http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

Using Semi-automated Term Extraction for IsiNdebele Health Terminology

Nomsebenzi Malele, *Department of African Languages, University of South Africa, Pretoria, South Africa (malelnj@unisa.ac.za) (https://orcid.org/0000-0001-8384-7853)*
and
Sonja Bosch, *Department of African Languages, University of South Africa, Pretoria, South Africa (seb@hbosch.com) (https://orcid.org/0000-0002-9800-5971)*

Abstract: IsiNdebele, also known as Southern isiNdebele, has a limited availability of language resources and specialised terminology, especially when compared to other members of the Nguni language family. This study therefore explores means of addressing the shortage of specialised terminology in isiNdebele by using semi-automatic term extraction methods. The focus is on health terminology, intended for communication with laypersons rather than between experts in the health field. Semi-automatic term extraction methods are employed, combining manual identification and extraction of data from available corpora with the use of a software tool named WordSmith Tools (WST). The study illustrates the necessity of utilising all functions of the WST, as they complement each other. Terms overlooked by one function may be captured by another. For instance, while the KeyWords function identified only a limited number of terms in this research, manual identification proved more fruitful. Interestingly, the Concord function emerged as particularly effective in identifying a greater number of terms. The use of the WST in this research highlights the viability of corpus-driven studies, even for resource-scarce languages like isiNdebele. Therefore, considering the limited resources available for isiNdebele, particularly the absence of specialised dictionaries, this collection of health terms exemplifies ideal candidates for inclusion in a general dictionary.

Keywords: ISINDEBELE, CORPUS-DRIVEN TERM EXTRACTION, HEALTH CORPORA, LANGUAGE FOR SPECIFIC PURPOSES (LSP), LANGUAGE FOR GENERAL PURPOSES (LGP), WORDSMITH TOOLS, WORD LIST, KEY WORDS, CONCORDANCE, SEMI-AUTOMATIC EXTRACTION

Irhunyezorhubhululo: Ukusebenzisa Indlela Yemitjhini Nezandla Ukutsomula Itheminoloji yesiNdebele Yezamaphilo. IsiNdebele, esibuye saziwe ngokobana siNdebele seSewula, sitlhayelelwa khulu mithombo yelimi kanye netheminoloji ekhethekileko khulukhulu, lokha umuntu nakasimadanisa namanye amalimi wabeNguni, isiNdebele esiyingcenye yawo. Ngalokho irhubhululweli lihlola iindlela zokuhlangabezana nalokhu kutlhayela kwetheminoloji ekhethekileko esiNdebeleni. Lokhu kwenziwa ngokusebenzisa iindlela zokutsomula amathemu kusetjeniswa imitjhini nezandla. Umnqopho werhubhululweli usetheminolojini yezamaphilo. Kuhloswe bona imikhulumiswano ibe lula hlangana nabantu abangasi lilitho kunokobana kube nemi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

khulumiswano elula hlangana nabocwephetjhe bomkhakha wezamaphilo. Njengombana sekuveziwe ngehla ukobana kusetjenziswa indlela yokutsomula amathemu ngomtjhini nangezandla, kilelirhubhululo, umtjhini osetjenzisweko ubizwa bona yiWordSmith Tools. Lelirhubhululo litjengisa ukuqakatheka kokusetjenziswa kwawo woke amathulusi weWordSmith Tools (WST) ngombana, womathathu aphelelisana kuhle khulu. Lokho kutjho bona amathemu angakhange alemukwe ngelinye ithulusi, ayalemukwa ngelinye. Isibonelo, njengombana iKeyWords ikghone ukulemuka amathemu ambalwa kangaka, ithulusi iConcord lona likwazile ukulemuka amathemu amanengi ngendlela erarako. Ukusetjenziswa kwe-WST kilelirhubhululo kuveza ngokusobala ukusebenza kuhle kwerhubhululo elisunduzwa yikhophasi nemalimini atlhayelelwa ziinstjenziswa njengesiNdebele. Ngalokho lokha nawutjhejisisa ukutlhayelelwa kwelimeli zizinto ezifana neenhlathululimezwi ezikhethekileko, lokhu kubuthelelwa kwamathemu wezamaphilo kuveza lawo mathemu angahle afakwe kusihlathululimezwi esivamileko.

Amagama aqakathekileko: ISINDEBELE, UKUTSOMULWA KWAMATHEMU OKUSUNDUZWA YIKHOPHASI, IKHOPHORA YEZAMAPHILO, ILIMI LOMNQOPHO OKHETHEKILEKO (LSP), ILIMI LOMNQOPHO OVAMILEKO (LGP), IWORDSMITH TOOLS, IWORD LIST, IKEYWORDS, ICONCORDANCE, UKUTSOMULA NGOMTJHINI NANGEZANDLA

1. Introduction and background

IsiNdebele, also known as Southern Ndebele (ISO 639-3: ndl)[1], is one of the twelve official languages of South Africa and is primarily spoken in the former kwaNdebele region of Mpumalanga. It belongs to the Nguni group of languages which includes isiZulu (zul), isiXhosa (xho), Zimbabwean Ndebele (nde) and Siswati (ssw). IsiNdebele exhibits a morphological complexity based on a robust noun class system resulting in the extensive use of prefixes and suffixes. These morphological elements play a crucial role in shaping the meaning of words and sentences. In Nguni languages a conjunctive orthography is employed which is characterised by a seamless and interconnected representation of linguistic elements (Taljard and Bosch 2006: 432-433), whereas the Sotho group of languages, including Sesotho sa Leboa (also known as Northern Sotho or Sepedi) employs a disjunctive orthography that introduces distinct boundaries between linguistic units.

Despite being a long-established spoken language, isiNdebele was first given full written status in 1985, when it was first introduced in classrooms. Prior to 1985, isiZulu rather than isiNdebele was the language of teaching for the children of the Ndebele people. In 1996, the language was examined for the first time as a subject for matriculation (Jiyane 1994: 1). In comparison to other official languages, especially those of the Nguni language family, isiNdebele performs poorly in terms of language resources. There is no language for specific purposes (LSP) dictionary available for isiNdebele. Only general-purpose (LGP) dictionaries exist, indicating the scarcity of specialised terminology in isiNdebele.

Terminology plays a crucial role in lexicography as it shapes the precision and clarity with which dictionaries and other lexical resources convey meaning.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

Due to the scarcity of language and lexicographic resources, general dictionaries could, according to Gouws and Prinsloo (2005: 61), include a broader selection of terms from clearly defined specialised fields. The treatment of such terms should be tailored to the layperson encountering them in everyday communication, rather than to experts in a specific field. In addition, the evolution of terminology reflects changes in language and society, influencing how new words are incorporated and defined. Thus, careful consideration of terminology is essential for maintaining the relevance and usefulness of lexicographical works.

Furthermore, the lack of standardised terminology has a detrimental effect on the growth of a language as this means that no new terms are created. Finlayson and Madiba (2002: 53) argue that for the terminology of a language to develop, intellectualisation must take place. They maintain that through this technique, underdeveloped African languages will develop more rapidly, and their terminology will carry the full weight of scientific rigour and clarity. In addition, they (ibid.) emphasise that intellectualisation ensures that language changes in a way that gives it the ability to carry and communicate all types of knowledge across all domains of life. The inclusion of terminology from well-defined specialised fields in general dictionaries ensures that the dictionaries reflect current and accurate language usage within a particular field. This approach is essential for keeping dictionaries up-to-date and relevant, particularly in rapidly evolving disciplines.

It is against this background that the current study examines the role of corpus-driven term extraction in filling the lexicographic gaps created by the lack of terms, with a particular focus on isiNdebele health terms. The research also establishes the success of the WordSmith Tools (WST) in identifying terms in a language with such a conjunctive orthography and the extent to which the WST reduces manual work.

In the next section an overview of similar research conducted in other African languages, will be reviewed. The focus will be on the types and sizes of corpora used, whether the corpora were written or spoken, and whether term extraction and analysis were done manually or with the assistance of software tools. This will be followed by a description of the resources that were used in this study. An exposition of the methodology used for term extraction of isiNdebele health terminology, a description of the results, and recommendations will be given. Finally, a conclusion and suggestions for future research will be presented.

2. Related work

The focus in this section is on semi-automatic term extraction conducted for other African languages. Of interest for this study are the types of tools used, the feasibility, practicality, and successes of various methodologies, emphasising the need for both computational and manual methods in terminological activities.

Taljard and De Schryver (2002) conducted a pioneering study on semi-automatic term extraction using basic corpus query software for African languages, particularly Sesotho sa Leboa. They employed three functions of the WordSmith Tools (WST): WordList, KeyWords, and Concord. Their findings revealed that the corpus query tool successfully identified 40% of multi-word linguistic terms that were overlooked manually. This led to the conclusion that semi-automatic term extraction significantly reduces human errors, proving to be both feasible and practical for African languages. Their findings also emphasised the fact that human beings will always remain the final judges in any terminological activity, whether that endeavour be manual or computational.

Building on the findings of Taljard and De Schryver, Prinsloo (2015) analysed corpora in Sesotho sa Leboa, English, and Afrikaans to assess the efficacy of restricted corpora for lexicographic endeavours. Using the Sketch Engine software, Prinsloo focused on frequencies and collocations, discovering that even lesser-resourced languages with limited, unbalanced corpora could yield results comparable to those of more resource-rich languages. This study challenged the notion of a "Big corpus", arguing that outcomes in languages with fewer resources such as the African languages, can be on par with those in languages with abundant resources.

Nkomo and Madiba (2011) employed semi-automatic term extraction to compile economics terminology for isiXhosa and Tshivenḓa. The study aimed to support concept literacy for students who are non-native English speakers. They used the WST and Multiconcord software, starting with the WordList function to generate a basic word list. After verification by economics lecturers, a final word list was created, and concordances were generated to identify term meanings in various contexts. This approach drew attention to the value of semi-automatic tools in educational settings.

Khumalo (2015) focused on the semi-automatic extraction of isiZulu linguistic terms with the goal of compiling dictionaries. Using both manual methods and the KeyWords function of the WST, Khumalo successfully extracted terms typical for the isiZulu linguistic domain. Terms identified by the tool were thereafter, manually verified. Khumalo (2015) outlined a term extraction process from raw corpora, without mentioning any efforts towards lemmatisation or morphological analysis to aid in term extraction. Notably, his article was published at a time when morphological analysers and lemmatisers for isiZulu were already accessible.

Ndhlovu (2014) investigated translation strategies for health terms from English to Zimbabwean isiNdebele using the parallel concordance tool, ParaConc. The English Ndebele Parallel Corpus (ENPC) was analysed to identify source terms and their equivalent translations in Zimbabwean Ndebele. ParaConc generated various data, including frequencies and potential translations, showing the tool's success. However, the researcher did not specify whether she manually verified the outcomes produced by the machine.

In a University of KwaZulu Natal (UKZN) project, a total of 1,863 terms was collected and subsequently deposited in the isiZulu term bank (Khumalo 2015:

495-499). WST (version 6), primarily utilising the KeyWords function was applied for doing searches on full words. The isiZulu National Corpus grew significantly, and later Sketch Engine was used to create, manage and analyse the corpus (Khumalo 2018).

Mawonga et al. (2014: 66-68) report on the role of the South African-Norwegian Higher Education Development (SANTED) project in the development of African languages in various higher education institutions. This project focused on promoting multilingualism and developing indigenous South African languages and involved collaboration between institutions such as Rhodes University (RU), the University of KwaZulu-Natal (UKZN), and Durban University of Technology (DUT). UKZN and DUT developed an English–isiZulu term list and glossary for fields like education, nursing, and psychology. RU created multilingual resources across various disciplines. The project manually extracted and developed 1,400 terms, which were made available to nurses and midwives, enhancing language use in professional fields (Engelbrecht et al. 2010: 249-267).

The Special Language Corpora for African Languages (SPeLCAL) project aimed to develop linguistic resources for South Africa's nine official African languages, focusing on technical dictionaries, glossaries, and research in areas like terminology and translation. The project compiled written texts from fields such as politics, health, education, law, and technology. SPeLCAL also supported Second Language Teaching (SLT). The English–Venda Parallel Corpus pilot project used Multiconcord software for corpus extraction and analysis. Although it is not explicitly mentioned that terms identified by the software were manually verified, Madiba (2004) emphasises that in this English–Venda parallel corpus pilot project, it was discovered that small corpora allow early human intervention (Madiba 2004: 141, 146).

In summary, the above review of related work reveals varying approaches to term extraction, some relying solely on manual methods while others combine manual and tool-assisted approaches. None of the studies and projects mentioned the use of lemmatised or morphologically decomposed words, suggesting that only complete words were analysed. This review highlights the potential and challenges of semi-automatic term extraction in African languages, underscoring the importance of human judgment alongside computational tools.

3. Resources

In this section the resources employed in this study, namely written corpora and WST as an integrated suite of programs used for corpus analyses, are discussed.

3.1 Corpora

Two types of corpora were collected for this study namely, monolingual, written corpora for general purpose and monolingual, written corpora for language for specific purpose (health). Both corpora were collected from *Vuk'uzenzele*[2] news-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

papers and the South African Centre for Digital Language Resources (SADiLaR) repository. The *Vuk'uzenzele* newspapers can be accessed on the Government Communications Information Systems (GCIS) website. Published monthly, this paper is freely accessible and serves to keep South African citizens informed about government initiatives and services. It deals with a variety of topics with special focus on health, education, safety and security and rural developments. The topics are written in English and all other African languages including, isiNdebele. SADiLaR on the other hand, focuses on research and development in the domains of language-related studies and language technologies in the humanities and social sciences, for all of South Africa's official languages. Access to the repository is open to anyone interested in language technologies. Resources in the repository are categorised into two groups: downloadable and non-downloadable resources. Downloadable resources encompass various formats, such as sound recordings in MP3, portable document format (PDF) documents, plain text (.txt) files, and Microsoft (MS) Word documents. These formats are presented as they were received by SADiLaR. Individuals with research data in the fields of humanities, social sciences, and languages may submit their data to the SADiLaR repository. Consequently, medical corpora collected by Malele (2021), focusing on the use of corpora in compiling an English–isiNdebele glossary of medical terms, as well as the glossary created during this project, are now accessible on the SADiLaR repository[3].

The reference corpus (RC) used in this study is a language for general purpose corpus with 147 417 running words. The size of the RC ensures that a wide variety of subjects is covered, and that their content is diverse. The analysis corpus on the other hand is a domain-specific corpus with 99 052 running words. This corpus comprises of a variety of health topics, including TB, HIV, Primary Health Care, to mention but a few.

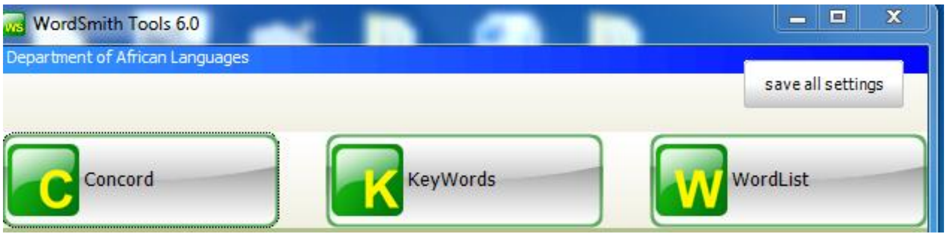


Figure 1: WordSmith Tools

3.2 WordSmith Tools

In this study, the WST tool (version 6.0) was employed. When reviewing literature where the same tool was used, the aim was to establish, among other aspects, the number of functions of the WST that various scholars had used. For instance, Nkomo and Madiba (2011) utilised only the WordList function and concordances of the WST, while Khumalo (2015) employed only the KeyWords function. Taljard and De Schryver (2002) made use of all the WST functions to extract and analyse Sesotho sa Leboa corpora. Similarly, the present study used all the functions of the WST to extract and analyse data from isiNdebele health corpora.

4. Methodology

The current research is rooted in the success of tried and tested methodologies employed in past projects and aims to refine and expand upon these proven approaches to further lexicographic research in African languages. Term extraction was conducted using a semi-automatic approach, combining both manual methods and the WST tool. The process involved utilising the tool to identify terms, which were then manually verified for accuracy. The following steps were followed:

Step 1: Collection of health and general texts: As mentioned earlier, both the isiNdebele health (monolingual, written and specialised) and general texts were collected from two sources namely, *Vuk'uzenzele* newspapers found on the website of the GCIS and also from SADiLaR's repository. All the texts (health and general) from *Vuk'uzenzele* newspapers were PDFs. From the SADiLaR platform, only health texts were collected. The health texts from SADiLaR were in two formats namely, PDF and plain text (.txt) format.

Step 2: Text conversion: All the texts were electronically collected and thereafter, text conversion for texts in PDF took place. The process of text conversion took place in two forms. Texts were firstly converted into MS Word. The purpose of this was to remove graphs, tables, and pictures before the process of loading data on WST could begin. After texts were converted into MS Word, the second step was to further convert the text into .txt format.

Step 3: Cleaning of the plain text format: The plain texts required manual cleaning, a task made easier by the .txt format. Cleaning is important as it effectively removes linguistic 'noise' which can arise from variations in grammar, structure, style, and clearly incorrect spelling in the use of basic language.

Step 4: Loading of corpus files on WST: After the converted texts were cleaned, the corpus files were then loaded on the WST for automatic term extraction. As illustrated in Figure 1, the WST has the following functions: Concord, KeyWords, and WordList all of which were used in identifying and analysing health terms from the given corpora.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

5. Discussion

For this study, the available NCHLT Lemmatiser (2018) as well as the NCHLT Morphological Decomposer (2018) tools for isiNdebele as described by Eiselen and Puttkammer (2014) were applied to the relevant corpora. The aim was to execute term extraction based on lemmas and not merely on full words. However, it was found that both these NCHLT tools for isiNdebele are not suitable for this purpose due to the unreliable quality of the output, with the result that raw, unlemmatised corpora had to be used. Table 1 and Table 2 represent an excerpt of the experiment with the NCHLT tools.

Word	**Lemma**	**Comment**	**Expected Lemmatisation**
abodorhodere	abodorhodere	x	dorhodere
babodorhodere	babodorhodere	x	dorhodere
bodorhodere	bodorhodere	x	dorhodere
nabodorhodere	nabodorhodere	x	dorhodere
njengodorhodere	njengodorhodere	x	dorhodere
nodorhodere	nodorhodere	x	dorhodere
udorhodere	dorhodere	Lemma correctly identified	
yobudorhodere	yobudorhodere	x	dorhodere

Table 1: Results of lemmatisation with NCHLT Lemmatiser

An excerpt of the output of the NCHLT Lemmatiser (2018) in Table 2 indicates inconsistences in the sense that only the basic, singular word *udorhodere* 'doctor' is lemmatised correctly, whereas the plural form *abodorhodere* 'doctors' and other forms with possessive and adverbial prefixes are not lemmatised at all. Similarly, the accuracy of the NCHLT Morphological Decomposer (2018) output appears flawed, as demonstrated by the examples in Table 3. While the basic singular word *udorhodere* 'doctor' is correctly decomposed, the plural form *abodorhodere* 'doctors' is incorrectly decomposed, and all other forms with possessive and adverbial prefixes remain undecomposed.

Word	Decomposition	Comment	Expected Decomposition
abodorhodere	a-bodorhodere	x	abo-dorhodere
babodorhodere	babodorhodere	x	ba-bo-dorhodere
bodorhodere	bodorhodere	x	bo-dorhodere
nabodorhodere	nabodorhodere	x	na-bo-dorhodere
njengodorhodere	njengodorhodere	x	njenga-u-dorhodere
nodorhodere	nodorhodere	x	na-udorhodere
udorhodere	u-dorhodere	Morphological decomposition correct	
yobudorhodere	yobudorhodere	x	yo-bu-dorhodere

Table 2: Results of morphological decomposition with NCHLT Morphological Decomposer

Accurate lemmatisation or morphological decomposition for isiNdebele with its conjunctive orthography, could have impacted the frequency of word counts and reduced much of the manual work required for term identification. For example, the keyword *dorhodere* 'doctor' appears with a frequency of 75 in the Frequency Word List (see Table 3). If the lemma *dorhodere* had been correctly identified in the other words listed in Tables 1 and 2 using a lemmatiser or morphological decomposer, this lemma could have replaced the unlemmatised forms in the Frequency Word List, thereby significantly increasing the frequency of *dorhodere* to over 250.

a. WordList / Frequency Ranked List function

To begin, it was necessary to create word lists of health terms. The text/file icon was chosen, followed by the 'Make a word list now' icon. The output came in three different formats, namely the frequency-ranked word list, the statistical analysis and an alphabetically ordered word list.

The frequency WordList function was the first function to be used. It is worth noting that the tool was provided with the unlemmatised form of texts. All functions of the WST were utilised for this article, as they complement each other. Terms possibly overlooked by one function could be picked by another. The frequency WordList was used to extract synonyms and terms with variant forms. Here are the identified terms with variant forms:

*ingogwan**a*** vs *ingogwan**i*** 'virus'
*udorhoder**e*** vs *udorhoder**a*** 'doctor'
*umulwan**i*** vs *umulwan**a*** 'germ'

Table 3 represents an excerpt from the Frequency Word List containing unlemmatised health terms.

N	**Word**	**Frequency**	**%**	**Texts**	**%**
90	ubulwele	86,00	0,09	1,00	100,00
91	womnyaka	86,00	0,09	1,00	100,00
92	endaweni	85,00	0,09	1,00	100,00
93	imithetho	85,00	0,09	1,00	100,00
94	ihlelo	84,00	0,08	1,00	100,00
95	njengombana	84,00	0,08	1,00	100,00
96	sakho	84,00	0,08	1,00	100,00
97	ukusebenza	83,00	0,08	1,00	100,00
98	weengazi	83,00	0,08	1,00	100,00
99	izinga	82,00	0, 08	1,00	100,00
100	ukuze	82,00	0,08	1,00	100,00
101	zoke	82,00	0,08	1,00	100,00
102	angeze	80,00	0,08	1,00	100,00
103	nofana	80,00	0,08	1,00	100,00
104	of	80,00	0,08	1,00	100,00
105	lezamaphilo	79,00	0, 08	1,00	100,00
106	tb	79,00	0,08	1,00	100,00
107	inomboro	78,00	0,08	1,00	100,00
108	nezokuphepha	78,00	0,08	1,00	100,00
109	nje	78,00	0,08	1,00	100,00
110	ubujamo	78,00	0, 08	1,00	100,00
111	kanti	77,00	0,08	1,00	100,00

112	iindleko	76,00	0,08	1,00	100,00
113	dorhodere	75,00	0,08	1,00	100,00
114	esibhedlela	75,00	0, 08	1,00	100,00
115	njalo	75,00	0,08	1,00	100,00
116	abasebenzi	74,00	0,07	1,00	100,00

Table 3: Frequency WordList

Table 3 illustrates that health terms are few as compared to non-health terms, for instance ranked number 90 is the term *ubulwele* 'disease', with 86 occurrences. Ranked number 105 is the term *lezamaphilo* 'of health' with 79 occurrences. Ranked number 106 is the term *tb* 'tb' with 79 occurrences. Ranked number 113 is the term *dorhodere* 'doctor' with 75 occurrences, and ranked number 114 is the term *esibhedlela* 'at the hospital' also with 75 occurrences. The most frequent words in the analysis corpus are function or grammatical words such as *njengombana* 'as it is', *nofana* 'or', *kanti* 'whereas', *ukuze* 'so that', *nje* 'now' or 'this way', *njalo* 'always'. Function words in isiNdebele include word classes such as adverbs, conjunctions and pronouns which serve a grammatical purpose in a sentence, but typically carry little lexical meaning. Unlike function words, content words hold lexical significance and typically represent tangible or intangible entities, actions, attributes, or concepts. In isiNdebele, content words include nouns, verbs and adjectives. Table 3 comprises mainly nouns (such as *imithetho* 'rules', *ihlelo* 'plan', *abasebenzi* 'workers', *iindleko* 'costs', etc.), but also (auxiliary) verbs (e.g., *angeze* 'he can/may/might not come'). This serves as evidence that function words or grammatical words dominate frequency lists. To preserve content words, function words can be excluded from the word list.

b. KeyWords function

To make the key word list, both the reference (general corpus) and analysis (health) corpus files were uploaded on the KeyWords function of the WordSmith Tools. The non-language specific WST was therefore relied on with the KeyWords function being chosen followed by 'Make a keyword list now'. A list of key words was then produced. The purpose of using the KeyWords function was to extract all term candidates from the health corpus. It was used to calculate words which are key in a text, that is, words used much more frequently or much less frequently in each corpus. Through this function, terms used in the analysis (health) corpus were identified. Here the frequency of each word in the word list of health corpus was compared with the frequency of the same word in the reference word list. The output was a list of key words, or words whose frequencies are higher in the analysis corpus than in the RC. Any word which is found to be most outstanding in its frequency in the text is then considered to be 'key'. Key words are presented in their order of the most outstanding word. The KeyWords function provided the term candidate list as illustrated in Table 4.

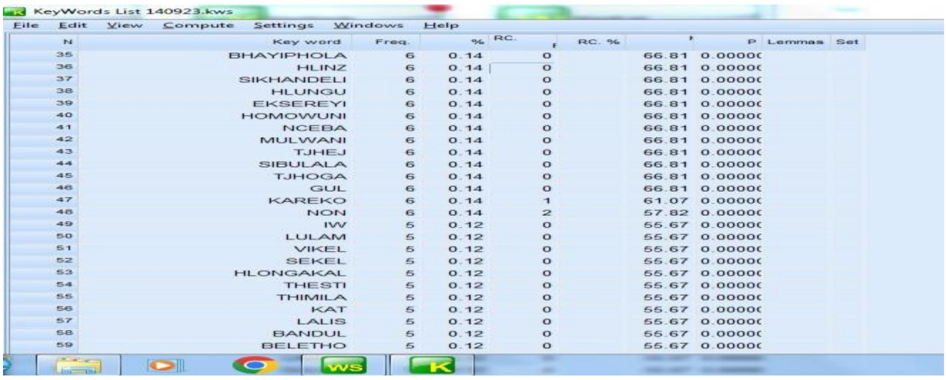
N	Key word	Freq.	%	RC.Freq.	RC. %	Keyness
1	mrhatjhi[4]	281,00	0,28	0,00		512,79
2	dorh	192,00	0,19	0,00		350,27
3	gems	188,00	0,83	0,00		342,97
4	begodu	822,00	0,45	498,00	0,34	262,45
5	khulu	448,00	0,16	187,00	0,13	239,87
6	HIV	154,00	0,19	13,00		202,92
7	ngamunye	184,00		28,00	0,02	198,92
8	amatshwayo	131,00	0,13	6,00		195,83
9	ingabe	169,00	0,17	26,00	0,02	181,84
10	umndeni	131,00	0,13	10,00		177,01
11	kobana	426,00	0,43	220,00	0,15	174,60
12	angabe	252,00	0,25	81,00	0,05	173,45
13	sista	93,00	0,09	0,00		169,61
14	beemali	122	0,12	12,00		154,04
15	umzimba	96,00	0,10	3,00		151,28
16	tb	79,00	0,08	0,00		144,07
17	womnyaka	86,00	0,09	2,00		139,80
18	dorhodere	75,00	0,08	0,00		136,77
19	lezamaphilo	79,00	0,08	1,00		134,34

20	ukudla	165,00	0,17	43,00	0,03	134,34
21	nezokuphepha	78,00	0,08	1,00		132,55
22	ngomkhawulo	70,00	0,07	0,00		127,65
23	esibhedlela	75,00	0,08	1,00		127,15
24	ubulwele	86,00	0,09	5,00		123,24
25	weengazi	83,00	0,08	4,00		123,02
26	udorhodere	72,00	0,07	1,00		121,76
27	iingazi	65,00	0,07	0,00		118,53
28	isana	69,00	0,07	1,00		116,37
29	emzimbeni	67,00	0,07	1,00		112,78
30	wesibhedlela	61,00	0,06	0,00		111,24

Table 4: First 30 words of the resultant key word list

The focus is on the health terms that appear in this first 30 list. Ranked number 2 is the word *dorh* 'doc' which is an abbreviation for *udorhodere* 'doctor'. Its frequency is 192 in the health corpus, and it is, zero (0.00) in the RC. Words which are key are 350.27. Ranked number 6 is the word HIV with the frequency of 154 in the health corpus, and the frequency of 13.00 in the RC. Its keyness is 202.92. Ranked number 19 is the word *lezamaphilo* 'of health' with the frequency of 79 in the health corpus, and the frequency of 1 in the RC. Its keyness is 134.34. Ranked number 24 is the word *ubulwelwe* 'disease' with the frequency of 86 in the health corpus and the frequency of 5 in the RC. It is 123.24 in keyness. Ranked number 25 is the word *weengazi* 'of the blood', with 83 frequencies in the health corpus and the frequency of 4 in the RC. It is 123.02 in keyness.

When considering the key words in Table 4, it is evident that most health terms, from this study's corpus have zero frequencies in the RC. This is evident in the case *of dorh* 'dr', *sista* 'sister', *TB* 'TB', *dorhodere* 'doctor', *iingazi* 'blood', *wesibhedlela* 'of the hospital'. It is important to note that the KeyWords function only managed to identify 138 health terms. The main challenge here is multi-words, that is strings of words (two or more words) that are considered to be one lexical unit. The KeyWords function neither identifies nor extracts multi-words. A multi-word such as *ikankere yomlomo wesibeletho* 'cervical cancer' appears as three separate words, with different ranks and frequencies. Therefore, only the word *ikankere* derived from the Afrikaans 'kanker' will appear as the health term. The word *umlomo* will just be translated as 'mouth' and *wesibeletho* will be translated as 'of the womb'. This means that the two words (*umlomo* and *wesibeletho*) will be rendered as non-health terms. Following the KeyWords function's identification of only 138 terms, we manually reviewed the word list of health terms and collected 582 health terms in total. This indicates that the KeyWords function overlooked many terms. Figure 2 clearly demonstrates the challenges associated with identifying multi-word terms.



N	Key word	Freq.	%	RC.	RC. %		P	Lemmas	Set
35	BHAYIPHOLA	6	0.14	0		66.81	0.0000(		
36	HLINZ	6	0.14	0		66.81	0.0000(		
37	SIKHANDELI	6	0.14	0		66.81	0.0000(		
38	HLUNGU	6	0.14	0		66.81	0.0000(		
39	EKSEREYI	6	0.14	0		66.81	0.0000(		
40	HOMOWUNI	6	0.14	0		66.81	0.0000(		
41	NCEBA	6	0.14	0		66.81	0.0000(		
42	MULWANI	6	0.14	0		66.81	0.0000(		
43	TJHEJ	6	0.14	0		66.81	0.0000(		
44	SIBULALA	6	0.14	0		66.81	0.0000(		
45	TJHOGA	6	0.14	0		66.81	0.0000(		
46	GUL	6	0.14	0		66.81	0.0000(		
47	KAREKO	6	0.14	1		61.07	0.0000(		
48	NON	6	0.14	2		57.82	0.0000(		
49	IW	5	0.12	0		55.67	0.0000(		
50	LULAM	5	0.12	0		55.67	0.0000(		
51	VIKEL	5	0.12	0		55.67	0.0000(		
52	SEKEL	5	0.12	0		55.67	0.0000(		
53	HLONGAKAL	5	0.12	0		55.67	0.0000(		
54	THESTI	5	0.12	0		55.67	0.0000(		
55	THIMILA	5	0.12	0		55.67	0.0000(		
56	KAT	5	0.12	0		55.67	0.0000(		
57	LALIS	5	0.12	0		55.67	0.0000(		
58	BANDUL	5	0.12	0		55.67	0.0000(		
59	BELETHO	5	0.12	0		55.67	0.0000(		

Figure 2: Resulting KeyWords function on multi-words

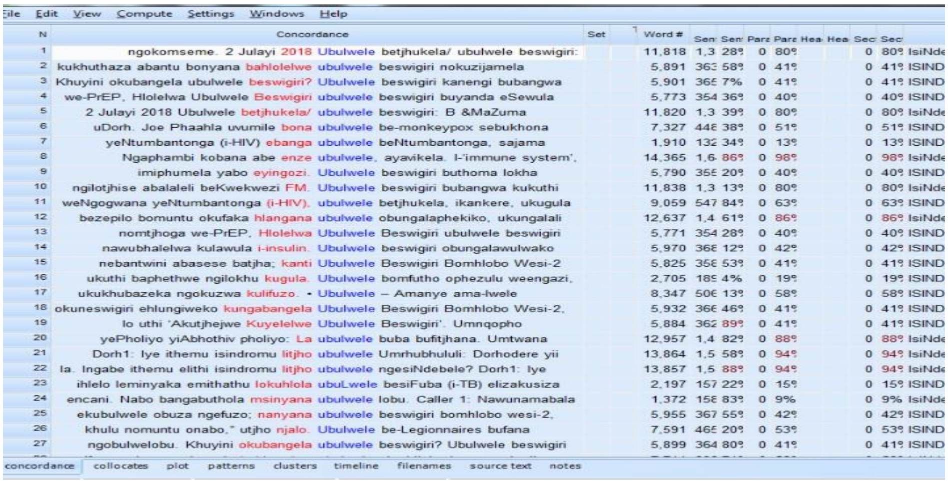
Figure 2 reflects the first version of the semi-automatically found term candidates. According to this figure, multi-words are not identified and extracted by the KeyWords function. Owing to the fact that the KeyWords function cannot sort the variants of the term candidate list, neither can it identify the multi-word term candidates, terms that resulted from the KeyWords function had to be manually validated. The researchers therefore had to manually sort certain terms and also manually identify multi-words. For instance, term number 37, *sikhandeli* loosely translates as 'preventer'. The full or complete word could be either *sikhandeli-magciwana* which loosely translates as 'preventer of germs', with 'antibiotic' as the correct health term, or *isikhandeli kuvuvuka* loosely translated as 'preventer of swollenness', with 'anti-inflammatory' being the correct health term. Term number 44 is also a multi-word of which the complete multi-word is actually *sibulala magciwana* loosely translated as 'the killer of germs'. The correct health term is 'antibiotic'. Figure 2 confirms that the KeyWords function indeed struggles to identify multi-word terms. Following this, we discuss the Concord function.

c. Concord function

To make a concordance, we started by choosing a text file. After a text file was

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

chosen, search words were entered. Figure 3 shows the concordance for the search word *ubulwele* 'disease'.



N	Concordance	Set	Word #	Sen	Sen	Para	Para	Hea	Hea	Sec	Sec	
1	ngokomseme. 2 Julayi 2018 Ubulwele betjhukela/ ubulwele beswigiri:		11,818	1,3	28%	0	80%			0	80%	IsiNde
2	kukhuthaza abantu bonyana bahlolelwe ubulwele beswigiri nokuzijamela		5,891	363	58%	0	41%			0	41%	ISIND
3	Khuyini okubangela ubulwele beswigiri? Ubulwele beswigiri kanengi bubangwa		5,901	365	7%	0	41%			0	41%	ISIND
4	we-PrEP, Hlolelwa Ubulwele Beswigiri ubulwele beswigiri buyanda eSewula		5,773	354	36%	0	40%			0	40%	ISIND
5	2 Julayi 2018 Ubulwele betjhukela/ ubulwele beswigiri: B &MaZuma		11,820	1,3	39%	0	80%			0	80%	IsiNde
6	uDorh. Joe Phaahla uvumile bona ubulwele be-monkeypox sebukhona		7,327	448	38%	0	51%			0	51%	ISIND
7	yeNtumbantonga (i-HIV) ebanga ubulwele beNtumbantonga, sajama		1,910	132	34%	0	13%			0	13%	ISIND
8	Ngaphambi kobana abe enze ubulwele, ayavikela. I-'immune system',		14,365	1,6	86%	0	98%			0	98%	IsiNde
9	imiphumela yabo eyingozi. Ubulwele beswigiri buthoma lokha		5,790	355	20%	0	40%			0	40%	ISIND
10	ngilotjhise abalaleli beKwekwezi FM. Ubulwele beswigiri bubangwa kukuthi		11,838	1,3	13%	0	80%			0	80%	IsiNde
11	weNgogwana yeNtumbantonga (i-HIV), ubulwele betjhukela, ikankere, ukugula		9,059	547	84%	0	63%			0	63%	ISIND
12	bezepilo bomuntu okufaka hlangana ubulwele obungalaphekiko, ukungalali		12,637	1,4	61%	0	86%			0	86%	IsiNde
13	nomtjhoga we-PrEP, Hlolelwa Ubulwele Beswigiri ubulwele beswigiri		5,771	354	28%	0	40%			0	40%	ISIND
14	nawubhalelwa kulawula i-insulin. Ubulwele beswigiri obungalawulwako		5,970	368	12%	0	42%			0	42%	ISIND
15	nebantwini abasese batjha; kanti Ubulwele Beswigiri Bomhlobo Wesi-2		5,825	358	53%	0	41%			0	41%	ISIND
16	ukuthi baphethwe ngilokhu kugula. Ubulwele bomfutho ophezulu weengazi,		2,705	189	4%	0	19%			0	19%	ISIND
17	ukukhubazeka ngokuzwa kulifuzo. • Ubulwele – Amanye ama-lwele		8,347	506	13%	0	58%			0	58%	ISIND
18	okuneswigiri ehlungiweko kungabangela Ubulwele Beswigiri Bomhlobo Wesi-2,		5,932	366	46%	0	41%			0	41%	ISIND
19	lo uthi 'Akutjhejwe Kuyelelwe Ubulwele Beswigiri'. Umnqopho		5,884	362	89%	0	41%			0	41%	ISIND
20	yePholiyo yiAbhothiv pholiyo: La ubulwele buba bufitjhana. Umtwana		12,957	1,4	82%	0	88%			0	88%	IsiNde
21	Dorh1: Iye ithemu isindromu litjho ubulwele Umrhubhululi: Dorhodere yii		13,864	1,5	58%	0	94%			0	94%	IsiNde
22	la. Ingabe ithemu elithi isindromu litjho ubulwele ngesiNdebele? Dorh1: Iye		13,857	1,5	88%	0	94%			0	94%	IsiNde
23	ihlelo leminyaka emithathu lokuhlola ubuLwele besiFuba (i-TB) elizakusiza		2,197	157	22%	0	15%			0	15%	ISIND
24	encani. Nabo bangabuthola msinyana ubulwele lobu. Caller 1: Nawunamabala		1,372	158	83%	0	9%			0	9%	IsiNde
25	ekubulwele obuza ngefuzo; nanyana ubulwele beswigiri bomhlobo wesi-2,		5,955	367	55%	0	42%			0	42%	ISIND
26	khulu nomuntu onabo," utjho njalo. Ubulwele be-Legionnaires bufana		7,591	465	20%	0	53%			0	53%	ISIND
27	ngobulwelobu. Khuyini okubangela ubulwele beswigiri? Ubulwele beswigiri		5,899	364	80%	0	41%			0	41%	ISIND

Figure 3: Concord on search word *ubulwele* "disease"

Figure 3 clearly illustrates the concord function, and the role it played in deriving more terms. See the following lines:

Line 1 *ubulwele betjhukela*	'disease of sugar; sugar diabetes'
Line 2 *bahlolelwe ubulwele*	'they were tested for a disease'
Line 3 *okubangela ubulwele*	'causes the disease'
Line 6 *ubulwele be monkey-pox*	'disease of monkey-pox; Monkey-pox'
Line 7 *ubulwele bentumbantonga*	'disease of AIDS; AIDS'
Line 12 *ubulwele obungalaphekiko*	'incurable disease'
Line 13 *ubulwele beswigiri*	'disease of sugar; sugar diabetes'
Line 16 *ubulwele bomfutho ophezulu*	'disease of high blood pressure; high blood pressure'

Through the Concord function, term identification is streamlined, leading to the automatic discovery of more terms. For example, searching for the word *ubulwele* 'disease' yielded various types of diseases. In line 1 *ubulwele betjhukela* 'sugar diabetes', line 6, *ubulwele be-monkey pox* 'monkey-pox', line 7 *ubulwele bentumbantonga* 'AIDS', line 16, *ubulwele bomfutho ophezulu* 'high blood pressure'. All these disease types were identified through the use of the Concord search.

Utilising the Concord function has contributed to a deeper understanding of the usage of terms in context. It aided researchers in examining words within their textual context, facilitating the identification of patterns of similarity or contrast in the words surrounding the search term. For instance, the search term *ubulwele* 'disease' is frequently followed by the type of disease, as observed. More-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

over, there appears to be a consistent pattern of words preceding the search term. In this instance, *ubulwele* 'disease' is preceded by *bahlolelwe* 'they were tested for' in line 2 *ebanga* 'that causes' in line 7. This suggests an association between the term 'disease' and terms such as 'tests', 'causes', and so forth.

Based on the given examples, the search term and its co-text are arranged so that the textual environment can be assessed and patterns surrounding the search term can be identified visually. Moreover, exploring concordances enables users to observe corpus occurrences, understand how meaning is constructed in texts, observe word co-occurrences, and recognise how they form meaningful patterns, without imposing predetermined notions on these units. As depicted in Figure 3 above, concordance analysis leads to the discovery of additional terms, simplifying the manual identification process. It also helps with collocates that is, the company that the key word keeps. Bowker and Pearson (2002: 124) state that collocates are words which typically occur in the vicinity of your search pattern. Collocates play an important role in corpus linguistics; they make it easier for the researcher or learner to understand the usage of words. Collocates further assist with the understanding of how two words come together meaningfully. Some concordances offer an additional facility which frequently ranks the words that appear in the vicinity of the search pattern.

Another significant aspect of the Concord function is its capability to identify multi-word expressions, a feature lacking in the Frequency WordList and KeyWords functions. Therefore, the Concord function produced the most effective results in terms of identifying both single and multi-word expressions. This is evident in the following lines:

Line 1: *ubulwele betjhukela/ubulwele beswigiri* 'diabetes' (loosely translated as 'disease of sugar')
Line 6: *ubulwele be-monkey pox* 'monkey pox' (loosely translated as 'disease of monkey pox')
Line 23: *ubulwele besifuba* 'TB' (loosely translated as 'disease of the chest')

The frequency WordList and KeyWords function would only identify *ubulwele* 'disease', *monkey-pox* 'monkey pox', *betjhukela* 'of sugar' and *besifuba* 'of the chest'.

In this section, the identification, extraction, and analyses of terms are discussed using the WST alongside manual methods. All the functions of the WST are fully utilised with manual verification of outcomes. Among these functions, the Concord function stands out for extracting more terms, including multi-word expressions not captured by the WordList and KeyWords functions.

Conclusion

Semi-automated term extraction in African languages yields valuable outcomes, particularly in addressing the deficiency of specialised terminology in isiNdebele. Terms extracted from specialised corpora or language for specific purpose (LSP) corpora, serve as an essential foundation for the compilation of subject field dictionaries by providing accurate and contextually relevant terms. This study has demonstrated the usefulness of semi-automatic term extraction in contributing

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

to this foundation. While using software like WST in the processing of raw corpora to extract and identify single-word and multi-word terms, the bulk of terms required manual identification. However, manual identification does not negate the corpus-driven nature of the approach. Corpus-driven term extraction proves pivotal in mitigating terminology shortages, even in resource-scarce languages like isiNdebele.

The utilisation of WST in extracting and analysing data from available corpora underscores its potential applicability across various fields of study such as law, economics, and religion. It is apparent from this study that employing all functions of WST is crucial as these functions complement each other. Terms overlooked by one function may be captured by another. The Concord function, particularly adept at identifying multi-word terms, proves most fruitful compared to Frequency WordList and KeyWords functions.

Health terms identified in this study are intended for communication with laypersons rather than between experts in the health field. Consequently, given the scarcity of resources in the field of isiNdebele dictionaries, especially the lack of dictionaries for specific purposes, this range of health terms represents typical candidates for incorporation in a general dictionary. Terminology is vital in lexicography, ensuring reliable interpretation across dictionary entries. Additionally, as language and society evolve, terminology reflects these changes, influencing potential incorporation of new words. Therefore, careful consideration of terminology is crucial for maintaining relevant and useful lexicographical works.

The findings of this research hold significance for isiNdebele translators, lexicography students, educators, and linguists, offering insights into the role of technology in terminology resource development. Moreover, it contributes to the standardisation of health terms inconsistently used for communication in healthcare institutions.

Presently, due to the scarcity of isiNdebele corpora containing health-related terms, a broader range of general health terms had to be compiled. Future research endeavours will need to focus on utilising larger corpora to gather terminology in additional specialised healthcare fields.

Acknowledgements

The insights shared by Gertrud Faass during the co-supervision of the original doctoral research, the assistance provided by Marissa Griesel with the NCHLT Lemmatiser and Morphological Decomposer tools, and the constructive comments from two anonymous reviewers are gratefully acknowledged.

Endnotes

1. Southern Ndebele should be differentiated from Northern Ndebele (ISO 639-3: nde), spoken in Zimbabwe (cf. https://iso639-3.sil.org/)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

2. https://www.vukuzenzele.gov.za
3. https://repo.sadilar.org/handle/20.500.12185/272
4. Although the first word with the highest frequency in Table 4 is a non-health term, namely *umrhatjhi* 'a radio DJ', it will be ignored for the purposes of this study because the relevant corpora were sourced from literacised radio interviews that all frequently included this term. This word has a frequency of 281.00 in the health corpus, and no frequency in the RC.

References

Bowker, L. and J. Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora.* London: Routledge.

Eiselen, R. and M. Puttkammer. 2014. Developing Text Resources for Ten South African Languages. Calzolari, N. et al. (Eds.). 2014. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC '14), Reykjavik, Iceland, May 26–31, 2014:* 3698-3703. Reykjavik, Iceland: European Language Resources Association (ELRA).

Engelbrecht, C., N.C. Shangase, S.J. Majeke, S.Z. Mthembu and Z.M. Zondi. 2010. IsiZulu Terminology Development in Nursing and Midwifery. *Alternation* 17(1): 249-272.

Finlayson, R. and M. Madiba. 2002. The Intellectualisation of the Indigenous Languages of South Africa: Challenges and Prospects. *Current Issues in Language Planning* 3(1): 40-61.

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: SUN PReSS.

Jiyane, D.M. 1994. *Aspects of isiNdebele Grammar*. Unpublished M.A. Dissertation. Pretoria: University of Pretoria.

Khumalo, L. 2015. Semi-automatic Term Extraction for an isiZulu Linguistic Terms Dictionary. *Lexikos* 25: 495-506.

Khumalo, L. 2018. Towards an isiZulu National Corpus. Du Plessis, A.H. and S.E. Bosch (Eds.). 2018. *African Association for Lexicography, 23rd International Conference, June 27–29, 2018, University of the Western Cape, Cape Town, South Africa: Abstracts and Programme:* 26-29. Cape Town: AFRILEX.

Madiba, M. 2004. Parallel Corpora as Tools for Developing the Indigenous Languages of South Africa with Special Reference to Venda. *Language Matters* 35(1): 133-147.

Malele, N.J. 2021. *The Use of Corpora in the Compilation of a Specialised English–isiNdebele Glossary of Medical Terms.* Unpublished D.Litt. et Phil. Thesis. Pretoria: University of South Africa.

Mawonga, S., P. Maseko and D. Nkomo. 2014. The Centrality of Translation in the Development of African Languages for Use in South African Higher Education Institutions: A Case Study of a Political Science English–isiXhosa Glossary in a South African University. *Alternation Special Edition* 13: 55-79.

NCHLT isiNdebele Lemmatiser. 2018.
https://repo.sadilar.org/handle/20.500.12185/303 [15 April 2024].

NCHLT isiNdebele Morphological Decomposer. 2018.
https://repo.sadilar.org/handle/20.500.12185/304 [15 April 2024].

Ndhlovu, K. 2014. Term-creation Strategies Used by Ndebele Translators in Zimbabwe in the Health Sector: A Corpus-based Approach. *Stellenbosch Papers in Linguistics Plus* 43: 327-344.

Nkomo, D. and M. Madiba. 2011. The Compilation of Multilingual Concept Literacy Glossaries at the University of Cape Town: A Lexicographical Function Theoretical Approach**.** *Lexikos* 21: 144-168.

Prinsloo, D.J. 2015. Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus. *Lexikos* 25: 285-300.

SADiLaR. 2024. *South African Centre for Digital Language Resources.* https://repo.sadilar.org/handle/20.500.12185/272/ [25 March 2024].

Scott, M. 2010. WordSmith Tools (Version 5.0). [Computer software]. Liverpool: Lexical Analysis Software.

Taljard, E. and S.E. Bosch. 2006. A Comparison of Approaches to Word Class Tagging: Disjunctively vs. Conjunctively Written Bantu Languages. *Nordic Journal of African Studies* 15(4): 428-442.

Taljard, E. and G.-M. de Schryver. 2002. Semi-automatic Term Extraction for the African Languages, with Special Reference to Northern Sotho. *Lexikos* 12: 44-74.

Vuk'uzenzele newspaper English–isiNdebele. Government Communication & Information System (GCIS). https://www.vukuzenzele.gov.za [25 March 2024].