audio platform providing podcasts and audio live streaming services for a variety of sources.

[3] NCHLT Speech text comprised the ASR training prompts.

Acknowledgements

The data harvesting work was enabled by funding received from the Department of Science and Innovation through the South African Centre for Digital Language Resources (SADiLaR).

References

Badenhorst, J. & de Wet, F. (2019), 'The usefulness of imperfect speech data for ASR development in low-resource languages', *Information* **10**(9).
URL: *https://www.mdpi.com/2078-2489/10/9/268*

Barnard, E., Davel, M. H., van Heerden, C., de Wet, F. & Badenhorst, J. (2014), The NCHLT speech corpus of the South African languages, *in* 'in Proc. SLTU', St Petersburg, Russia.

Giannakopoulos, T. (2015), 'pyaudioanalysis: An open-source python library for audio signal analysis', *PloS one* **10**(12).

Groenewald, H. J. & du Plooy, L. (2010), 'Processing parallel text corpora for three South African language pairs in the Autshumato project', *AfLaT 2010* p. 27.

Kumar, A. & Aggarwal, R. K. (2020), 'Hindi speech recognition using time delay neural network acoustic modeling with i-vector adaptation', *International Journal of Speech Technology* pp. 1–12.

McKellar, C. A. & Puttkammer, M. J. (2020), 'Dataset for comparable evaluation of machine translation between 11 south african languages', *Data in brief* **29**, 105146.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. et al. (2011), The Kaldi speech recognition toolkit, *in* 'IEEE 2011 workshop on Automatic Speech Recognition and Understanding (ASRU)', number EPFL-CONF-192584, Hilton Waikoloa Village, Big Island, Hawaii.

Puttkammer, M., Schlemmer, M., Pienaar, W. & Bekker, R. (2014*a*), 'NCHLT Afrikaans text corpora'.

Puttkammer, M., Schlemmer, M., Pienaar, W. & Bekker, R. (2014*b*), 'NCHLT Tshivenda text corpora'.

Szaszáak, G. & Pierucci, P. (2019), A comparative analysis of domain adaptation techniques for recognition of accented speech, *in* '2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)', IEEE, pp. 259–264.

Vaněk, J., Michálek, J. & Psutka, J. (2019), Tuning of acoustic modeling and adaptation technique for a real speech recognition task, *in* 'International Conference on Statistical Language and Speech Processing', Springer, pp. 235–245.

Using MonoConc Pro to teach and learn lexical collocations in Xitsonga

Mlambo Respect & Matfunjwa Muzi
South African Centre for Digital Language Resource (SADiLaR), North-West University, Potchefstroom, South Africa
Respect.Mlambo@nwu.ac.za &
Muzi.Matfunjwa@nwu.ac.za

Abstract

Few language resources have been developed for indigenous languages in South Africa. Surprisingly, these are also official languages which constitutionally share the same language status as other languages. One of the major challenges for the development of basic language resources is the lack of digital corpora that can be used to train and develop the resources. Such a challenge has impeded the use of technology for research, learning, and teaching domains in indigenous languages. In this study, we used MonoConc Pro, a concordancer, to demonstrate how language users can utilise the software to display lexical collocation from a corpus for teaching and learning purposes. We illustrated how corpus can be used simultaneously with language technology to teach and learn aspects of linguistics in a form of lexical collocations in Xitsonga. An Autshumato Xitsonga Monolingual Corpus (AXMC) that was retrieved from the South African Centre for Digital Language Resources repository was used as data for analysis. The AXMC is a corpus that was collected and semi-automatically aligned at the sentence level during the Autshumato project. To search for lexical collocations, we interrogated the AXMC corpus using the MonoConc Pro program. A semi-automatic search for collocates of Xitsonga adjectives *lavakulu, letikulu, lavantsongo,* and *letinene* was conducted. The study found that lexical collocations or words that co-occur with adjectives are nouns, adjectives, possessives, and relatives. It was also observed that each adjective frequently collocates with certain nouns belonging to a specific class. The results obtained suggest a practical way in which language technologies can be used to explore corpora and examine language patterns for teaching and learning. We hope that this line of study will lead to the use of modern language resources to examine linguistic traits in indigenous languages.

Keywords: MonoConc Pro, Lexical collocations, Xitsonga, Concordancer, Language resources

1 Introduction

There is a significant gap in the development of South African official indigenous languages, particularly in the use of language resources for research, teaching, and learning. Some languages, such as English and Afrikaans, have received preferential attention in terms of development and use of contemporary language resources, whereas indigenous languages with fewer speakers, such as Siswati, isiNdebele, Tshivenda, and Xitsonga, are still in the process of development (Finlayson and Madiba 2002 as cited by Mlambo et al., 2021, p. 82). The absence of contemporary language resources and appropriate corpora required for building digital tools for indigenous languages has hampered research, learning and teaching in these languages using modern technologies. Furthermore, political, cultural, and socio-economic issues have hindered the development and integration of contemporary language resources into indigenous languages (Unesco, 2019). Meanwhile, languages with such resources and corpora have benefited in the research, learning, and teaching domains (Yunus, 2017, p. 297).

There have been some developments in South Africa with the establishment of language research infrastructure, such as the South African Centre for Digital Language Resources (SADiLaR), which provides corpora and builds digital tools for all 11 official languages. This centre facilitates the creation, administration, and dissemination of contemporary language resources, as well as relevant digital tools, which are made publicly available to communities for teaching, learning, and research purposes (https://www.sadilar.org). However, other languages are still underserved in some of SADiLaR's contemporary language resources.

Therefore, additional efforts are needed to develop and promote available language resources for all indigenous languages.

The study focuses on lexical word combinations known as collocations in linguistics. Collocations, or the co-occurrence of words, are crucial in language teaching and learning (Jafarpour, 2013, p. 57). We employ Xitsonga corpus, which is a language with a disjunctive writing system, to demonstrate how a concordancer can be utilised to teach as well as to learn Xitsonga lexical collocations. We pay attention particularly to the parts of speech that collocate with adjectives. Barlow's MonoConc Pro was used to illustrate how lexical collocations can be viewed and learned from a corpus using a concordancer. This technique was coined by (John, 1991) as Data-Driven Learning, in which learners or students of a language use electronic corpora to actively investigate a language and its grammatical patterns such as collocations and prepositional colligations (Lewis, 2000).

Concordancers such as WordSmith Tools, Multiconcord, and ParaConc have primarily been investigated in South African languages by scholars such as Taljard and De Schryver (2002), Madiba (2004), Moropa (2007), Ndhlovu (2016), Shoba (2018), and Mlambo (et al., 2021) for identification and extraction of terminology lists for dictionary compilations. Despite the benefits that have been observed, these concordancers have not been employed in teaching and learning grammatical aspects of Xitsonga. Therefore, this study attempts to close the gap in the use of corpus-based methods and concordancers to teach and learn lexical collocation in Xitsonga.

2 Related work

According to Moehkardi (2002), all languages contain words that frequently co-occur with other words(s) in units and their co-occurrences correspond to particular grammatical norms of a specific language, and such words are known as collocations. Benson et al. (1986) classified collocations into two types: grammatical collocations and lexical collocations. Grammatical collocations are made up of a noun, an adjective or a verb, plus a particle (a preposition, an adverb, or a grammatical structure such as an infinitive, gerund, or clause) (Bahns, 1993). The grammatical collocations are occasionally idiomatic and lexicalised as single units since their meanings do not correspond to the real meanings of the parts. Meanwhile, lexical collocations do not include prepositions, infinitives, or relative clauses, but instead consist of different combinations of nouns, adjectives, verbs, and adverbs (Bahns, 1993). However, in this study we are only interested in lexical collocations.

Pirmoradian and Tabatabaei (2013) researched Iranian English first language (EFL) students. Their study's goal was to examine the impact of using the Collins Collocation Dictionary (CCD) as a concordancer tool on Iranian EFL university students' acquisition of English lexical collocations. A total of 30 students was randomly assigned to one of the two groups: the experimental group and the control group. The experimental group was tasked with identifying miscollocations in 10 lexical collocations. Meanwhile, the control group was instructed to identify lexical collocations in the texts they were given. Pirmoradian and Tabatabaei (2013) discovered that using the CCD as a concordancer had a substantial impact on the subjects' overall performance of students in learning English lexical collocation. As CCD was used as a concordance tool, students' results on lexical collocation improved considerably when compared to those without the concordance tool.

The findings also demonstrated that there is a substantial difference between the impacts of concordance approaches and traditional methods on students when it comes to learning English lexical collocation for Iranian EFL learners. This means that students who utilise a concordancer to learn lexical collocations are more likely to outperform students who learn collocations through traditional methods.

The current study also draws from the research by Jafarpour et al. (2013), who compared the effects of the corpus-based approach with the effects of the traditional approach to learning collocations.

The study was conducted on two groups of English second language (L2) students. The experimental group implemented a concordancer, while the control group used a traditional technique. In terms of proficiency and collocation competency, the individuals in each group were at the same level when the experimentation began. Jafarpour et al. (2013) observed that using a corpus-based approach to teaching collocations and their use in writing is more beneficial for L2 learners than traditional techniques. In general, the findings revealed that using concordancers to teach collocations is more fruitful and efficient than using traditional collocation teaching methods.

From the literature consulted, no study has investigated lexical collocation in Xitsonga either with the aid of language resources or using a corpus-based approach. This reveals the critical need for employing modern language tools to explore grammatical trends in indigenous languages by using corpora. As a result, we believe that our study will pave the way for the use of concordancers and corpora to aid in the teaching and learning of Xitsonga lexical collocations and other linguistics traits respectively.

3 Methodology

This is qualitative research that employs an Autshumato Xitsonga Monolingual Corpus (AXMC) acquired from an online SADiLaR repository as data for analysis. The AXMC is a corpus that was collected and semi-automatically aligned at the sentence level during the Autshumato project. This project was funded by the Department of Arts and Culture and aimed to create tools, resources, and corpora for all South African languages (Groenewald & Fourie, 2009, p. 191).

MonoConc Pro which is a software invented by Michael Barlow was used to analyse the data. The software is utilised as a text-searching program with a user-friendly interface. Like other concordancers, MonoConc Pro has numerous capabilities such as constructing word lists (in both alphabetical and frequency order), creating concordance output, handling huge corpora quickly, providing collocation information, and operating with tagged or untagged texts (Reppen, 2001, p. 32). The primary purpose of using MonoConc Pro in our study is to demonstrate its utility on how lexical collocations can be viewed from a corpus using the concordancer for purposes of lexical competence in Xitsonga. To display collocations, the AXMC was imported in its text format into MonoConc Pro. Then, the 'search engine' function on the software was utilised to explore collocations of a particular adjective. When the search process was complete, all collocates of the entered adjectives were shown as results.

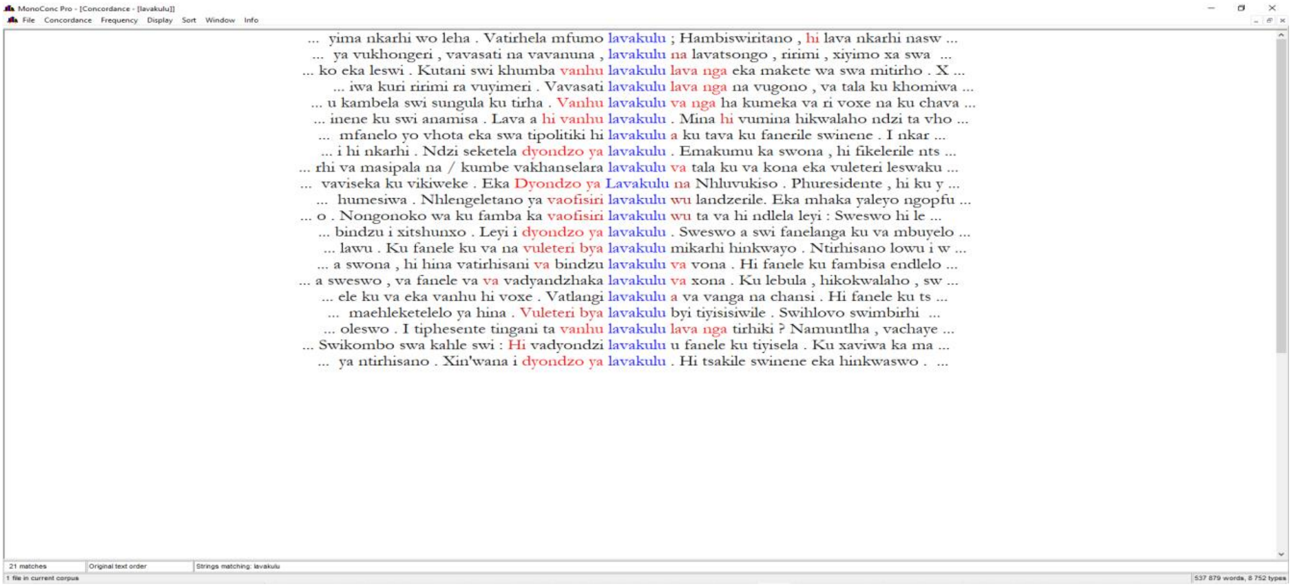
4 Data presentation and analysis

This section demonstrates how MonoConc Pro can be used to view lexical collocations for teaching and learning outcomes. MonoConc Pro was chosen for its functionality and features that enable us to find and visualise the lexical collocations in a language. The search engine function of MonoConc Pro was utilised to illustrate ways in which lexical collocations could be examined from corpora. A description of the operation of MonoConc Pro and its search engine feature in identifying collocations follows.

4.1 Search engine

The search engine is the main MonoConc Pro feature that was used to interrogate the corpus to view lexical collocations. To search for a specific adjective, we used the search function found on the MonoConc Pro interface bar. Searching for the adjective in Xitsonga resulted in the retrieval of all occurrences of the searched adjective together with its collocates. The target adjective was displayed in blue with the correct collocates highlighted in red on its left side. Using the search engine is beneficial in sorting words and viewing the collocation immediately to the left or right of the target word as this gives us vital information on how the word has been used in the context (Reppen, 2001, pp. 33-34).

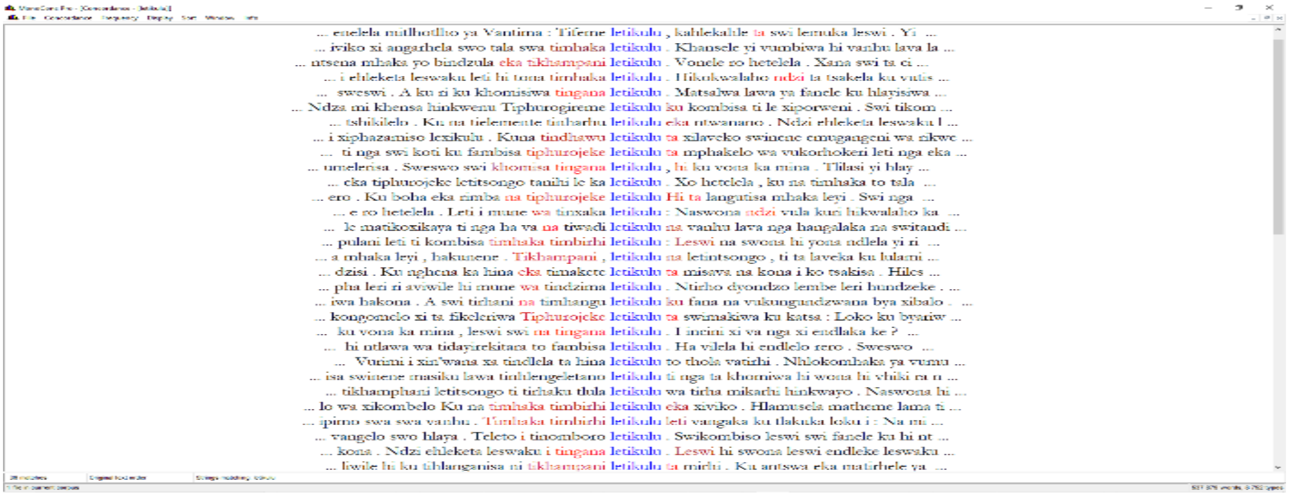
Results of the searched Xitsonga adjectives *lavakulu, letikulu, lavantsongo,* and *letinene* are given in a concordance format in figures. All instances in which the searched adjectives have been used are presented, and these occurrences are utilised to see all words in the corpus that collocate with the searched adjectives. An adjective is a lexical category that modifies a noun by providing additional information about the referent (Radford et al., 2009, p. 130). Adjectives in Xitsonga are constructed by adjectival concords and adjectival stems. The adjectives that are shown in this study have adjectival concords *lava-* and *leti-*, which are derived from demonstrative pronouns and adjectival stems *-kulu* (big), *-ntsongo* (small), and *-nene* (good). As illustrations, the results of the lexical collocations for the searched Xitsonga adjectives *lavakulu, letikulu, lavantsongo,* and *letinene* are presented consecutively in Figures 1, 2, 3, and 4.


MonoConc Pro - [Concordance - [lavakulu]]
File Concordance Frequency Display Sort Window Info
21 matches
Original text order
Strings matching: lavakulu
1 file in current corpus
537 879 words, 8 752 types

```
...  yima nkarhi wo leha . Vatirhela mfumo lavakulu ; Hambiswiritano , hi lava nkarhi nasw ...
...  ya vukhongeri , vavasati na vavanuna , lavakulu na lavatsongo , ririmi , xiyimo xa swa  ...
... ko eka leswi . Kutani swi khumba vanhu lavakulu lava nga eka makete wa swa mitirho . X ...
... iwa kuri ririmi ra vuyimeri . Vavasati lavakulu lava nga na vugono , va tala ku khomiwa ...
... u kambela swi sungula ku tirha . Vanhu lavakulu va nga ha kumeka va ri voxe na ku chava ...
... inene ku swi anamisa . Lava a hi vanhu lavakulu . Mina hi vumina hikwalaho ndzi ta vho ...
...  mfanelo yo vhota eka swa tipolitiki hi lavakulu a ku tava ku fanerile swinene . I nkar ...
... i hi nkarhi . Ndzi seketela dyondzo ya lavakulu . Emakumu ka swona , hi fikelerile nts ...
... rhi va masipala na / kumbe vakhanselara lavakulu va tala ku va kona eka vuleteri leswaku ...
...  vaviseka ku vikiweke . Eka Dyondzo ya Lavakulu na Nhluvukiso . Phuresidente , hi ku y ...
...  humesiwa . Nhlengeletano ya vaofisiri lavakulu wu landzerile. Eka mhaka yaleyo ngopfu ...
... o . Nongonoko wa ku famba ka vaofisiri lavakulu wu ta va hi ndlela leyi : Sweswo hi le ...
... bindzu i xitshunxo . Leyi i dyondzo ya lavakulu . Sweswo a swi fanelanga ku va mbuyelo ...
... lawu . Ku fanele ku va na vuleteri bya lavakulu mikarhi hinkwayo . Ntirhisano lowu i w ...
... a swona , hi hina vatirhisani va bindzu lavakulu va vona . Hi fanele ku fambisa endlelo ...
... a sweswo , va fanele va va vadyandzhaka lavakulu va xona . Ku lebula , hikokwalaho , sw ...
... ele ku va eka vanhu hi voxe . Vatlangi lavakulu a va vanga na chansi . Hi fanele ku ts ...
...  maehleketelelo ya hina . Vuleteri bya lavakulu byi tiyisisiwile . Swihlovo swimbirhi  ...
... oleswo . I tiphesente tingani ta vanhu lavakulu lava nga tirhiki ? Namuntlha , vachaye ...
... Swikombo swa kahle swi : Hi vadyondzi lavakulu u fanele ku tiyisela . Ku xaviwa ka ma ...
...  ya ntirhisano . Xin'wana i dyondzo ya lavakulu . Hi tsakile swinene eka hinkwaswo .  ...
```

Figure 1: Results of the search adjective 'lavakulu'

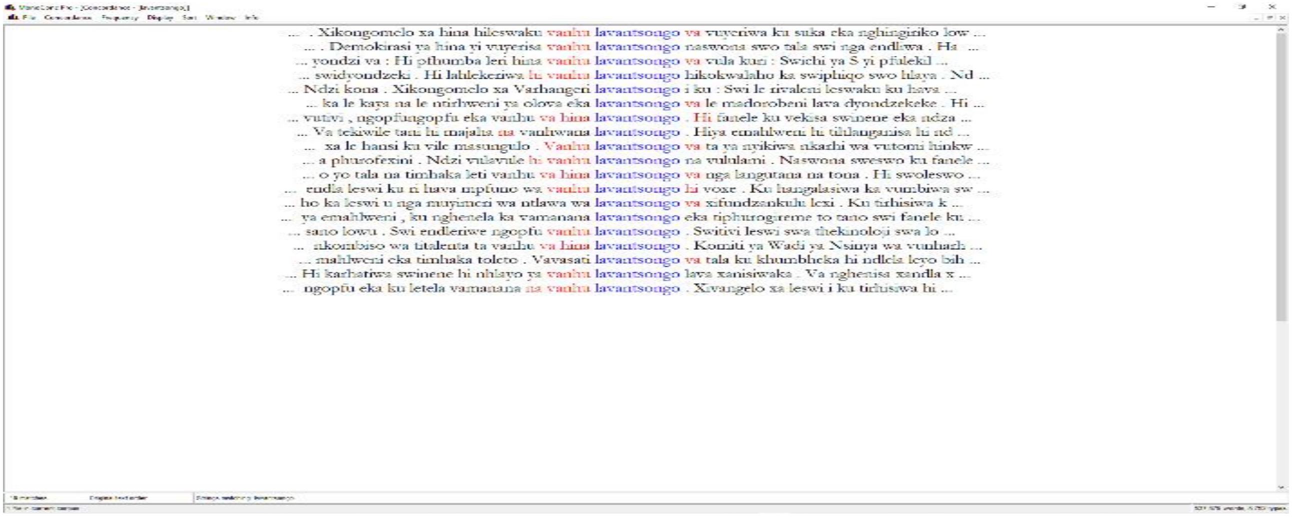
The adjective *lavakulu* has collocated with the nouns *vanhu* (people) and *vaofisiri* (officers) in Figure 1. These nouns belong to class 2 which is known as the human class that uses the plural prefix *va-*. The queried adjective also collocates with subject nouns *vuleteri* (guidance) and *dyondzo* (lesson) and their concords *ya* and *bya*. The software was able to display all collocates of the searched adjective. However, some were not highlighted.



```
... enelela nitlhotlho ya Vantima : Tiferne letikulu , kahlekahle ta swi lemuka leswi . Yi  ...
... iviko xi angarhela swo tala swa timhaka letikulu . Khansele yi vumbiwa hi vanhu lava la ...
... ntsena mhaka yo bindzula eka tikhampani letikulu . Vonele ro hetelela . Xana swi ta ci ...
... i ehleketa leswaku leti hi tona timhaka letikulu . Hikokwalaho ndzi ta tsakela ku vutis ...
...  sweswi . A ku ri ku khomisiwa tingana letikulu . Matsalwa lawa ya fanele ku hlayisiwa ...
... Ndza mi khensa hinkwenu Tiphurogireme letikulu ku kombisa ti le xiporweni . Swi tikom ...
... tshikilelo . Ku na tielemente tinhanhu letikulu eka ntwanano . Ndzi ehleketa leswaku l ...
... i xiphazamiso lexikulu . Kuna tindhawu letikulu ta xilaveko swinene emugangeni wa nkwe ...
...  ti nga swi koti ku fambisa tiphurojeke letikulu ta mphakelo wa vukorhokeri leti nga eka ...
... umelerisa . Sweswo swi khomisa tingana letikulu , hi ku vona ka mina . Tlilasi yi hlay ...
... eka tiphurojeke letitsongo tanihi le ka letikulu . Xo hetelela , ku na timhaka to tala  ...
... ero . Ku boha eka rimba na tiphurojeke letikulu Hi ta langutisa mhaka leyi . Swi nga  ...
... e ro hetelela . Leti i mune wa tinxaka letikulu : Naswona ndzi vula kuri hikwalaho ka  ...
...  le matikoxikaya ti nga ha va na tiwadi letikulu na vanhu lava nga hangalaka na switandi ...
... pulani leti ti kombisa timhaka timbirhi letikulu : Leswi na swona hi yona ndlela yi ri  ...
... a mhaka leyi , hakunene . Tikhampani , letikulu na letintsongo , ti ta laveka ku lulama ...
... dzisi . Ku nghena ka hina eka timakete letikulu ta misava na kona i ko tsakisa . Hiles ...
... pha leri ri aviwile hi mune wa tindzima letikulu . Ntirho dyondzo lembe leri hundzeke . ...
... iwa hakona . A swi tirhani na timhangu letikulu ku fana na vukungundzwana bya xibalo .  ...
... kongomelo xi ta fikeleriwa Tiphurojeke letikulu ta swimakiwa ku katsa : Loko ku byariw ...
...  ku vona ka mina , leswi swi na tingana letikulu . I incini xi va nga xi endlaka ke ?  ...
...  hi ntlawa wa tidayirekitara to fambisa letikulu . Ha vilela hi endlelo rero . Sweswo  ...
...  Vurimi i xin'wana xa tindlela ta hina letikulu to thola vatirhi . Nhlokomhaka ya vumu ...
... isa swinene masiku lawa tinhlengeletano letikulu ti nga ta khomiwa hi wona hi vhiki ra n ...
... tikhamphani letitsongo ti tirhaka tlula letikulu wa tirha mikarhi hinkwayo . Naswona hi ...
... lo wa xikombelo Ku na timhaka timbirhi letikulu eka xiviko . Hlamusela matheme lama ti ...
... ipimo swa swa vanhu . Timhaka timbirhi letikulu leti vangaka ku tlakuka loku i : Na mi ...
... vangelo swo hlaya . Teleto i tinomboro letikulu . Swikombiso leswi swi fanele ku hi nt ...
... kona . Ndzi ehleketa leswaku i tingana letikulu . Leswi hi swona leswi endleke leswaku ...
... liwile hi ku tihlanganisa ni tikhampani letikulu ta mirhi . Ku antswa eka matirhele ya  ...
```

Figure 2: Results of the search adjective 'letikulu'

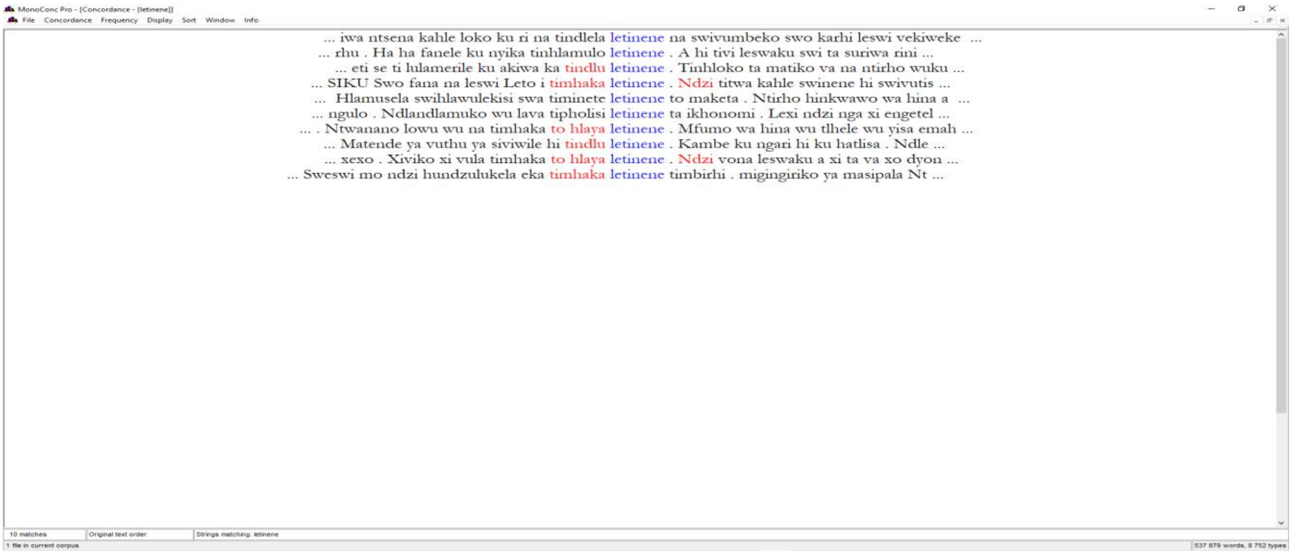
In Figure 2, the searched adjective *letikulu* also collocates with different nouns such as *tikhaphani* (companies), *tindhawu* (places), *tiphurhojeke* (projects), *timhaka* (news), *and tingana* (disgrace) that are categorised in noun class 10. This class uses the plural prefix *ti-*. The figure also shows that the adjective *letikulu* can collocate with the adjective *timbirhi* (two) that gives numerical information about the nouns *timhaka* (news). It is interesting to observe that the software accurately identified *timhaka* (news) and *timbirhi* (two) as collocates of *letikulu*.



```
... . Xikongomelo xa hina hileswaku vanhu lavantsongo va vuyeriwa ku suka eka nghingiriko low ...
... . Demokirasi ya hina yi vuyerisa vanhu lavantsongo naswona swo tala swi nga endliwa . Ha ...
... yondzi va : Hi pfhumba leri hina vanhu lavantsongo va vula kuri : Swichi ya S yi pfulekil ...
... swidyondzeki . Hi lahlekeriwa hi vanhu lavantsongo hikokwalaho ka swiphiqo swo hlaya . Nd ...
... Ndzi kona . Xikongomelo xa Varhangeri lavantsongo i ku : Swi le rivaleni leswaku ku hava ...
... ka le kaya na le ntirhweni ya olova eka lavantsongo va le madorobeni lava dyondzekeke . Hi ...
... vutivi , ngopfungopfu eka vanhu va hina lavantsongo . Hi fanele ku vekisa swinene eka ndza ...
... Va tekiwile tani hi majaha na vanhwana lavantsongo . Hiya emahlweni hi tihlanganisa hi nd ...
...  xa le hansi ku vile masungulo . Vanhu lavantsongo va ta ya nyikiwa nkarhi wa vutomi hinkw ...
... a phurofexini . Ndzi vulavule hi vanhu lavantsongo na vululami . Naswona sweswo ku fanele ...
... o yo tala na timhaka leti vanhu va hina lavantsongo va nga langutana na tona . Hi swoleswo ...
...  endla leswi ku ri hava mpfuno wa vanhu lavantsongo hi voxe . Ku hangalasiwa ka vumbiwa sw ...
... ho ka leswi u nga muyimeri wa ntlawa wa lavantsongo va xifundzankulu lexi . Ku tirhisiwa k ...
...  ya emahlweni , ku nghenela ka vamanana lavantsongo eka tiphurogireme to tano swi fanele ku ...
... sano lowu . Swi endleriwe ngopfu vanhu lavantsongo . Switivi leswi swa thekinoloji swa lo ...
...  nkombiso wa titalenta ta vanhu va hina lavantsongo . Komiti ya Wadi ya Nsinya wa vunhazh ...
... mahlweni eka timhaka toleto . Vavasati lavantsongo va tala ku khumbheka hi ndlela leyo bih ...
... Hi karhatiwa swinene hi nhlayo ya vanhu lavantsongo lava xanisiwaka . Va nghenisa xandla x ...
...  ngopfu eka ku letela vamanana na vanhu lavantsongo . Xivangelo xa leswi i ku tirhisiwa hi ...
```

***Figure 3**: Results of the search adjective 'lavantsongo'*

In Figure 3, the words that collocate with the searched adjective *lavantsongo* are similar to those in Figure 1. The noun *vanhu* (people) is the main word that has collocated with the adjective *lavantsongo*. Besides collocating with the noun *vanhu* (people), the adjective *lavantsonga* also collocates with possessive qualificative *va hina* (of ours). This possessive qualificative was formed by combining the possessive concord *va* (of) and possessive stem *hina* (ours) which is derived from the absolute pronoun. From this figure, we observed that almost all the lexical collocations that involve the searched adjective were highlighted.



```
... iwa ntsena kahle loko ku ri na tindlela letinene na swivumbeko swo karhi leswi vekiweke ...
... rhu . Ha ha fanele ku nyika tinhlamulo letinene . A hi tivi leswaku swi ta suriwa rini ...
... eti se ti lulamerile ku akiwa ka tindlu letinene . Tinhloko ta matiko va na ntirho wuku ...
... SIKU Swo fana na leswi Leto i timhaka letinene . Ndzi titwa kahle swinene hi swivutis ...
...  Hlamusela swihlawulekisi swa timinete letinene to maketa . Ntirho hinkwawo wa hina a ...
... ngulo . Ndlandlamuko wu lava tipholisi letinene ta ikhonomi . Lexi ndzi nga xi engetel ...
... . Ntwanano lowu wu na timhaka to hlaya letinene . Mfumo wa hina wu tlhele wu yisa emah ...
... Matende ya vuthu ya siviwile hi tindlu letinene . Kambe ku ngari hi ku hatlisa . Ndle ...
... xexo . Xiviko xi vula timhaka to hlaya letinene . Ndzi vona leswaku a xi ta va xo dyon ...
... Sweswi mo ndzi hundzulukela eka timhaka letinene timbirhi . migingiriko ya masipala Nt ...
```

10 matches | Original text order | Strings matching: letinene
1 file in current corpus | 537 879 words, 8 752 types

***Figure 4:** Results of the search adjective 'letinene'*

From Figure 4, the occurrences show that all words which collocate with the adjective *letinene* are the same as in Figure 2. The collocates *tindlu* (houses) and *timhaka* (news) are nouns that are classified under class 10 in the Xitsonga noun class system. The remaining collocate *to hlaya* (many) that is highlighted in the figure is a relative qualificative. It was formed by a relative concord *to* and relative stem *hlaya* (count) which is derived from a verb.

5 Conclusion

In the study, we used MonoConc Pro to explore and visualise collocates of adjectives from the Xitsonga monolingual corpus. We deduce that lexical collocations or words that co-occur with adjectives, namely *lavakulu, letikulu, lavantsongo* and *letinene*, are nouns, adjectives, possessives, and relatives. It was discovered that each adjective often collocates with particular nouns that belong to a specific noun class system. The adjectives that were studied collocate with nouns in classes 2 and 10 of the Xitsonga nouns class system. The utilisation of the MonoConc Pro software has helped us to realise that adjectives do not only co-occur with nouns but also with other parts of speech. The overall results have demonstrated the usefulness of concordancers such as MonoConc Pro to analyse Xitsonga language patterns such as lexical collocations. The study also exhibits the significance of concordancers and corpora in increasing language proficiency and collocational competence awareness among speakers of the language in question. The use of such language technology provides a practical approach to teaching and learning lexical collocations with the aid of corpora, other than relying on traditional methods of teaching and learning linguistic traits of indigenous languages.

References

Bahns, J 1993, 'Lexical collocations: a contrastive view', *ELT Journal,* vol. 47, no 1, pp. 56-63.

Benson, M, Benson, E, & Ilson, R 1997, The BBI dictionary of English word combinations, John Benjamins Publishing Company, Amsterdam.

Finlayson, R & Madiba M 2002, 'The intellectualisation of the indigenous languages of South Africa: Challenges and prospects', *Current Issues in Language Planning,* vol. 3, no 1, pp. 40-61.

Groenewald, HJ & Fourie, W 2009, 'Introducing the Autshumato Integrated Translation Environment' *13th Annual Conference, EAMT Proceedings 2009*, Barcelona, Spain, May 14-15, 2009, pp. 190-196.

Jafarpour, AA, Hashemian, M & Alipour, S 2013, 'A corpus-based Approach toward teaching collocation of synonyms', *Theory and Practice in Language Studies*, vol. 3, no 1, pp. 51-60.

Johns, T 1991, 'Should you be persuaded: Two examples of data-driven learning', in T. Johns & P. King (Eds.), Classroom Concordancing. *English Language Research Journal,* vol. 4, pp. 1-16.

Lewis, M 2000, 'Language in the lexical approach', in M Lewis, *Teaching collocation: Further developments in the lexical approach,* Language Teaching Publications, Hove, pp. 126-54.

Madiba, M 2004, 'Parallel corpora as tools for developing the indigenous languages of South Africa with special reference to Venda', *Language Matters,* vol. 35, no 1, pp. 133-147.

Mlambo, R, Skosana, N & Matfunjwa, M 2021, 'The extraction of terminology list using ParaConc for creating a Quadrilingual dictionary', *Southern African Linguistics and Applied Language Studies*, vol. 39, no 1, pp. 82-91.

Moehkardi, R 2002, 'Grammatical and lexical English collocations: Some possible problems to Indonesian learners of English', *Humaniora,* vol. 14, no 1, pp. 53-62.

Moropa, K 2007, 'Analysing the English-Xhosa parallel corpus of technical texts with ParaConc: A case study of term formation processes', *Southern African Linguistics and Applied Language Studies,* vol. 25, no 2, pp. 183-205.

Ndhlovu, K 2016, 'Using ParaConc to extract bilingual terminology from parallel corpora: A case of English and Ndebele', *Journal of Literary Criticism, Comparative Linguistics and Literary Studies,* vol. 37, no 2, pp. 1-12.

Pirmoradian, M & Tabatabaei, O 2012, 'The enhancement of lexical collocation learning through concordancing: A case of Iranian EFL learners', *The Modern Journal of Applied Linguistics*, vol. 4, no 4, pp. 185-200.

Radford, A, Martin, A, David, Herald, C, Andrew, S 2009. Linguistics: An Introduction, Cambridge University Press, Cambridge.

Reppen, R 2001, 'Review of MonoConc Pro and Wordsmith tools', *Language Learning and Technology*, vol. 5, no 3, pp. 32-36.

Shoba, FZ 2018, 'Exploring the use of parallel corpora in the compilation of specialized bilingual dictionaries of technical terms: A case study of English and isiXhosa', PhD thesis, University of South Africa.

Statistics South Africa 2012, *Census 2011: Census in brief*, Statistics South Africa, Pretoria.

Taljard, E & De Schryver GM 2002, 'Semi-automatic term extraction for the African languages, with special reference to Northern Sotho', *Lexikos,* vol. 12, pp. 44-74

Training Cross-Lingual embeddings for Setswana and Sepedi

Makgatho, Mack
Dept. of Computer Science, University of Pretoria
mack.letladi1@gmail.com

Marivate, Vukosi
Dept. of Computer Science, University of Pretoria
vukosi.marivate@cs.up.ac.za

Sefara, Tshephisho
Council for Scientific and Industrial Research
tsefara@csir.co.za

Wagner, Valencia
Sol Plaatje University
valencia.wagner@spu.ac.za

Abstract

African languages still lag in the advances of Natural Language Processing techniques, one reason being the lack of representative data, having a technique that can transfer information between languages can help mitigate against the lack of data problem. This paper trains Setswana and Sepedi monolingual word vectors and uses VecMap to create cross-lingual embeddings for Setswana-Sepedi in order to do a cross-lingual transfer.

Word embeddings are word vectors that represent words as continuous floating numbers where semantically similar words are mapped to nearby points in n-dimensional space. The idea of word embeddings is based on the distribution hypothesis that states, semantically similar words are distributed in similar contexts (Harris, 1954).

Cross-lingual embeddings leverages monolingual embeddings by learning a shared vector space for two separately trained monolingual vectors such that words with similar meaning are represented by similar vectors. In this paper, we investigate cross-lingual embeddings for Setswana-Sepedi monolingual word vector. We use the unsupervised cross lingual embeddings in VecMap to train the Setswana-Sepedi cross-language word embeddings. We evaluate the quality of the Setswana-Sepedi cross-lingual word representation using a semantic evaluation task. For the semantic similarity task, we translated the WordSim and SimLex tasks into Setswana and Sepedi. We release this dataset as part of this work for other researchers. We evaluate the intrinsic quality of the embeddings to determine if there is improvement in the semantic representation of the word embeddings.

Keywords: cross-lingual embeddings, word embeddings, intrinsic evaluation

1 Introduction

Many African languages have insufficient language resources (data, tools, people) (Abbott & Martinus 2019, Martinus & Abbott 2019, Nekoto et al. 2020, Sefara et al. 2021) and fall into the classification of low resource languages (Ranathunga et al. 2021) in the Natural Language Processing (NLP) field. This lack of resources makes it harder to capitalise on the recent advances in many NLP downstream tasks such as Neural Machine Translation (Cho et al. 2014), Large Language Models (Devlin et al. 2018, Howard & Ruder 2018), Q&A systems (Kwiatkowski et al. 2019), etc. There may be more downstream approaches to deal with some of these challenges such as Transfer Learning (Ruder et al. 2019), Data Augmentation (Marivate & Sefara 2020*a*), Multilingual Models (Hedderich et al. 2020), etc. Additionally, the lack of research attention to existing NLP techniques results in difficulties finding a benchmark (Abbott & Martinus 2019). In this work, we focus on word representations through word embeddings and how we can leverage one language to assist in the representation of another related language. These embeddings can then be used to develop tools for other downstream tasks.

Word Embeddings are a mathematical technique to learn general language vector representations from a large amount of unlabelled text using co-occurring statistics. In recent years, monolingual word embeddings techniques are increasingly becoming an important resource in NLP. Word embeddings are widely used in NLP problems such as sentiment analysis (Socher et al. 2013), named-entity-recognition (Guo et al. 2014), parts-of-speech tagging, and document retrieval. Word2Vec is a vector training model proposed by Mikolov et al. (2013). Word2Vec produces a low-dimensional real-value vector representing the meaning of a word. The word vector represents grammatical and semantic properties, which results in words with similar semantic relations being close to each other. The word vector representation method incorporates the semantic relationship between words which is not possible through representations such as Bag-Of-Words of TFIDF. Word embeddings are better than both methods because they map all the words in a language into a vector space of a given dimension, the words are converted into vectors and allow multiple linear operations and have the property of preserving analogies (Mikolov et al. 2013, Pennington et al. 2014).

Cross-lingual word embeddings have been receiving more and more attention from the NLP community, mainly because it has provided a path to effectively align two disjoint monolingual embeddings with no bilingual dictionary for unsupervised techniques or no more than a small bilingual dictionary for supervised techniques (Lample et al. 2018, Artetxe et al. 2020). Cross-lingual techniques also enable knowledge transfer between languages with rich resources and low resources. For languages lacking bilingual parallel corpus with other languages, cross-lingual embeddings can be utilised to train high-quality cross-lingual embeddings (Lample et al. 2018). This can aid in accelerating the progress of applying NLP to low-resourced languages. Artetxe et al. (2018) created the cross-lingual unsupervised or supervised word embedding (VecMap library) approach for training cross-lingual word embedding models. The approaches can be used to construct cross-language word vectors with or without a bilingual dictionary.

The majority of South African languages lag bilingual parallel corpus with other languages. In this work, we aim to investigate how cross-lingual embeddings could be used to improve the state of one or both languages. We used data (corpus) from different domains to train Word2Vec and fastText (Bojanowski et al. 2016) monolingual embeddings. When using VecMap, the two embeddings are aligned. VecMap requires two monolingual word vectors from source and target (Artetxe et al. 2018). To evaluate the effectiveness of the cross-lingual embedding for Setswana and Sepedi, we use intrinsic evaluation (Bakarov 2018) through Setswana and Sepedi versions of WordSim (Finkelstein et al. 2001) and Simlex (Hill et al. 2015). This is following on an approach that has been used for Yoruba and Twi (Alabi et al. 2019). We also release the dataset for this benchmark of human semantic similarity task.

This paper is structured as follows; the next section is a review of related work that is done on cross-lingual word vectors. Followed by data collection in Section 3. Section 4 discusses methodology followed to train cross-lingual word vectors using VecMap. The evaluation of the word vectors is discussed in Section 5. Section 5.1 explains the results while Section 6 discusses the findings and finally, conclusions and future work can be found in Section 7.

2 Background and Related Work

Cross-lingual word embeddings (CLWEs) are becoming popular in NLP for two reasons: Cross-lingual word embeddings can transfer knowledge from rich-resourced languages to low-resourced; The technique can also infer the semantics of words in a multiple language environment. Conneau et al. (2018) show that word embeddings spaces can be aligned without any cross-lingual supervision. The alignment is based on solely unaligned datasets of each language. Using adversarial training, they were able to initialise a linear mapping between a source and a target space, which they use to create a synthetic parallel dictionary. First, they propose a simple criterion that is used as an unsupervised validation matric. Second, they propose the similarity measure cross-domain similarity local scaling (CSLS), which mitigates the hubness problem and increases the word translation accuracy. The hubness problem is defined by Dinu et al. (2015) as:

> ”neighbourhoods of the mapped elements are strongly polluted by hubs, vectors that tend to be near a high proportion of items, pushing their correct labels down the neighbour list.”

In the work done by Adams et al. (2017), the research looked at applying CLWEs to Yongning Na, a Sino-Tibetan language. The research focused on determining if the quality of CLWEs depends on having large amounts of data in multiple languages and if initialising the parameters of neural network language models (NMLM) can improve language modelling in a low-resourced context. The research scaled down the available monolingual data of the target language to about 1000 sentences. The quality of intrinsic embedding was assessed by taking into consideration correlation with human judgement on the WordSim353 (Finkelstein et al. 2001) test set. They went further to perform language modelling experiments by initialising the parameters for long short-term memory (LSTM) (Hochreiter & Schmidhuber 1997) by training across different language pairs. The research showed that CLWEs are resilient even when target language training data is scaled-down and that initialisation of NMLM parameters leads to good performance.

Artetxe & Schwenk (2019) introduced an architecture that can be used to learn multilingual sentence representations for more than 90 languages. The languages belonged to 30 different families. The research used a single BiLSTM encoder with a shared Byte Pair Encoding (BPE) vocabulary coupled with an auxiliary decoder and trained on parallel corpora. They learn a classifier using English annotated data only and transfer it to any language without modification. The research mainly focused on vector representations of sentences that are general for the input language and the NLP task.

Alabi et al. (2019) worked on massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. Authors compare two types of word embeddings obtained from curated corpora and a language-dependent processing. They move further to collect high quality and noisy data for the two languages. They quantify that improvements that is based on the quality of data and not only on the amount of data. In their experiments, they use different architectures to learn word representations both from characters and surface forms. They evaluate multilingual BERT on a down stream task, specifically named entity recognition and WordSim-353 word pairs dataset.

Feng et al. (2018) investigates a cross-lingual knowledge transfer technique to improve the semantic representation of low-resourced languages and improving low resource named-entity recognition. In their research, neural networks are used to do knowledge transfer from high resource language using bilingual lexicons to improve low resource word representation. They automatically learn semantic projections using a lexicon extension strategy that is designed to address out-of lexicon problem. Finally, they regard word-level entity type distribution features as an external language independent knowledge and incorporate them into their neural architecture. The experiment is done on two low resource languages (Dutch and Spanish) to demonstrate the effectiveness of these additional semantic representations.

Banerjee et al. (2021) show that initialising the embedding layer of Unsupervised Neural Machine Translation (UNMT) models with cross-lingual embeddings shows significant improvements in BLEU score. Authors show that freezing the embedding layer weights lead to better gains compared to updating the embedding layer weights during training. They experimented using Denoising Autoencoder (DAE) and Masked Sequence to Sequence (MASS) for three different unrelated language pairs (for English-Hindi, English-Bengali, and English-Gujarati). The analysis shows the importance of using cross-lingual embedding as compared to other techniques.

The literature shows that there is a substantial amount of work done on cross-lingual transfer and empirical proof that the method improves the performance of models. The literature does not relay solely on intrinsic evaluation but the solutions are applied to some downstream tasks. In the next section, we detail the data used for conducting experiments.

3 Data collection

Training data is very important for implementing powerful and accurate models, and clean training data can make a difference between a good and great model. The data needs to be very imperative because the quality of the alignment depends on the quality of the monolingual embeddings, i.e. data used to create the initial monolingual embeddings before mapping.

We use data collected from different domains for training word vectors:

- **JW300 bible** (Agić & Vulić 2019): A biblical-domain data set containing parallel corpus for low-resourced languages.
- **Wikipedia**
- **National Centre for Human Language Technology (NCHLT) text corpus** (Eiselen & Puttkammer 2014): The dataset contains clean textual data in Sepedi and Setswana. The data set was constructed by harvesting existing data such as online publications, online news, web crawling and crowd-sourcing.
- **SABC News Data in Setswana and Sepedi** (Marivate et al. 2020, Marivate & Sefara 2020*b*): The data set contains news titles collected from online social media.

National Centre for Human Language Technology (NCHLT) data is used for training monolingual word vectors. For preprocessing, we changed all words to lowercase, removing brackets, digits, punctuations, and white spaces.

Table 1: Corpus size for the Setswana and Sepedi Datasets

	Sepedi	Setswana
Number of tokens:	2133972	3000682
Unique words:	93461	107606

In this section we dealt with how we collected the data used to training our monolingual embeddings for both languages and what approach we took to pre-process the data before training the models. In the next section we discuss the approach taken to train the monolingual embeddings and how VecMap was used to training the cross-lingual embeddings.

4 Training monolingual and cross-lingual embeddings (VecMap)

In this section, we present the methods (frameworks) used to train monolingual and cross-lingual embeddings. We describe the parameters used to train word2Vec and fastText embeddings. We also look into VecMap, the framework that we used to align monolingual embeddings.

CLWEs have proved to perform very well for low-resourced languages. The main idea is to do a cross-lingual transfer from the source language to the target, such that we have a single representation for a pair of languages where semantically similar words are closer to one another. In order to use VecMap two monolingual embeddings are required, we train fastText and word2Vec vectors. We use the following parameters for fastText and word2Vec in Table 2. The definition of the parameters are as follows: skipGram - training method, dim - size of word vectors, minCount - minimal number of word occurrences, ws - size of the context window, and epoch - number of epochs or iterations.

4.1 Word2Vec

The word2Vec (Mikolov et al. 2013) algorithm is a two-layer neural network that vectorises words to

Table 2: Parameters for FastText and Word2Vec

Parameter	Value
skip-gram	true
dim	300
minCount	1
ws	4
epoch	100

processes text. The algorithm takes as input a text corpus and returns feature vectors that represent words in that corpus as a set of vectors. Word2Vec trains words against neighbouring words based on a window size context. It trains the words using two methods: skip-gram or continuous bag of words (CBOW), skip-gram uses a word to predict a target context and CBOW uses context to predict a target word. The experiment uses skip-gram to train monolingual embeddings. We use word vectors that were trained using Word2Vec. These correspond to monolingual embeddings of dimension 300 trained on Sepedi and Setswana corpora.

4.2 FastText

FastText (Bojanowski et al. 2016) is a supervised prediction-based technique based on the word2Vec family of algorithms (Mikolov et al. 2013). It predicts tags through context and represents each word as an *n*-gram of characters, instead of learning vectors for words directly. The fastText model has three layers: input layer, hidden layer, and output layer. Input is a number of words and their *n*-gram features, these features are used to represent a single document. The hidden layer is the superimposed average of multiple feature vectors. The hidden layer solves the maximum likelihood function, then constructs a Huffman tree according to the weights and model parameters of each category, and uses the Huffman tree as the output.

4.3 VecMap

VecMap (Artetxe et al. 2020) is an open-source framework to learn CLWEs written in Python. There are two techniques to do cross-lingual embeddings with VecMap, supervised (recommended if you have a large training dictionary) and unsupervised (recommended if you have no seed dictionary and do not want to rely on identical words). In this work, we align word embedding using VecMap[1]. The approach is fully unsupervised. The steps we followed to build our cross-lingual word embeddings model are motivated by the authors of VecMap Artetxe et al. (2020). The assumption is that we have a monolingual corpus for source and target languages. The word representations is learned independently for each language (monolingual embeddings for each language), and then mapped to a common vector space.

In this section, we presented word2Vec, fastText and VecMap. We also described the parameters used to train word2Vec and fastText embeddings. In the next section, we present experimental results and perform some analyses.

5 Evaluation

We evaluate the quality of Setswana and Sepedi word vector representations on two different benchmarks Simlex and WordSim. The datasets (Simlex and WordSim) contain pairs of Setswana and Sepedi words that have been assigned similarity ratings by humans. They give a similarity score between a pair of words corresponding to their relatedness. Cosine similarity is used to collect a score from the model in order to check how close the score is to the human score, we use Spearman to measure correlation. Spearman index measure the dependence of two variables, the correlation of two statistical variables is evaluated using monotonic equation. We manually translate the WordSim and Simlex word pairs dataset from English into Setswana and Sepedi. We are releasing a dataset of Setswana and Sepedi translated WordSim and Simlex as part of this project at `https://github.com/dsfsi/embedding-eval-data` and archived on Zenodo at `https://zenodo.org/record/5673974`.

Table 3: FastText Monolingual Results

Monolingual fastText	Coverage	Spearman
Sepedi(Simlex)	94.58	40.39
Sepedi(WordSim)	81.29	46.15
Setswana(Simlex)	95.22	33.23
Setswana(WordSim)	95.38	44.80

Table 4: Word2Vec Monolingual Results

Monolingual word2Vec	Coverage	Spearman
Sepedi(Simlex)	79.49	25.96
Sepedi(WordSim)	84.49	23.57
Setswana(Simlex)	95.32	31.52
Setswana(WordSim)	95.38	35.11

5.1 Results

This section presents the results of the experiments conducted to show the efficiency of the proposed technique with a couple of experiments. We first present the monolingual evaluation task for word2Vec and fastText and then present the cross-lingual evaluation task for Setswana and Sepedi. The evaluations of cross-lingual evaluation task is based on two embedding methods fastText and word2Vec.

In Table 3 and Table 4, we show the Spearman's correlation for word vectors trained on fastText and word2vec. The correlation scores calculate the similarity between word vectors. Table 5 and Table 6 scores are obtained from using Setswana and Sepedi monolingual vectors and using VecMap to align the two vectors to the same vector space.

The results at Table 3 and Table 4 show the coverage and Spearman results. Coverage refers to the total number of in vocabulary words (words that are found both in the model and evaluation dataset). We can see that the coverage is lower for word2Vec but a little higher for fastText (we expected coverage for fastText to be 100 percent). The Simlex and WordSim similarity score for monolingual fastText embeddings in Table 3 is higher, this is expected due to the coverage percentage also being very high as compared to the coverage value in Table 4.

Table 5: Word2Vec Crosslingual Results

Monolingual word2Vec	Coverage	Spearman
Setswana-Sepedi(Simlex)	90.76	31.14
Setswana-Sepedi(WordSim)	68.56	40.87

Table 6: FastText Crosslingual Results

Crosslingual fastText	Coverage	Spearman
Setswana-Sepedi(Simlex)	91.19	30.44
Setswana-Sepedi(WordSim)	68.84	36.33

6 Discussion

The main purpose of this research is to show that it is possible to do cross-lingual transfer from the source language to the target. In essence we wanted to check if cross-lingual alignment can improve the word representation for the target language. The results on Table 4 shows that the Spearman's correlation value for the target language when using word2Vec is low, this is also due to coverage percentage, but fastText based-embeddings perform better on Table 3 and has a higher coverage percentage, as stated upove we expected 100 percent coverage. Table 5 shows that we improved the representation of words after cross-lingual alignment for word2Vec based-embeddings. The Spearman's value has increased for both Simlex and Wordsim. We expected to improve the results for fastText embeddings but in this case word2Vec actually yielded better results.

7 Conclusion

In this paper, VecMap was used to align Setswana-Sepedi to the same vector space. Through this work, we wanted to use cross-lingual (VecMap) technique to enable knowledge transfer between languages with rich resources and low resources. The results show that it is possible to align two monolingual embeddings to get cross-lingual embeddings. We mapped Setswana to Sepedi and used Spearman's to check correlation. Interestingly we get different results on fastText and word2Vec-based embeddings though we used the same data to train the embeddings.

In future work, it would be interesting to use the cross-lingual embedding on a downstream task like translation or sentiment analysis specifically for low-resourced languages.

8 Acknowledgements

We would like to acknowledge ABSA for sponsoring the industry chair and it's related activities to the project.

References

Abbott, J. & Martinus, L. (2019), Benchmarking neural machine translation for southern african languages, *in* 'Proceedings of the 2019 Workshop on Widening NLP', pp. 98–101.

Adams, O., Makarucha, A., Neubig, G., Bird, S. & Cohn, T. (2017), Cross-lingual word embeddings for low-resource language modeling, *in* 'Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers', pp. 937–947.

Agić, Ž. & Vulić, I. (2019), JW300: A wide-coverage parallel corpus for low-resource languages, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 3204–3210.
URL: *https://aclanthology.org/P19-1310*

Alabi, J. O., Amponsah-Kaakyire, K., Adelani, D. I. & España-Bonet, C. (2019), 'Massive vs. curated word embeddings for low-resourced languages. the case of yor\ub\'a and twi', *arXiv preprint arXiv:1912.02481* .

Artetxe, M., Labaka, G. & Agirre, E. (2018), A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 789–798.

Artetxe, M., Ruder, S. & Yogatama, D. (2020), On the cross-lingual transferability of monolingual representations, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', pp. 4623–4637.

Artetxe, M. & Schwenk, H. (2019), 'Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond', *Transactions of the Association for Computational Linguistics* **7**, 597–610.

Bakarov, A. (2018), 'A survey of word embeddings evaluation methods', *arXiv preprint arXiv:1801.09536* .

Banerjee, T., au2, R. M. V. & Bhattacharyya, P. (2021), 'Crosslingual embeddings are essential in unmt for distant languages: An english to indoaryan case study'.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016), 'Enriching word vectors with subword information', *arXiv preprint arXiv:1607.04606* .

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014), 'Learning phrase representations using rnn encoder-decoder for statistical machine translation', *arXiv preprint arXiv:1406.1078* .

Conneau, A., Lample, G., Ranzato, M., Denoyer, L. & Jégou, H. (2018), 'Word translation without parallel data'.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018), 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .

Dinu, G., Lazaridou, A. & Baroni, M. (2015), 'Improving zero-shot learning by mitigating the hubness problem'.

Eiselen, R. & Puttkammer, M. J. (2014), Developing text resources for ten south african languages., *in* 'LREC', pp. 3698–3703.

Feng, X., Feng, X., Qin, B., Feng, Z. & Liu, T. (2018), Improving low resource named entity recognition using cross-lingual knowledge transfer, *in* 'Proceedings of the Twenty-Seventh In-

ternational Joint Conference on Artificial Intelligence, IJCAI-18', International Joint Conferences on Artificial Intelligence Organization, pp. 4071–4077.
URL: *https://doi.org/10.24963/ijcai.2018/566*

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppin, E. (2001), Placing search in context: The concept revisited, *in* 'Proceedings of the 10th international conference on World Wide Web', pp. 406–414.

Guo, J., Che, W., Wang, H. & Liu, T. (2014), Revisiting embedding features for simple semi-supervised learning, *in* 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)', pp. 110–120.

Hedderich, M. A., Adelani, D., Zhu, D., Alabi, J., Markus, U. & Klakow, D. (2020), Transfer learning and distant supervision for multilingual transformer models: A study on african languages, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', pp. 2580–2591.

Hill, F., Reichart, R. & Korhonen, A. (2015), 'Simlex-999: Evaluating semantic models with (genuine) similarity estimation', *Computational Linguistics* **41**(4), 665–695.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural Computation* **9**, 1735–1780.

Howard, J. & Ruder, S. (2018), Universal language model fine-tuning for text classification, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 328–339.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K. et al. (2019), 'Natural questions: a benchmark for question answering research', *Transactions of the Association for Computational Linguistics* **7**, 453–466.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L. & Jégou, H. (2018), Word translation without parallel data, *in* 'International Conference on Learning Representations'.

Marivate, V. & Sefara, T. (2020*a*), Improving short text classification through global augmentation methods, *in* 'International Cross-Domain Conference for Machine Learning and Knowledge Extraction', Springer, pp. 385–399.

Marivate, V. & Sefara, T. (2020*b*), 'South african news data'.
URL: *https://doi.org/10.5281/zenodo.3668495*

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R. & Modupe, A. (2020), Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi, *in* 'Proceedings of the first workshop on Resources for African Indigenous Languages', pp. 15–20.

Martinus, L. & Abbott, J. Z. (2019), 'A focus on neural machine translation for african languages', *arXiv preprint arXiv:1906.05685* .

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* 'Advances in neural information processing systems', pp. 3111–3119.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabenamualu, S. K., Osei, S., Sackey, F. et al. (2020), Participatory research for low-resourced machine translation: A case study in african languages, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings', pp. 2144–2160.

Pennington, J., Socher, R. & Manning, C. (2014), Glove: Global vectors for word representation, Vol. 14, pp. 1532–1543.

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M. & Kaur, R. (2021), 'Neural machine translation for low-resource languages: A survey', *arXiv preprint arXiv:2106.15115* .

Ruder, S., Peters, M. E., Swayamdipta, S. & Wolf, T. (2019), Transfer learning in natural language processing, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials', pp. 15–18.

Sefara, T. J., Zwane, S. G., Gama, N., Sibisi, H., Senoamadi, P. N. & Marivate, V. (2021), Transformer-based machine translation for low-resourced languages embedded with language identification, *in* '2021 Conference on Information Communications Technology and Society (ICTAS)', IEEE, pp. 127–132.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y. & Potts, C. (2013), Recursive deep models for semantic compositionality over a sentiment treebank, *in* 'Proceedings of the 2013 conference on empirical methods in natural language processing', pp. 1631–1642.

An Open Source System for Crowd Sourcing an African Language Short Story Corpus

Muite, Benson K.
Kichakato Kizito
benson_muite@emailplus.org

Abstract

Many African languages have few open access corpora for use in developing technological applications such as grammar checkers, spell checkers, speech to text, text to speech and machine translation tools. This may lead to a decline in all cultural traits associated with the peoples that speak these languages. To enable collection of textual corpora, and long term preservation of positive cultural characteristics, the design considerations and implementation of an open source online short story competition collection and evaluation system are described. The system is written in PHP, and can be relatively cheaply deployed on shared hosting servers available from many African hosting providers. This allows for the possibility of a decentralized collection of stories, as well as adaptation, and improvements of the software to different types of short story competitions. The software has been used for two short story competitions across the African continent with the aim of providing stories suitable for children. Holding the competition online has enabled participation from a wide variety of locations, but most of the submissions have came from African countries with relatively good information technology infrastructure. Preparation for a third competition is in progress.

Keywords: Crowd sourcing, African Literature, Corpus Creation

1 Introduction

Africa has a great variety of languages and language dialects. Many of these have rich oral traditions, and have only recently been used for written electronic communication, such as social media, email and internet forums. Many African countries have chosen to use English or French as the language of education, however the internet has democratized access to spaces for public expression, and with this freedom, people have begun to express themselves in their own languages - the Indigenous Tweets project (Scannell 2021) documents the use of languages that are not widespread on one social media platform. Expecting this trend to grow for at least some African languages, there will be interest and need for digital language processing tools for African languages. To develop such tools, language data is required. One way to collect this data is by crowd sourcing. Complementary multilingual short story collection efforts include StoryWeaver in India (Story Weaver 2021) and the African Story Book in South Africa (Saide 2021). Neither of these at present has openly available software for other people to use in crowd sourcing short stories, and they only focus on literacy, without much effort made for use of the collected materials to enable the development of natural language processing tools.

2 Crowd Sourcing Platform Design Considerations

While most people in Africa, uses telephones for communication rather than computers, internet access is growing. Websites enable easier communication for longer pieces of writing such as short stories, and so are a good first platform choice. Further considerations for a web based platform are:

- Many of Africa's languages are closely linked to ethnic identity. This can be a source of tension and has caused conflict. The platform should allow anonymous submissions and/or submission by a proxy.
- Many African countries have a strong focus on English and French as languages of education. Translation from African languages has also primarily focused on translating to English or French, yet many African languages are closely related and may have concepts not well expressed in English and French, so it is important to develop corpora that enable cre-

ation of translation tools between African languages.

- Many authors and translators making a submission may not know English, so the submission platform should be available in several languages, ideally every language submissions can be made in.
- Since many African languages are not used in or studied at school, the language ability of people who submit may not be very high, but they may be interested in improving it. Links to available digital resources should be made available and efforts should be made to allow for feedback on submissions.
- The software should be easy and inexpensive to deploy within Africa to allow for independent organization of short story competitions.
- The software should and be easy for software developers based in Africa to contribute to.
- The software should as much as possible respect privacy of people submitting stories and translations, and enable them to keep personally identifying information private should they choose to do so.

3 Crowd Sourcing Platform Implementation

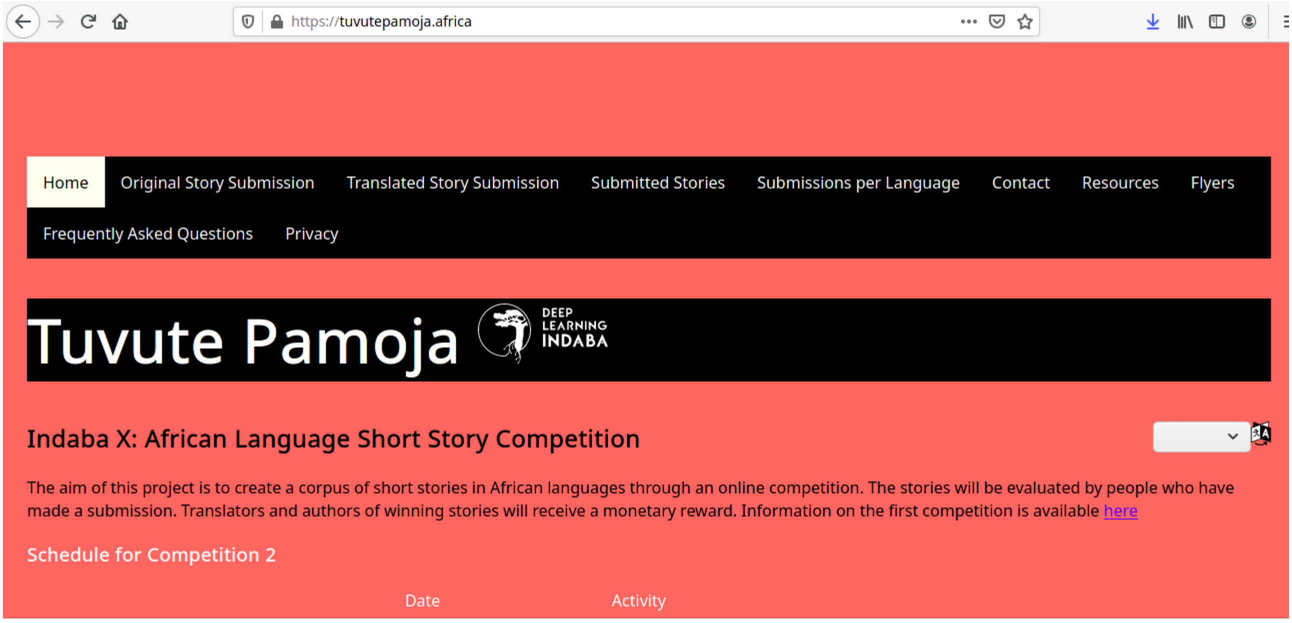
The first consideration is the programming language to use. Python (Python Software Foundation 2021), Ruby (Ruby Community 2021), PHP (PHP Community 2021) and Javascript (ECMA International 2021) were considered. Python and Ruby have great use in web development, however they have poor support for shared hosting in Africa and typically require higher specification servers than PHP - shared hosting support can be improved, but in the time frame of this project, this was not feasible. Javascript is also used for developing the server side of a web application, in addition to making the client side of a web application interactive, however, it again has much less support for shared hosting in Africa. Thus, PHP was chosen as the programming language to use for server programming, with some amount of Javascript, in particular to create soft keyboards. Many cloud providers in Africa use the software cPanel (cPanel LLC 2021) for managed shared hosting with upto 10 GB of storage for prices between $1 and $5 per month. This has support for hosting a PHP web application and a database.

The software consists of:

- A database of original stories along with the contact details of the author or their proxy
- A database of translated stories, with the contact details of the translator or their proxy
- A database of authors, translators or their proxies and the languages in which they can vote
- A database with votes for and comments on each original story or translated story
- A server side PHP program to allow for stories and translations to be submitted
- Client side HTML and Javascript components to enable input of special characters using a soft keyboard and translation of the website, submission of stories, submission of translations, viewing of stories and voting

The website can be viewed at `https://tuvutepamoja.africa`. Figures 1 and 2 show the front page of the website in English and in Kiswahili. Figure 3 shows a soft keyboard suitable for entering Yorùbá which has a number of accented characters which many people who are not trained typists need to see to be able to use in their writing.

The software was openly developed using version control on `https://notabug.org/tuvutepamoja` and is released under the GNU general public license (Free Software Foundation 2021*b*). The repository `https://notabug.org` hosts the development of GnuSocial (GnuSocial Community 2021), an open source social networking site also written in PHP which can be translated into African languages. An attempt was made to use GnuSocial to enable communication between



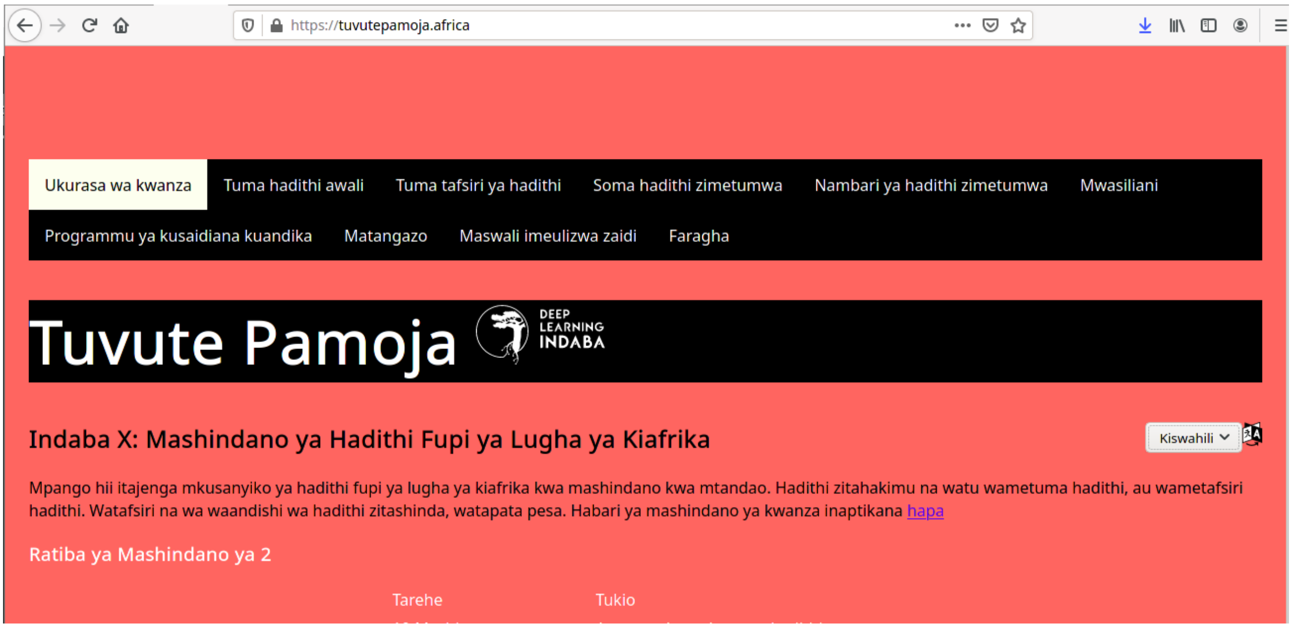
Indaba X: African Language Short Story Competition

The aim of this project is to create a corpus of short stories in African languages through an online competition. The stories will be evaluated by people who have made a submission. Translators and authors of winning stories will receive a monetary reward. Information on the first competition is available here

Schedule for Competition 2

Date	Activity

Figure 1: English front page.



Indaba X: Mashindano ya Hadithi Fupi ya Lugha ya Kiafrika

Mpango hii itajenga mkusanyiko ya hadithi fupi ya lugha ya kiafrika kwa mashindano kwa mtandao. Hadithi zitahakimu na watu wametuma hadithi, au wametafsiri hadithi. Watafsiri na wa waandishi wa hadithi zitashinda, watapata pesa. Habari ya mashindano ya kwanza inaptikana hapa

Ratiba ya Mashindano ya 2

Tarehe	Tukio

Figure 2: Kiswahili front page.

people interested in promoting the competition in African languages that they know, however engagement on this platform was low. At present an online forum using the open source software FluxBB (FluxBB Community 2021), is being used to enable public communication by people interested in the project. The forum is available at `https://ingxoxo.tuvutepamoja.africa/`.

For many possible contributors, translation of software is not as exciting as writing their own short stories, so this has not yet been successful. Translating flyers advertising the short story competition was less of a task and many people contributed to this effort, enabling use of social media platforms

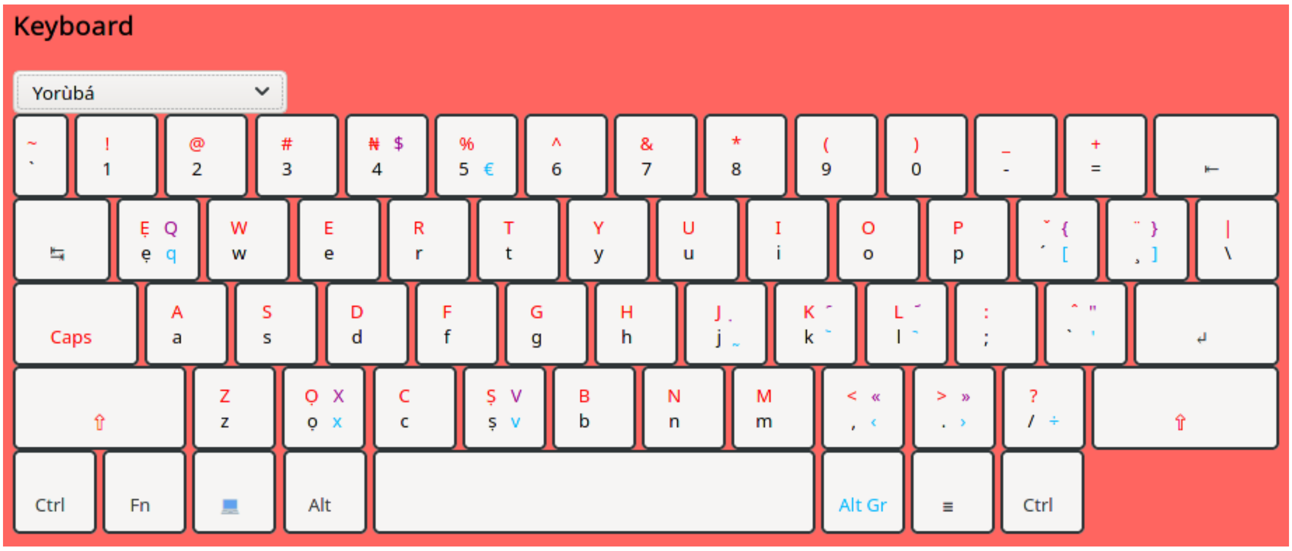

Keyboard
Yorùbá
~ ` ! 1 @ 2 # 3 ₦ $ 4 % 5 € ^ 6 & 7 * 8 (9) 0 _ - + =
Ẹ Q ẹ q W w E e R r T t Y y U u I i O o P p ˇ { ´ [¨ } ¸] | \
Caps A a S s D d F f G g H h J j K k L l : ; ^ " ` '
Z z Ọ X ọ x C c Ṣ V ṣ v B b N n M m < « , ‹ > » . › ? / ÷
Ctrl Fn Alt Alt Gr ≡ Ctrl

Figure 3: Soft Keyboard - Yorùbá shown.

by people willing to promote their language.

4 Further Work

Word limits were used to indicate the length of acceptable submissions. However, a word in one language may correspond to many more words in another language. Further work will involve finding a reasonable measure of estimating equivalent content between languages rather than using word counts.

Voting on suitability of stories by participants can be contentious. For some African languages it is easy to identify a panel of experts. For others, consensus on the written form of the language, and on different dialects is still developing. Community models for participation are required to further enhance this. In many countries where one language dominates, such as Rwanda, language standardization has been easier to accomplish. Many African countries are multilingual, and some languages are used across borders, making intergovernmental cooperation very helpful in language standardization efforts. Cooperation with institutions pursuing African language standardization work, would also be helpful, though many such institutions, when they exist are often poorly funded. For languages with a small number of speakers who form a minority section of the population of a country, community cultural organizations or social media groups are the only available and easily accessible resource.

Voting is done by issuing story authors and translators with an id which they enter on the website to be able to cast their votes. Their votes are then stored in a separate database from the id with the aim of enabling anonymity when there are more than two voters for stories in a particular language. It would be good to improve the system so that it is encrypted and the administrator cannot determine what each participant voted for. This is technically challenging since one can also possibly use server logs and database timestamps to correlate activity. Nevertheless, electronic voting has been done for other purposes which have greater security and identity verification requirements, so it should be possible to incorporate this in the current software.

The software has been partially translated into Kiswahili. At present this is done by creating a list of strings for each user facing component on the website. Many open source software projects use files formatted for Gettext (Free Software Foundation 2021*a*) to store translations. It would be good to use such a system.

Further work on making the soft keyboards easier to use is also needed, in particular for Ethio-Semitic languages such as Amharic. Soft keyboards were added to the software to enable use of orthography that may not be available on standard English and French computer keyboards available on computers in many African countries. Not all languages have standardized orthography, so adaptation and collaboration with linguists is required to do this well. A number of contributors indicated using keyboards on their mobile phones which have all the letters they need for their languages. Enabling creation, submission and viewing of stories through a mobile application in addition to a website may also be something to consider adding.

This note has mainly focused on the technical considerations in creating software to host an online competition for stories in African languages to collect a corpus. An additional article will analyze the collected corpora, and include viewpoints of those who helped obtain, write and translate short stories.

5 Conclusion

It is expected that many of Africa's languages will become extinct, or merged and standardized with other similar languages that are used in the same region, for example as has happened with Runyakitara. Documenting the rich cultural legacy is important for improved understanding and possible lessons that can be learned from the current linguistic diversity. Documenting the cultures is also helpful to enable cross-cultural understanding. The digital age requires digital tools, and while most of the technological developments enabling digitization have occurred in English speaking countries, these have been adapted for local conditions in countries with different cultural contexts. Notable examples of digital natural language processing tools have originated in Russia, China and Japan, some of these tools have been partly developed by commercial social media companies with a strong local presence. By providing open source software to enable short story competitions for African languages, it is hoped to stimulate further collection of short stories at relatively low cost, improve reading and writing skills in African languages, and develop natural language processing tools for African languages.

Acknowledgements

This work has benefited from a $4,000 grant provided by the International Research Development Centre administered by Knowledge for All (`https://www.k4all.org`) through the 2020 IndabaX-AI4D Innovation Grant program. The author thanks William Agbo, Liané Van Den Bergh, Audrey Mbogho, Cascious Mofokeng, Itaru Ohta, Kevin Scannell, Juan Steyn, Benito Trollip, Lilian Wanzare, and Constantine Yuka for many helpful discussions and assistance in running the competitions. The author also wishes to acknowledge all the people who participated in the competitions, by either submitting a story, submitting a translation, soliciting contributions from others, translating a flyer or checking submitted original and translated stories.

References

cPanel LLC (2021), 'cpanel website'.
URL: *https://www.cpanel.net*

ECMA International (2021), 'Javascript specification'.
URL: *https://www.ecma-international.org/publications-and-standards/standards/ecma-262/*

FluxBB Community (2021), 'Fluxbb website'.
URL: *https://fluxbb.org*

Free Software Foundation (2021*a*), 'Gettext website'.
URL: *https://www.gnu.org/software/gettext/*

Free Software Foundation (2021*b*), 'Gnu general public license'.
URL: *https://www.gnu.org/licenses/gpl-3.0.en.html*

GnuSocial Community (2021), 'Gnusocial website'.
URL: *https://www.gnusocial.rocks/*

PHP Community (2021), ‘Php website’.
URL: *https://www.php.net/*

Python Software Foundation (2021), ‘Python website’.
URL: *https://www.python.org/*

Ruby Community (2021), ‘Ruby website’.
URL: *https://www.ruby-lang.org/*

Saide (2021), ‘African story book website’.
URL: *https://www.africanstorybook.org/*

Scannell, K. (2021), ‘Indigenous tweets’.
URL: *http://indigenoustweets.com/*

Story Weaver (2021), ‘Story weaver website’.
URL: *https://storyweaver.org.in/weave_a_story_campaign*

Wordsmith Tools as an Enabler for Text Analysis

Rooweither Mabuya

South African Centre for Digital Language Resources
North West University

Roo.Mabuya@nwu.ac.za

Abstract

The process of intellectualization has been characterized as a planned process of accelerating the growth and development of South Africa's indigenous languages to enhance their effective interface with modern developments, theories and concepts (Finlayson & Madiba, 2002). Corpora are used to develop empirical knowledge about language. A corpus is a collection of naturally occurring texts derived from real life language use in either written or spoken form (cf. Sinclair, 1991). These texts are then processed, stored and accessed by means of computers for use in developing electronic resources of a language. It is thus an important resource in the development of isiZulu Human Language Technologies. Thus, specialized corpora were created for this study and were used as the basis to enquire into the sensitive use of isiZulu with reference to language phenomenon isiHlonipho. Corpus linguistics provides a more objective view of language than introspection, intuition and anecdotes. This study profited from the use of the WordSmith Tools version 6 software program suite, which allowed the researcher to use different programs simultaneously to analyse data.

Keywords: Corpus Building, Wordsmith Tools, Language for Specific Purposes, isiHlonipho

1 Introduction

It must be noted that this study is corpus based and therefore presents a detailed discussion on the composition of the corpora, that is, the Reference Corpus (RC) and the Analysis Corpora (AC) that were developed for this study. It also discusses the computational software used to query the corpora. Semi-automatic extraction methods and concordance queries are also explained. The Corpus Linguistics (CL) method was used to analyse gender sensitivity in isiZulu.

IsiZulu is the most widely spoken first language in South Africa, with 9 million speakers (Census, 2011) and is used in the media, and in the national and provincial parliaments. IsiHlonipho is generally defined as a practice that involves the use of a particular vocabulary and manner of speaking that is specific to a particular gender (Makoni, 2014).

Four LSP corpora were created for this study and were used as the basis to enquire into the sensitive language use of isiZulu. These were created from the following book titles: *Insila kaShaka*, *Bafa Baphela*, *Umdonsiswano*, and *Amandl' Esambane*, WordSmith Tools is an integrated suite of programs that is used for examining how words behave in a corpus.

2 Corpus Building

Corpus linguistics is a linguistic method that is based on the creation of a corpus. Franz (1996: 7) notes that. "Corpus based linguistics focuses on naturally occurring spoken or written language, as opposed to individual example sentences that are designed to illustrate grammatical theory". A corpus is thus a collection of naturally occurring texts derived from real life language use in written or spoken form. It is then processed, stored and accessed by means of computers. The corpus becomes an accurate form of linguistic data that mirrors the language under investigation.

Ngcobo and Nomdebevana (2010:187) indicate that, "one of the requirements for the development of a language is the planning of its corpus". A language is preserved for posterity and is also developed through corpus building, which aids its accessibility to the public and enables the development of human language technologies. A corpus is authentic language data which is designed and collected according to a specific sampling protocol or procedure.

It is important to note that many corpora of different sizes and typologies have been developed for different languages that are spoken globally. Some of these are available through a repository called the Sketch Engine

platform. Sketch Engine is a platform for corpora management and analysis and creation of corpora that is accessed through payment of a subscription. It has corpora in ninety (90) languages, with some like English having multiple corpora. African languages on Sketch Engine are Afrikaans, Arabic, Igbo, Setswana, Swahili and Yoruba.

There are very few corpora in African languages. Some corpora that have been developed for African languages do not appear on Sketch Engine and some require permission to access them. This is compounded by the fact that most of these corpora do not reside on the continent but are hosted and kept in European institutions (Khumalo, 2015:24).

Within the South African context, the University of Pretoria (UP) has led a massive effort to develop local corpora. Since this work started in the early 1990s, UP has developed corpora in all the country's official languages. Table 1 shows these corpora and their size.

LGP Corpus Name	**Acronym**	**Size**
Pretoria isiNdebele Corpus	PNC	1, 959, 482
Pretoria siSwati Corpus	PSwC	4, 442, 666
Pretoria isiXhosa Corpus	PXhC	8, 065, 349
Pretoria isiZulu Corpus	PZC	5, 783, 634
Pretoria English Corpus	PEC	12, 799, 623
Pretoria Afrikaans Corpus	PAfC	11, 602, 276
Pretoria Xitsonga Corpus	PXic	4, 556, 959
Pretoria Tshivenda Corpus	PTC	4, 117, 176
Pretoria Setswana Corpus	PSTC	6, 130, 557
Pretoria Sesotho sa Leboa Corpus	PSC	8, 749, 597
Pretoria Sesotho Corpus	PSSC	4, 513, 287

Table 1: Corpora in South African languages (De Schryver and Prinsloo, 2000)

The University Language Planning and Development Office (ULPDO) at the University of KwaZulu-Natal (UKZN) embarked on a massive isiZulu National Corpus (INC) building exercise as one of the university's major initiatives to develop isiZulu as a language of research, teaching and learning. The INC was piloted in 2014 at an impressive number of 1.3 million tokens. At the end of 2017, it reached the milestone of over 20 million tokens. This surpasses the earlier Pretoria Zulu Corpus (PZC) at 5.7 million tokens and the Ukwabelana Corpus with 100 000 tokens. The PZC is an LGP corpus and the Ukwabelana is a LSP annotated one that is used as a learning resource. The INC is an LGP monitor corpus which is balanced in terms of text and thematic content.

Corpora are characterised according to medium, design, size, language variables, and mark-up and annotation (McEnery et al., 2012). Each of these components will be further discussed as follows:

2.1 Medium

This refers to the fact that the text can either be printed or handwritten and that the text that makes up the corpus can be electronic or digitised speech.

2.2 Design method

The design refers to the manner in which the corpus collection plan is executed. The method used to collect the corpus can either be balanced or opportunistic. Balance refers to the representativeness of the texts selected to make up the corpus. The opportunistic method relies on any texts made available to the compilers for addition to the corpus.

2.3 Size

The size can either be fixed or open ended, meaning that it is a monitor corpus. The former refers to a typology where the size of the corpus is planned, and once the collection reaches the set target, corpus building stops. The latter refers to a big and continuously growing corpus. A bigger corpus is considered to be better and was used in this study as an RC. This study thus used both the small (AC) and the big (RC).

2.4 Language variables

This refers to the language in which the corpus text files are written. A corpus builder constructing a monolingual corpus will have one language from which to draw his/her texts, and may choose to tag all references to other languages in the texts as foreign. Language variables differ in many ways because the researcher has to choose between a monolingual and a multilingual corpus. Other decisions in this regard include whether the corpus will be synchronic or diachronic, a general language corpus or a language for special purpose, and a written or spoken language.

2.5 Mark-up and annotation

In a raw corpus, the texts that comprise it are collected and stored as plain text for use in their original form. Corpus mark up or annotation is the process of adding interpretive text or further analysis by way of tags on the data to enhance its reusability. A corpus can thus be raw (plain), marked-up or annotated.

One of the main characteristics of a corpus is representativeness. This is characterized as “the extent to which a sample includes the full range of variability in a language or language variation” (Biber, 1993). The corpus should thus strive to mirror the language as it exists.

Corpus linguistics can be used to investigate many kinds of linguistic questions. It has the potential to yield highly interesting, fundamental and often surprising new insights into language and has become one of the most common methods in linguistic investigation. Corpus representativeness depends on two factors. The first is balance that entails the range of genres and registers included in the corpus sampling, while the second is the techniques used to select the text extracts for each genre.

Texts in a corpus need to be converted into electronic format in order for the corpus to be compatible with the data handling software program. Hardcopy materials need to be scanned for electronic access to the corpus. In this study the data that makes up the corpus was selected, analysed, cleaned, and then stored electronically using the Wordsmith Tools version 6 software program suite. It is important to remember that any document prepared for corpus analysis is only a representation of the original one. According to Römer and Wulff (2010), “corpus linguistics can assist the researcher to assess and describe a linguistic phenomenon in a maximally objective and hence largely theory-neutral fashion”.

The advantage of a CL approach is that it studies the words in context. It assists in comparing the use of words in different documents and in determining how words work together. It is used to analyse and research a number of linguistic questions and offers insights into the dynamics of language, which has made it one of the most widely used linguistic methodologies. Corpus linguistics relies on computers for speed, accuracy, reliability and verification by others.

3 Wordsmith Tools

This section presents a detailed discussion on the software program used to query the LSP corpus. A variety of corpus analysis tools are available. As the CL field advances, several off-the-shelf software tools have been developed and packaged with graphical user interfaces like day-to-day software. These are popularly known as corpus managers, corpus browsers or corpus query systems. They provide facilities for searching for language forms and sequences, and analyzing corpus chunks.

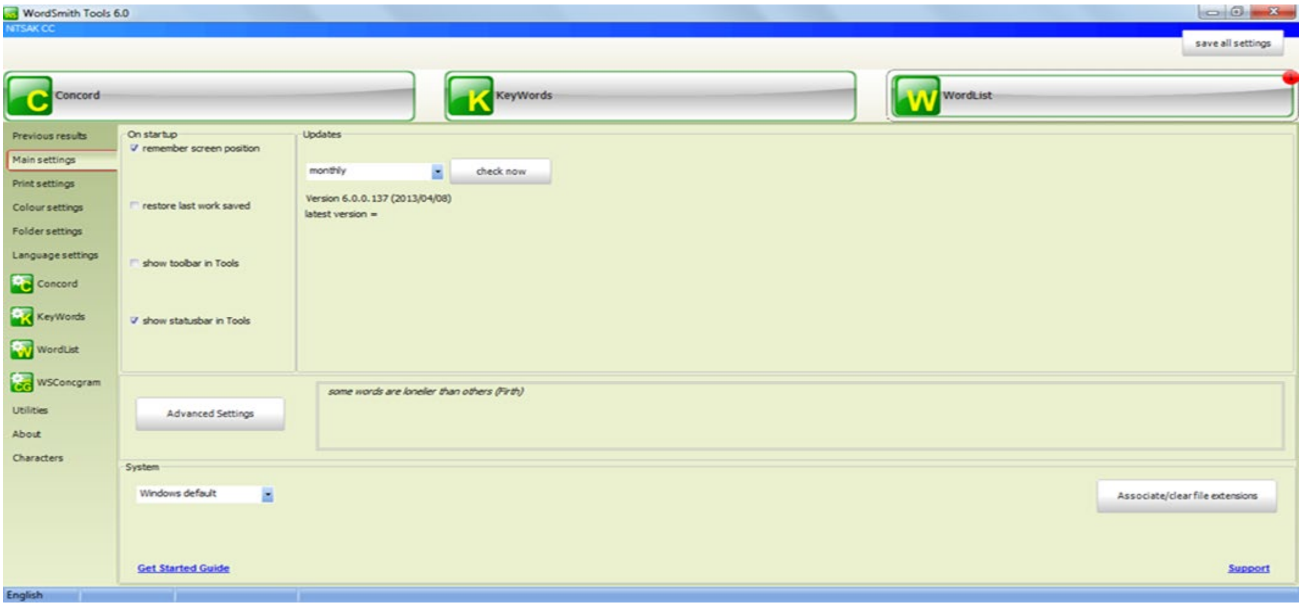
Corpus managers may be web or desktop-based, commercially marketed or open-source, and designed for specific corpora or generic.

Desktop software solutions include WordSmith (Scott, 1996), Antconc (Anthony, 2011), MonoConc Pro (Barlow, 2000), CasualConc (Imao, 2013), Concapp (Greaves, 2007), and aConCorde (Roberts, Al-Sulaiti, and Atwell, 2006). This study employed the WS Tools software solution.

Wilkinson (2011) states that "most corpus analysis programs include a concordance which finds all the occurrences of a search word, or search pattern, and displays them in the centre of your screen as a keyword in context (KWIC) display, together with a span of co-text to the left and right".

WordSmith Tools is an integrated suite of programs for examining how words behave in a corpus. It was developed by Mike Scott with the first version released in 1996. As a suite of programs, WordSmith Tools contains three programs that are used to query a corpus, namely, the concordance, key word, and word list suites. WordSmith Tools version 6 is a suite of programs that enables manipulation of corpus data according to frequency, word list, concordance, collocations and KWIC. Otlogestwe describes these tools as follows; a wordlist tool can be used to produce wordlists or word-cluster lists from a text. A concord can give any word or phrase in context so one can see what other words occur in its vicinity. Keywords calculate words which are used much more frequently in a text (Otlogestwe, 2014: 270). Figure 1 below shows the interface of WS Tools and the three main functions.

Figure 1: Interface of Concordance, Keyword and Wordlist functions



On startup
☑ remember screen position
□ restore last work saved
□ show toolbar in Tools
☑ show statusbar in Tools

Updates
monthly
check now
Version 6.0.0.137 (2013/04/08)
latest version =

some words are lonelier than others (Firth)

System
Windows default

Creating a word list is useful because this is where and how the investigation is prompted and generated with high and low frequency ratios. According to Römer and Wulff (2010), "[…] a useful first step in approaching a corpus or text is to generate a list of all the words that occur in it together with their frequencies". A frequency list provides a range of diverse types of words, tokens, or forms which make up a corpus. The word list in Wordsmith Tools version 6 is able to create alphabetical, high-to-low ratio sorted lists.

While a word list highlights what is frequent in a corpus or text, it does not tell us what is significant or unusually frequent. The keyword list enables the researcher to identify the most outstanding or unexpectedly frequent words in a corpus. It compares a frequency wordlist based on the corpus under analysis (AC) with another frequency wordlist based on an RC.

Table 2 lists the acronyms used in the Wordsmith Tools program that are used when testing for keyness of words when the AC is compared with the RC.

Column	**Description**
Keyword	List of keywords
Freq.	Number of occurrences of each keyword within the source text(s) in which these key words are key
%	Percentage value of the frequency of the keyword in the source text
RC Freq.	Number of occurrences of each keyword in the reference corpus (RC)
RC %	Percentage value of the frequency of the key word in the reference corpus
Keyness	Statistical calculation that factors in the frequency of a word in each wordlist and limits it with the probability value (p)
P	Value used in statistics to indicate the probability of obtaining a wrong result; a high p value implies a high chance of that word not being a key word

Table 2: Acronyms used in the Wordsmith Tools program

A concordance is an alphabetical list of the primary words used in a book or body of work. It lists every instance of each word with its immediate context. Barnbrook (1996) notes that a concordancy list "provides a simple way of placing each word back in its original context, so that the details of its use and behaviour can be properly examined". Concordances are usually displayed in KWIC format, with the search word (or phrase) shown in the middle of the screen and some context to left and right of it. A concordance highlights a list of particular words or a sequence of words in context and is at the centre of CL as it enables access to many important language patterns in texts. In a now famous observation, Firth (1957:179) stated that "you shall know a word by the company it keeps". This means that collocates will occur in a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text.

According to Wodak and Meyer (2016), the CL methodology provides statistical and coherence and concise analytical views on written information, computing frequencies and measures of statistical significance, as well as presenting data extracts so that the researcher can access single occurrences of search words, qualitatively examine their collocational environments, describe salient patterns and identity discourse functions.

4 Collections

The current study comprised of a small, targeted LSP corpus collected from four isiZulu literature text books, namely, Insila KaShaka by JL Dube (second male author in the history of isiZulu writing), Bafa Baphela by Joyce Jessie Gwayi (first female author in the history of isiZulu writing), Umdonsiswano by Condy Nxaba (contemporary male author), and Amandl' Esambane by Babhekile Ngcobo (contemporary female author). P Lamula's UZulu kaMalandela was the first piece of fiction written by a Zulu person in 1924 and JL Dube's Insila kaShaka was the second in 1930. The first book by a Zulu author was Magema Fuze's Abantu abamnyama lapho abavela khona in 1922 which was not a literary work. This required that the researcher use JL Dube's book for this study as it was easy to access. Each of these books was used as an AC. An LSP corpus is a language specific corpus, which

looks into language that is restricted to a particular domain (Bowker and Pearson, 2002). The advantage of using LSP corpora is that it is easier to identify specialized terms, thus providing a large amount of information with regard to the words, frequency, structure and style in the specialised language. Language for Specific Purposes is also useful in picking up the difference between standard language and specific registers as well as neologisms (new terms) that emerge in the language (Weisser, 2016). The AC is a technical corpus, meaning that it is domain specific. The motivation for choosing these texts was basing the analysis on traditional/old Zulu and contemporary Zulu. Insila kaShaka with 19 425 tokens and Bafa Baphela with 22 103 tokens fall under the traditional era of writing whilst Umdonsiswano with 49 341 tokens and Amandl' Esambane at 31 251 tokens follow the contemporary way of writing. The LSP corpora were used to compare gender sensitivity in traditional literature to that in contemporary literature to investigate any language shift and ensure a balance when investigating the corpus. An RC is a non-technical corpus that is designed according to Oostdijk et al., (2021) to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference material. The model for selection usually defines a number of parameters that provide for the inclusion of as many sociolinguistic variables as possible and prescribes the proportions of each text types that is selected (Oostdijk et al., 2013).

The texts collected for the AC and RC were scanned and converted to plain text format for WordSmith Tools compatibility. The data analysis is enhanced by the extent to which the corpus data has been processed and annotated for ease of reusability. According to John Sinclair, a corpus can either be annotated, meaning that linguistic analysis has been performed on the text or orthographic, meaning that the texts have not been linguistically analysed. He labels the latter a raw corpus (Sinclair, 1991). The corpus used in this study is raw and hence not annotated.

Four Zulu texts were analysed to test for gender sensitivity using a set of words that co-occur known as collocates. Linguistic analyses that use CL methods and tools thus do not represent the entire language but sample texts that mirror the language that needs to be analysed. Corpus materials were collected from the carefully selected bodies of published literature described above. The data was collected by gathering texts from each novel/literary work. The corpus was collected and analysed using the Wordsmith Tools software program (WS Tools). The study profited from the key data handling functionalities of WS Tools such as the Frequency List tool, and the Concordance tool. To enable a closer examination of the corpus data, a Word List and a Frequency List was created from the entire corpus using WS Tools. The Concordance analysis was then conducted on selected content words that were regarded as instructive and informative to the study.

The corpus was analysed and queried in relation to gender sensitivity. One of the advantages of CL and Critical Discourse Analysis is that they enable qualitative and quantitative analysis. The quantitative analysis seeks to measure and qualify the data in order to numerically represent a given reality and examines the overall types and tokens, the wordlist and the frequency of a key content word. According to Weisser (2016), type is a representative word in the frequency list of a corpus and token refers to the individual occurrence of a particular type, that is, the number of words in a corpus. A wordlist lists the words that occur in the corpus, either in alphabetical order or in the order of frequency. A frequency list records the words in the corpus according to the number of times they are used. The qualitative analysis seeks to elaborate on the quantitative analysis by identifying the most frequent and less frequent words. The salient features that emerged from the corpus were in relation to how the texts relayed their discourse on gender sensitivity. This was done by looking at the words that occur before and after the context word and is known as the KWIC. The study thus profited from a judicious mixture of quantitative and qualitative approaches.

5 Querying

The AC is an LSP corpus with 122 121 running words and the RC is an LGP corpus with 2 039 691 tokens. Each book was analysed for keyness in isolation and against gender keywords. Keyness refers to the extent to which texts show a specific stylistic profile when compared with another set of texts. It is quantified by measuring the positive or negative differences in the lexis of author(s) in juxtaposition to the lexis of the texts in a reference corpus with which the former is compared. The researcher assembled and analysed five different corpora. They are labelled Insila kaShaka (AC1), Bafa Baphela (AC2), Umdonsiswano (AC3), Amandl' Esambane (AC4), and lastly the IsiZulu National Corpus, labelled RC.

To determine the words' keyness, the frequency ratios of words in the AC was compared to the frequency ratios of those words as they appear in a general corpus (the RC). According to Mason and Platt (2006:159) "we count how often a particular lexical item occurs in the text. Then we work out how often we would expect the item to occur in a text of that length, using the item's frequency in a large corpus as an indication. The ratio between the observed frequency and the expected frequency then tells us whether a word is significant or not".

Words that occur more frequently in the AC than in the RC have a positive keyness value while those appearing less frequently have a negative value. This revealed the words which are sensitive to gender in isiZulu. The keyness analysis provided clear evidence of the distinction between male speech and female speech (or its representation) in isiZulu. In the English language it is understood that collocates which accompany male/boy are usually positive words like strong, tall, brave, etc. Those that accompany female/girl are associated with being negative such as weak, slim, scared, etc. Using the CL approach, similar paradigmatic patterns were analysed in this study since the Zulu culture is known to be patriarchal. According to Atanga et al. (2013) use of isiHlonipho is linked to patriarchal control and regulation of women's behaviour. The linguistic custom of isiHlonipho is linked to a strict code of behaviour (ukuhlonipha). Mathonsi and Gumede (2006) note that, Ukuhlonipha requires a

Zulu woman to refrain from directly and publicly voicing her opinion as a sign of her femininity. Women are known to be marginalized in their speech in the Zulu language and their speech is different from that of males due to the use of isiHlonipho. However, this phenomenon is no longer widely practised in the speech of most modern Zulu women and they virtually use the same speech register as men due to gradually shifting hierarchies, status and power.

6 Discussion

In the data analysis, we first looked at words that characterize maleness to check for any interesting details by creating a wordlist. It is interesting that the following words are high in frequency ratios; *uJeqe*, *inkosi*, *uShaka*, and *izinkomo* and it is apparent that their function is closely related to that of maleness and power.

A concordance list was created for the analysis of the word *izinkomo*. In traditional Zulu culture, cattle are associated with power and wealth. In this analysis, the notion of *izinkomo* belonging to male members of society is apparent; wealth is inherited by males from generation to generation.

N	Concordance
3	Injongo yethu ukuyophanga ***izinkomo*** laphaya njengoba wonke
4	Useziqoqe zonke ***izinkomo*** zesizwe; amanye amabutho
5	abafana ababafice belusa ***izinkomo*** zesizwe sakwaMoloi edlelw
6	alelile, bazikhomba zonke ***izinkomo*** zesizwe sakwaMoloi, beben
7	hweni akhe awashiye elusa ***izinkomo***. Endleleni wathuma enye
11	khotha eyikhothayo." ***Izinkomo*** zikaMatiwane zase zihamba
12	nje indaba yokuba amudle ***izinkomo*** uMatiwane. uMatiwane wa
16	dwa inkosi yenu ukungipha ***izinkomo*** bese ibuye izilanda ngale
17	into esingaba sisayenza ***izinkomo*** sezingezami." Nanxa yayis
18	amaHlubi eyintshontshela ***izinkomo*** zayo yaqala ukubona
19	honisa kwelamaHlubi ezabo ***izinkomo***. Ngenxa yabo manje nge
20	zisebenzisa ukuze abuyise ***izinkomo*** zami azintshontshayo."

Table 3: Concordance List of AC 2

Power relations are evident in the selected concordance lines above (see Table 3) where *izinkomo* is shown to belong to the nation and as a form of wealth in "[…]useziqoqe zonke ***izinkomo*** zesizwe" (*he has collected all the cattle that belong to the nation*).

Power relations are discursive, and discourse in itself constitutes society and culture, are evident in this excerpt, where cattle are represented as a sign of power and wealth in traditional Zulu society. It is also evident in the concordances in line 16 that cattle have a transactional value akin to contemporary monetary value.

It is clear from the collocates in Table 3 that *izinkomo* (cattle) are associated with the brute, imperial power of seizing cattle (cf. line 3). *Izinkomo* are also associated with a specific gender, highlighted as collocates *amabutho* (warriors), *abafana* (boys), *inkosi* (king), and *–ntshontsh-* (theft). It is clear from the associated meaning that *izinkomo* represents maleness and power in Zulu culture and society. Using the CL analysis, the collocates thus highlight the portrayal of power relations. Cattle are a form of wealth and such wealth is associated with maleness and power.

N	Concordance
1	omdala. Impilo inzima, ***ndodana*** . Kodwa-ke ningalilahli ithemba
3	muhle impela umsebenzi wakho, ***ndodana*** . Isebenza kanjena-ke indoda
4	"Kahle, kahle, sekwanele, ***ndodana*** . Usubonga sengathi sengik
5	ngiphelile yinsini. "Bamba-ke, ***ndodana*** . Usuyoqala ukufundela uku
7	shayela. "Yindawo yami le, ***ndodana*** . Isuka la ize iyothiywa
8	emzini owodwa. Uthini ke ***ndodana*** ? Uyayithatha le ndawo? Ka
15	"Sengathi ibambe ngako, ***ndodana*** ," kuhleba uNkwanyana ebheka

Table 4: Concordance List of AC 3

Table 4 presents the concordance list of the word *ndodana* (son) taken from AC 3. The collocates accompanying *ndodana* in line 1, clearly show that the father is encouraging the son not to lose hope/faith. Positivity is emphasized. In line 3, the son receives praise and is celebrated for his manhood. The emphasis is that 'doing well' is the hallmark of manhood or maleness "muhle impela umsebenzi wakho, ***ndodana***. Isebenza kanjena-ke indoda" (*your work is indeed impressive, son. This is precisely how a man works*). It is therefore observed that in Zulu culture and society, success and a positive impression are associated with a man.

The advantage of analysing concordances is that they enable one to closely study the texts. This is because the search item is isolated and highlighted. It is also accompanied by context on either side to figure out the meaning of an unknown word and to disambiguate problematic senses of words. It is an attested fact in linguistics that we know more about a word through the company it keeps.

Concordances list all instances of a word found in the selected corpus which saves time as one does not have to go through each text file separately and extract relevant examples.

It is important to note that there seems to be no distinction in the use of gender sensitive language between the two authors writing in the same period, that is, traditional early writing. It is instructive that the AC 2 corpus evidence by the female author shows the highest frequent word *amabutho 4* (warriors) as number 4 in the wordlist. Although *indlovukazi, 5* (Mother of the King/Wife of the King) is at number 5 in the frequency wordlist, interesting references like the traditional instrument of war, *ngemikhonto, 134* (with spears) and names like *Bhekimpi, 150* (Looking after the army/Ready for attack) are words associated with maleness.

These examples clearly show that both the male and female authors' narratives or discourses are framed within the larger socio-cultural influences that reflect the world as dominated by male power and influence. Even the word *indlovukazi* is defined through its association with maleness, either as the "mother of the King" or "wife of the King". The corpora AC 2 and AC 3 therefore clearly show that the use of isiZulu is gendered in that images of power, wealth, precision, success, strength, war, conquering, etc. are all associated with the male gender.

Given the history of patriarchy and a culture that views females as weaker, it is clear that isiZulu is a prejudicial language where women are portrayed as weaker than the opposite sex. The *isiHlonipho* register was offered as an apt example of the skewed relations in language use between the sexes in isiZulu.

The existence of *isiHlonipho* as a gender specific register begs the question of whether isiZulu as a language has a way of referring to both sexes without offending either. It was evident from the corpora that males or maleness is viewed positively, and is associated with progress, while females or femininity are either absent from the discourse or are viewed through the lenses of maleness such as references like *indlovukazi*. It is clear that females will remain inferior to males due to patriarchal norms that are deeply embedded in Zulu culture. The corpus clearly displayed in the collocates, that words which refer to females are mostly accompanied by a diminutive, in contrast to those that refer to males.

It is further notable that there is no corpus evidence from the contemporary literature by both the male and female authors that suggest a marked departure from a historical gendered language skewed towards celebrating maleness towards more neutral and gender sensitive language use.

7 Conclusion

In order to document and make explicit attitudes towards women, a multi-method approach incorporating mainly qualitative methods and supported by quantitative methods, was used to analyse the data. The advantage of such an approach is that the richness and precision of qualitative analysis is combined with statistically reliable and generalizable results (Schmied, 1993). In general, qualitative research is used to describe and answer questions from a participant's point of view. The data is used to identify items, explain aspects of usage and provide real-life examples of such. The inductive process of qualitative studies proceeds by way of general questions, the collection of enormous amounts of data, carefully observing the data and then presenting the findings, which may produce tentative answers about what was observed (Glesne and Peshkins, 1992).

This research therefore incorporated a collection and organization of textual information into a corpus, and an investigation and querying of the corpus in order to discern relations and social practices to do with an isiZulu speech community or society. It is notable that CDA is concerned with a careful and lengthy investigation of major causes events and results of issues thereto. It accordingly, requires a record of point by point connections between content, talk, society and culture.

8 References

Adolphs, S. 2000. *Introducing electronic text analysis.* New York: Routledge.

Anthony, L. 2011. *AntConc* (Version 3.2. 2) [Computer Software]. Tokyo, Japan: Waseda University.

Baker, P. and McEnery, T. 2005. A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts, *Journal of Language & Politics*, 4:2, 197–226.

Barlow, M. 2000. *Monoconc Pro 2.0*: Athelstan.

Barnbrook, G. 1996. *Language and Computers: A Practical Introduction to the Computer Analysis of Language.* Edinburgh: Edinburgh University Press.

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus linguistics: Investigating language structure and use.* Cambridge University Press.

Bower, L. and Pearson, J. 2002. *Working with Specialized Language: A practical guide to using copora.* London: Routledge.

Chitauro-Mawema, M. 2006. Gender Sensitivity in Shona Language Use: A lexicographic and corpus-based study of words in context.

Dowling, T. 1988. *IsiHlonipho Sabafazi: The Xhosa women's language of respect. A sociolinguistic exploration.* MA Thesis, University of Cape Town, South Africa.

Finlayson, R. 1982. Hlonipha – the women's language of avoidance among the Xhosa. *South African Journal of African Languages*, 1:1, 35-60.

Franz, A. 1996. *Automatic Ambiguity Resolution in Natural Language Processing: An Empirical Approach.* Tokyo: Springer.

Glesne, C. and Peshkins, A. 1992. *Becoming Qualitative Researchers: An introduction.* White Plains, NY: Longman.

Hammer, A. and Damascelli, A. T. 2002. *Corpus Linguistics and Computational Linguistics: An overview with Special Reference to English.* Torino: Celid.

Khumalo, L. 2015a. Advances in Developing corpora in African languages. *Kuwala, 1:*2, 21-30.

Khumalo, L. 2015b. Semi-automatic Term Extraction for an isiZulu Linguistic Terms Dictionary. *Lexikos, 25* (AFRILEX-reeks/series 25:2015), 495-506.

Khumalo, L. n.d. Corpora as agency in the intellectualization of African languages.

Krishnamurthy, R. 2011. Accessing all areas: Corpus Analysis methods in inter-disciplinary applications.

Luthuli, T. 2007. *Assessing Politeness, Language and Gender in Hlonipha.* MA Thesis, University of KwaZulu-Natal, South Africa.

Mabuya, R. 2018. *A corpus linguistic analysis of gender sensitive language in isiZulu.* MA Thesis, University of KwaZulu-Natal, South Africa.

Makoni, B. 2014. Feminizing linguistic human rights: use of isihlonipho sabafazi in the courtroom and intra-group linguistic differences, *Journal of Multicultural Discourses*, 9:1, 27-43.

Mason, O. and Platt, R. 2006. Embracing a new creed: Lexical patterning and the encoding od ideology. *College Literature* 33:1, 155-170.

Mathonsi, N. and Gumede, S.H. 2006. Communication through performance: Izigiyo Zawomame as Gendered Protests. *Journal of Southern African Linguistics and Applied Language Studies*, 24:4, 483-494.

McEnery, T. and Wilson, A. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Ngcobo, M.N. and Nomdebevana, N. 2010. The Role of Spoken Language Corpora in the Intellectualisation of Indigenous Languages in South Africa. *Alternation: Interdisciplinary Journal for the Study of the Arts and Humanities in Southern Africa.* 17:1, 186-206.

Oostdijk, N., Reynaert, M., Hoste, V. and Schuurman, I. 2013. The construction of a 500 million word reference corpus of contemporary written Dutch: *Essential Speech and Language Technology for Dutch.* (eds) Spyns, P and Odijk, J. Springer: Heidelberg.

Otlogetswe, T.J. 2014. Extracting business terms for dictionary subject label. In *African Languages and Linguistic Theory.* CASAS Book Series 109. Cape Town: CASAS.

Scott, M. 1996. *WordSmith tools*: Oxford: Oxford University Press.

Sinclair, J.M. 1991. *Corpus concordance collocation*. Oxford: Oxford University Press.

Sinclair, J. M. 2004. *How to use corpora in language teaching* (Vol. 12). John Benjamins Publishing.

Spiegler, S., Van Der Spuy, A., and Flach, P. A. 2010. *Ukwabelana: An open-source morphological Zulu corpus*. Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics.

Spiegler, S. 2011. *Machine learning for the analysis of morphologically complex languages*. University of Bristol.

Swales, J. 2000. Languages for Special Purposes. *Annual Review of Applied Linguistics,* 20, 59-76. Cambridge: Cambridge University Press.

Weisser, M. 2016. *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*. United Kingdom: Wiley Blackwell.

Wilkinson, S. 2011. Analysing focus group data. In D. Silverman (Ed.), *Qualitative research (3rd ed., pp. 168–184).* London, UK: Sage.