Using Corpora in Online isiZulu Language Teaching

Langa Khumalo
ORCID iD: https://orcid.org/0000-0002-2694-9105

Abstract

The impact of the Covid-19 global pandemic on higher education in South Africa has inspired the academy to adopt new pedagogies in the teaching of African languages. This chapter shows that through the use of an isiZulu corpus, African language courses can be offered online, using digital humanities methodologies such as the AntConc concordance program. African languages are resource-scarce languages (Bosch *et al.* 2007; Pretorius & Bosch 2003; Keet & Khumalo 2014). This scarcity includes the paucity of exhaustive grammatical descriptions, the compilation of both large and specialized corpus resources, and the development of machine-readable lexicons. A corpus is thus carefully designed and systematically collected natural language data from a variety of text types and sources following a particular set of principles, which constitute a sample that statistically reflects the use of that particular language, and is processed, stored and accessed by means of computers. This chapter argues for a novel way of teaching African languages, particularly isiZulu, using corpora and lexical software as open-source online resources. UKZN has developed the isiZulu national corpus (INC) to be the biggest African language corpus, as well as two other corpus typologies, the English-isiZulu Parallel Corpus (EiPC) and the IsiZulu Oral Corpus (IOC), that are available as digital resources for language research and language teaching. Using AntConc, which is a freeware concordance program for Windows, African languages courses can be offered online in response to the Covid-19 education lockdown.

Keywords: Covid-19, Corpus, human language technology, language, teaching, digital humanities.

DOI: https://doi.org/10.29086/978-0-9869936-4-0/2020/AASBS03
Chapter URL pp. 37 - 53: http://alternation.ukzn.ac.za/Files/books/series-03/02-khumalo.pdf

1 Introduction

The Covid-19 global pandemic, which primarily affects health, has had a telling effect on many sectors, including the higher education sector. The global crisis has resulted in many countries taking unusual measures such as closing schools, colleges and universities in a massive lockdown in order to stem the spread of the novel coronavirus. The higher education sector has been compelled to come up with responsive measures to address the disruption of the academic programmes. Efforts are under way to migrate to online teaching and learning on a hitherto untested and unprecedented scale for most universities in South Africa. Student assessments have provided a rare challenge, with some institutions opting to cancel them altogether. A further challenge is networked communication, which is not available to most students, is very poor, is exorbitant and is disrupted by power interruptions. Other hidden challenges are the complexities of home schooling, with some homes just not suitable for any learning to take place privately.

The migration of teaching and learning to remote learning via online platforms presents a further challenge to the teaching of African languages. Most African languages have no digital presence. This means that they have no processed natural language data that is stored in a reusable format online. They are in this sense viewed as resource-scarce languages. The paucity of resources includes lack of exhaustive linguistic descriptions, the absence of large and specialized corpora, and machine-readable lexicons. As a result, the development of human language technologies and other computational resources has been scarce for most African languages. Crucially, because of the generally diminished status and limited use of these languages, attraction of funding resources is also poor (Bosch *et al*. 2008; Pretorius & Bosch 2003; Keet & Khumalo 2014). This is notably the bane of most African languages. It is our argument that computational solutions and funding resources are an important precursor to the development of African languages. Computational tools in the current context of the Fourth Industrial Revolution (4IR) and Big Data analytics are important enablers in accelerating the introduction and use of African languages in the academy.

The University of KwaZulu-Natal (henceforth UKZN) has made advances in the development of isiZulu as one of its two official languages (UKZN Language Policy 2014). It is notable that isiZulu is the most popular

South African language by first language speakers, with about 25% of a population of about 58 million (www.gov.za). Since 2014, UKZN has developed the isiZulu national corpus (INC) to be the biggest African language corpus with 31.7 million running words, and two other corpus typologies, the English-isiZulu Parallel Corpus (EiPC) and the IsiZulu Oral Corpus (IOC), that are available as digital resources for language research and language teaching. We thus argue in this position paper that using these corpus resources, UKZN can teach isiZulu online in response to the Covid-19 education lockdown.

2 Corpora and Digital Humanities Methods

Digital Humanities methods involve the use of novel computational methods, such as computer software and carefully processed and machine-readable data to solve research problems in the humanities and social sciences or to challenge existing theoretical assumptions. In order to address the disruption that has been brought about as a result of the Coronavirus pandemic, the imperative exists to transform pedagogies at our universities. This entails using technology to improve the student experience through digitizing content and using digital methodologies in teaching and learning. It is in this sense that the advancing scholarship of Digital Humanities comes to the fore. The scholarship of Digital Humanities promotes and advances digital research and teaching across all arts and humanities disciplines using cutting-edge technological resources, providing scholars with new ways of looking at old problems, while simultaneously advancing (new) knowledge and novel pedagogies. Digital Humanities provides a bridge between the traditional practices of research and technology-driven research to scholars straddling quintessential humanist approaches and modern digital methodologies, tools and frameworks to support them in novel avenues of enquiry. It is a growing scholarship in the Humanities and Social Sciences that is spurred by advances in computing and digital spheres and providing new collaboration with engineering and computer (and techno) sciences. African language corpora (**corpus**, *singular*) as digital resources provide African linguists with an opportunity to teach African languages in a novel and technology-driven way. A corpus is carefully designed and systematically collected natural language data from a judiciously selected variety of text types and sources following a

particular set of principles, which constitutes a sample that statistically reflects the use of that particular language, and is processed, stored in a reusable format in order to be accessed by means of computers (Khumalo, *forthcoming*). The size of the corpus and the source from which it is created depend on the intended purpose.

The INC is a Language for General Purpose (LGP) with 31.7 million running words of written text data, currently the largest African language corpus. It is an organic monitor corpus with a sufficient balance in terms of text types and thematic content. The INC is a web-based open-source resource. The URL for the INC is: https://iznc.ukzn.ac.za/.

Table 1. The INC Statistics.

	File name (IsiZulu)	File name (English)	Number of files	Word tokens
1.	Inoveli	IsiZulu novels	487	9 679 532
2.	Isolezwe	Isolezwe newspaper	489	7 289 832
3.	UmAfrika	UmAfrika newspaper	76	5 735 962
4.	Ilanga	Ilanga newspaper	970	4 361 605
5.	Izindaba zabantu	Izindaba zabantu newspaper	43	2 376 468
6.	Ezasegagasini	Metro ezasegagasini newspaper	61	782 377
7.	Ibhayibheli	The Bible	1	435 481
8.	Umthetho	The Hansard	194	367 998
9.	Zulumanuscripts	Dissertations	19	287 684
10.	Umngenelo	Literature competition short stories	16	130 698
11.	Ingede	Ingede newspaper	14	125 972
12.	Uhlelo	Zulu grammar textbooks	2	84 416
13.	Zulusimama	kznonline newsletter	4	34 525
14.	Umthethosisekelo	The Constitution of the Republic	1	32 410
15.	Amanothi	Lecture notes	4	27 447
16.	Zuluplay	The play amaseko	1	16 694
17.	Amantshontsho	A Bible lesson	1	4 974
18.	Inganekwane	Folktales	1	1 709
		Total number of files in the INC	**2 384**	
		Total number of tokens in the INC		**31 775 784**

Table 1 shows the genre of text assorted as filenames, the number of the files of each genre and the tokens, also known as running words.

The INC is one of the three corpus typologies. The second is the isiZulu Oral Corpus (IOC). UKZN embarked on fieldwork across KwaZulu-Natal Province's nine districts, collecting oral speech data of 1 000 hours of digital recordings. This oral corpus is being processed for addition into the INC. The addition of the oral corpus will help to achieve a healthy balance of both written and spoken text types in the INC. The third corpus is a parallel corpus, which is a collection of texts that are translated from one language L_1 (the source language) to the other language L_2 or language L_X (the target language(s)).

In 2017 UKZN initiated a process to create a bilingual parallel corpus of English and isiZulu. The imperative to create the English-IsiZulu Parallel Corpus (EiPC) was inspired by the massive textual production in the two languages, as a result of conformity to the language policy, which stipulates that whenever possible, administrative and academic information must be made available in the University's two official languages. The EiPC currently has 15 503 parallel-sentence data in both English and isiZulu. It is on the basis of the EiPC that a Data-Driven Machine Translation (DDMT) approach will be employed to build a machine translation tool. The DDMT approach uses the data theory as a framework to curate and train the data that is based on the corpus such as the EIPC. For an under-resourced language such as isiZulu, the exigency exists to create such a tool in order to automate the translations between the two languages, since human translation cannot cater for the translation demand (Kituku *et al.* 2016:1).

Corpora have been widely used in linguistics research and beyond. They are at the core of many human language technologies like spellcheckers, machine learning, translators and lexicons. For many developed languages, corpora are increasingly used to provide excellent facilities for teaching and learning. Furthermore, since there is a connection between language and culture, by analysing a corpus, much can be learned about the communities represented in a corpus. Specialised corpora have also been applied in disease surveillance (Brownstein, Freifeld & Madoff 2009) and customer sentiment analysis in business (Pak & Paroubek 2010).

Perhaps it is useful to briefly highlight and reference some contestations in Corpus Linguistics (CL). While it has become prevalent in CL studies to use large bodies of processed natural languages data in order to

address theoretical issues, it is still contested whether CL is a tool, method(ology), discipline, theory or framework. Leech (1992: 106), Tognini-Bonelli (2001: 1) and Teubert (2005: 2) view CL as a theory. McEnery and Wilson (1996), Meyer (2002), Bowker and Pearson (2002) and McEnery *et al.* (2006: 7f.) view CL as a methodology.

> […] corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning, it certainly has a theoretical status. Yet theoretical status is not theory in itself (McEnery *et al.* 2006: 7f.).

It is evinced in this chapter that CL as a methodology is useful in language teaching and modelling grammatical descriptions.

3 Corpora for Language Teaching

Corpora can be used at multiple levels in order to contribute to effective language teaching in first or additional language learning. Corpora provide valuable information on the frequency of use of both grammatical or functional words, and content or lexical elements. Corpora are useful in providing additional information to intuition by providing evidence on attested use of language, and thus can influence the content and design of language modules positively. Corpora provide real, natural languages usage examples, which are not immediately available to a language teacher who uses illustrative, made-up examples that are based on intuition. Corpora can enhance language description as language learning moves away from introspection-based research, resulting in improved pedagogical grammars and more informative dictionaries.

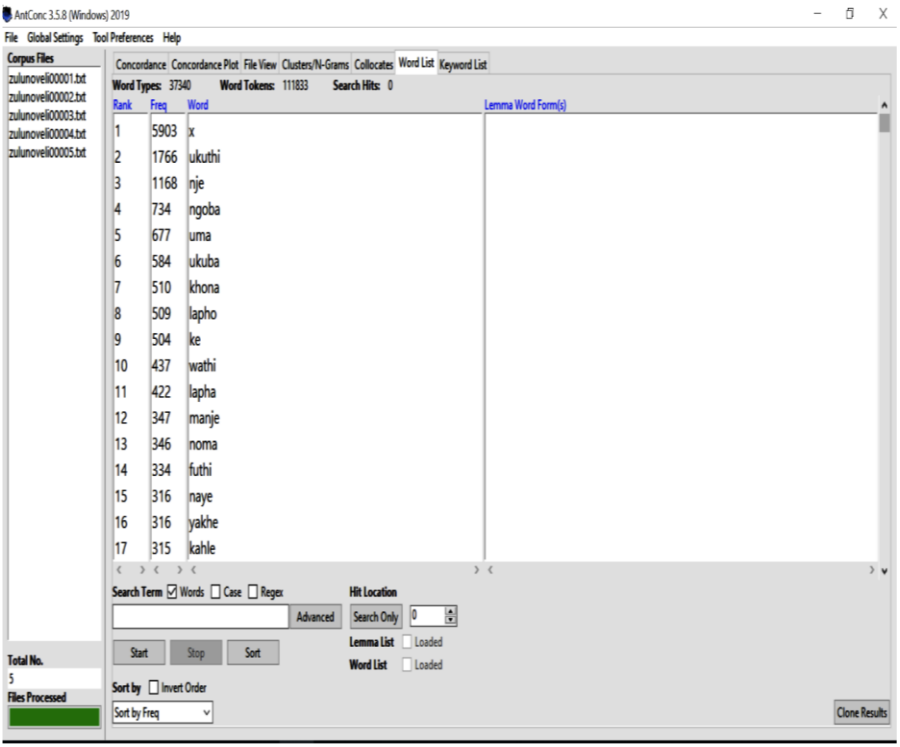
In a study on corpus-based teaching of Northern Sotho, Taljard (2012) observes that it is unsurprising that the pedagogy of language teaching for African languages lags far behind when compared to languages such as English. The traditional pedagogical material available for most African languages are notably inadequate in that they are premised on the structural model of grammatical description with no reference to frequency of use, real natural language usage and the communicative value of grammatical struc-

tures. The selection and sequencing of learning material are criticised in this study for its reliance on anecdotal evidence or on the intuition of the language teacher, which evidentially is often wide off the mark when compared to corpus data. It is the argument in this chapter that the use of corpus data provides the language teacher with guidance on both macro- and micro-level with regard to the content of the curriculum (Taljard 2012).

In order to access digital data that are stored in a computer, one needs a text analysis software programme that can assist in querying the corpus data. There are several software programmes such as Sketch Engine, WordSmith Tools, and AntConc. Sketch Engine, which has been developed by Lexical Computing since 2003, is the most advanced tool, with corpus management functions and complex text analysis systems. The WordSmith Tools, which was developed by Mike Scott and was first released in 1996, currently has version 7. It is an integrated suite of three main programmes, which include the WordList, Concord and Keyword. AntConc, which was developed by Laurence Anthony, is a freeware, multiplatform tool for carrying out corpus linguistics research and data-driven learning. As a multiplatform tool it can run on any computer with MS Windows, Macintosh OS X, and Linux. Sketch Engine and WordSmith Tools are proprietary; hence, prohibitive for massive online student learning. Because AntConc is a freeware tool and can be used on different operating system, it is ideal for isiZulu online teaching and learning.

Antconc contains seven tools. The user can easily access them and navigate from one tool to the other. These tools include the Concordance tool, the Concordance Plot tool, the File View tool, the Clusters tool, the Collocates tool, the Word List tool, and the Keyword List tool. These tools are useful in analysing various word units, word-in-context, word clusters, words and their preferred grammatical or semantic collocates (or combinations), most frequent and less frequent words, etc. In this study we sampled the INC in order to have a smaller sample corpus that we queried, using AntConc in order to demonstrate that isiZulu corpus can be used through a freeware software programme in teaching and learning isiZulu at UKZN. A sample corpus of about 110 000 running words generated from five (5) written texts was queried. Figure 1 is a screenshot showing a word list generated using the AntConc software. It shows the most frequent words generated from the sample corpus.

Figure 1. A wordlist of the most frequent words



The wordlist in Figure 1 flows from the most frequent word to the least frequent in a descending order. ***Rank*** stands for the number the word occupies in the list of the words that are in the word list, and ***Freq.*** is the number of times a word occurs in the corpus. The ten most frequent words in the sample corpus of the INC are *ukuthi*, *nje*, *ngoba*, *uma*, *ukuba*, *khona*, *lapho*, *ke* and *wathi*. All these words are function or grammatical words belonging to a closed word class. The closed word classes are concords, connectives, numerals, pronouns, etc. Studies in corpus linguistics show that frequency lists are commonly dominated by function words. However, this information is not intuitively available to a language learner. Hence the importance of a corpus in language learning.

Following this corpus evidence, the first task in language learning would be to identify closed word classes in isiZulu, and then putting all the words in the frequency list in their correct word classes. Then a concordance search, done by clicking on the isolated word (e.g. click on line 2 *ukuthi*), would show the function word *ukuthi* as used in its typical context in the corpus (see example 1). This provides new insights to the learners as they see various contexts in which the word is used in natural language.

Example 1. Concordance lines for *ukuthi*

1.	angibazondi 267 abantu besilisa ngoba ngiyazi	ukuthi	abababi bonke. Ngikholelwa ekutheni inhliziyo
2.	liyobe limshonele. Watshela amabutho akhe	ukuthi	ababheke kuzo zonke izindonga zendawo
3.	esiyinkinga emzini wakho. Nawe uzibonele	ukuthi	abafowethu bakwaDubazana banjani. Siyabhubha
4.	uNyokana akungenanga ngisho iphela. Okusobala	ukuthi	abagqekezi babone ukuthi angibalulekile, ngiyimp
5.	Angihlalanga khona kuyona ngoba ngibona	ukuthi	abaNguni laba osengitshela ngabo ngeke
6.	nazo zithule ngoba mhlawumbe zingazi	ukuthi	abaningi besizwe bakhalelani kayikho into
7.	ihlane igcwele amathambo bangaqondi	ukuthi	abantu abaningi kangaka bebebulawa yini
8.	nengifikelwa yizwi ngakho elithi angikusho,	ukuthi	abantu bakithi bonke bedukile, babalekelene
9.	yawakhuluma kukuba induna yayo seyiyitshelile	ukuthi	abantu bayo (inkosi) mabasale sebehlala
10.	thanda mngani wami. Ukhulunyiswa ngukwazi	ukuthi	abantu besilisa kabethembeke kangakanani.

It is notable that ***x*** is ranked as number 1 on the word list in Figure 1. However, this is not a word. ***X***, is the number of times that numbers have been counted in the corpus. Again, using this corpus freeware a concordance search for more context will illuminate this, as shown in example 2 below. The concordance examples in both example 1 and example 2 can be exported to a word document.

Example 2. Concordance line for ***x***

1.	ukungalaleli lokhu kwenu. Ngibeke u\	x	96a, abeke u\x96e;
2.	ngilahlekelwe enye ingxenye yami?\x94 \	x	93A, dade, wahlupheka\x96ke
3.	phaphama uThabekhulu, wethuka, wadonsa amathe.\	x	93A, emsebenzini!\x94 Wethuka noMaZondi,
4.	aphuma uMakhosazana, watholoza encika ngonina. \	x	93A, kanti usunentombi engaka! Bonke
5.	phansi lithi, \x93Makabongwe uThixo.\	x	94 3.3.A Ucingo lwangena ngehora leshumi
6.	\x96u aqhubeke abeke u\	x	96aa; ngibeke u\x96ee,

The Word List Tool allows one to export the word list into a word document.

Table 2. Showing 20 most frequent words

Rank	Frequency	Word
1	5903	x
2	1766	ukuthi
3	1168	nje
4	734	ngoba
5	677	uma
6	584	ukuba
7	510	khona
8	509	lapho
9	504	ke
10	437	wathi
11	422	lapha
12	347	manje
13	346	noma
14	334	futhi
15	316	naye
16	316	yakhe
17	315	kahle
18	308	phansi
19	300	Ujeqe
20	291	Njengoba

Table 3. Showing 10 most frequent content words

Rank	Frequency	Word
10	437	wathi
19	300	ujeqe
23	262	inkosi
27	212	abantu
28	206	udubazana
33	191	ethi
34	188	umazondi
35	184	amehlo
39	173	uthabekulu
43	161	ikhanda

It is also possible to export the entire word list without the limitation of the screen as shown in the screen shot in Figure 1. Table 2 shows an expanded word list. While most of the words in Table 2 are still function words, *ujeqe* (19) is a content word. Unlike grammatical words whose function is structural, content words refer to word units that carry a semantic content, which contributes to the meaning of a sentence they occur in. Nouns and verbs are the main examples of these types of words.

It is possible in a corpus class to further isolate content words from function words. This again can be a linguistic exercise that students can do

computationally. Table 3 below shows the 10 most frequent content words in the sample corpus of the INC that we queried. The content words are of interest to linguists and lexicographers. Linguists study their morphology, their morphosyntax and semantics. Lexicographers are interested in them because content words are the words that make up dictionaries. Since lexicography is the science of dictionary making, the frequency, concordance and collocation of content words are of scientific interest to a lexicographer. This computational approach to language study and language analysis makes the work of a linguist and that of a lexicographer more efficient, error-free and faster than it would be if it were done manually through human effort.

For example, one of the crucial decisions a lexicographer has to make in dictionary making is which word to include, or which word to exclude in a dictionary. With the availability of online resources such as AntConc, Summers (1996: 261) posits that 'all aspects of lexicography are influenced by frequency'. Headword selection for a particular dictionary becomes informed by the frequency through a statistical analysis, rather than a subjective intuitive exercise of a lone lexicographer. Frequency lists also provide developers of second language teaching material with the most relevant words, phrases, and expressions to teach. Frequency lists also shed more light on the most common words in isiZulu linguistic domain. These words may be the ones which characteristically typify the domain. According to Kilgarriff (1997: 135) 'The more common it is, the more important it is to know it'.

Content words provide a whole gamut of both linguistic and cultural information that is contained in a language. While the corpus enables one to isolate verbs and nouns as shown in Table 4 and Table 5[1], the frequency list suggests more insights. Table 4 is a list of the 12 most frequent verbs in the sample INC corpus that we queried. The first six most frequent verbs in this list are a variant of the main verb ***-thi*** (say). This is a monosyllabic verb, which can be inflected in a variety of ways, as shown in the word list. This verb and

[1] This can be enhanced if the corpus is annotated, which means some kind of linguistic analysis has been performed on the text. This includes marking up text with parts-of-speech markers or tags, which makes data retrieval from the corpus more precise, and fast.

its verb forms provide an interesting lesson in inflectional morphology and verb conjugation in isiZulu.

Table 4. Showing frequency list of verbs

Rank	Frequency	Word
10	437	wathi
33	191	ethi
71	113	athi
91	99	bathi
98	96	uthi
99	95	yathi
128	76	wabona
153	66	wahlala
154	66	wangena
155	66	waphuma
237	48	wasukuma
3529	4	sukuma

Table 5. Showing frequency list of 5 nouns

Rank	Frequency	Word
23	262	inkosi
35	184	amehlo
43	161	ikhanda
57	129	izinkomo
70	115	amabutho

The nouns in Table 5 also provide interesting insights. A noun is made up of two formatives, a prefix and a stem. This is an interesting aspect of the morphology of the nouns in isiZulu. While each noun in isiZulu is allocated a specific noun class, there are some nouns that are difficult to classify (Keet & Khumalo 2017), which makes it an interesting linguistic unit of study. The isiZulu noun class system is a distinct pairing of singular and plural nominal forms. However, there are interesting examples of nouns that do not seem to take a singular form. Because of the agglutinating nature of isiZulu, coupled with a conjunctive writing system, which glues together elements of an isiZulu word, as a result, a number of noun class prefixes in isiZulu are phonologically conditioned and yet others are homographs.

There are interesting studies that seek to explicate the generation of, and semantic motivation for, the various noun class assignments, not just in isiZulu but in Bantu studies. It is our argument that access to massive corpus data can lead to new insights in this area. It is also interesting that the focus on the noun covers areas such as phonology, morphology and semantics. Corpus studies also provide for an extended focus. For instance, in Table 5 above, the noun *inkosi* (the king), *izinkomo* (cattle) and *amabutho* (the warriors) are ranked 23, 57 and 70 in the frequency list. These words are all in the top 100 most frequent words, and top five most frequent nouns in the sample corpus. This suggests that *inkosi* (the king) occupies a very important seat in the organization of the Zulu people. The high frequency of *izinkomo* (cattle) also suggests that they must have a significant cultural influence in the Zulu social organization. The word *amabutho* (warriors) makes reference to the history of the Zulu people, when armies where organised in terms of *amabutho* (warriors). In order to study these highly frequent words further and deeper, the concordance tool provides more context, and a student delves into a deeper sociocultural worldview using evidence that is immediately available by means of an electronic corpus. The use of a corpus in language teaching clearly extends beyond language structure and second language learning, but extends into very interesting sociocultural and historical topics.

The corpus-based approach is also interesting for both linguists and lexicographers in the area of semantics and sense disambiguation, respectively. In the area of semantics, homographs present interesting challenges, particularly for second language learners, when words are spelt the same way but have different meanings. Table 6 is an example of a homograph in isiZulu.

Table 6. Showing frequency of the word noun *inyanga*

Rank	Frequency	Word
201	53	inyanga

The word *inyanga* is ranked just outside the top 200 most frequent words in the sample corpus. A concordance search, done by clicking on the word *inyanga*, produces the following concordances lines shown in Example 3 below.

Example 3. Concordance lines for *inyanga*

1.	nani zanuse ngoba kufike lapha	inyanga	ebhulayo neyelaphayo yaseSwazini. Ngayizwa
2.	oyenzayo wesabeka bathi abazange bayibone	inyanga	enje kwaZulu. Wanikwa ukuba angene
3.	ukuba ibe yisibonda somhlaba. "Kukhona	inyanga	enkulu kuphela, ubulawu bezulu. Lapha
4.	yokuficwa kukaJeqe enjengofileyo ogwini nokuthi	inyanga	enkulu yasesiqhingini imthathile, kuhle angesabi,
5.	unina weNkosi nabo bangene kukhona	inyanga	enkulu yeNkosi. Kwaphela isonto uJeqe
6.	singabe sisaba bikho isidingo sokulinda	inyanga	ezayo, ungagcina usuqale ngalo isonto
7.	eNkosi nenyanga bevimbeleke kuleyondlu, ibaphethe	inyanga	ibagcaba, iphalazisa iNkosi ngezintelezi ezinkulu
8.	kade ilala endlini yamakhosi namadlozi.	Inyanga	ilandele uJeqe bayolala elawini likaJeqe.
9.	uphuze utshwala besundu nobamaganu, uyokufa	inyanga	ingakapheli. Kodwa ungakadluli kusasa
10.	bangena emotweni eyangena umgwaqo lapho	inyanga	ishona ithi gqwambi ngale kwezintaba.

The dominant sense as shown in concordance lines 1, 2, 3, 4, 5, 6, and 8 is that of *inyanga* as a traditional doctor or traditional healer. In line 9 *inyanga* refers to the month, and line 10 *inyanga* is referring to the moon. In lexicography such data provide usage examples that are typical in natural language. It is also very useful in disambiguating various senses, as shown above. This is important to a lexicographer, as meaning reference is an intrinsic part of a dictionary.

4 Conclusion

The global lockdown of education institutions as a result of the coronavirus pandemic has forced many institutions to rethink the mode of delivering their course offerings. This entails preparing pedagogical materials to be made available in a structured way and available on an online platform that is accessible in digital forms. The challenge that this mode of delivery poses in the teaching of African languages is that most African languages are under-resourced. They do not have resources online that are immediately available to learners such as corpora, spellcheckers, and morphological and grammatical analysers.

However, we have shown in this study that UKZN has made progress in developing isiZulu, which has put it in a good position to be offered online. The INC is a very important resource that can be used in the teaching and learning of isiZulu via the digital mode of delivery. Using novel Digital Humanities approaches that infuse computational approaches in the teaching

of courses in the Humanities and Social Sciences, we have shown that the teaching of isiZulu phonology, morphology, morphosyntax, and semantics can be done online using the AntConc freeware and the open-source INC. This study also shows that the corpus approach can extend the learning and teaching of isiZulu to lexicography and lexicographic practice, sociocultural and historical spheres.

Acknowledgements
I am grateful to Ms. Neo Putini for corpus access, sampling and corpus queries. I am also grateful for comments from the two anonymous reviewers.

References

Bosch, S.E., L. Pretorius & J. Jones 2007. Towards Machine-Readable Lexicons for South African Bantu languages. *Nordic Journal of African Studies* 16,2: 131 - 145.

Bowker, L. & J. Pearson 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge. https://doi.org/10.4324/9780203469255 PMCid:PMC1732030

Brownstein, J.S., C.C. Freifeld & L.C. Madoff 2009. Digital Disease Detection Harnessing the Web for Public Health Surveillance. *New England Journal of Medicine* 360: 2153 - 2157. https://doi.org/10.1056/NEJMp0900702 PMid:19423867 PMCid:PMC2917042

isiZulu National Corpus. languagematters@ukzn. Available at: https://iznc.ukzn.ac.za/

Khumalo, L. *forthcoming*. Corpora as Agency in the Intellectualization of African Languages. In Kaschula, R. & H.E. Wolff (eds.): *The Transformative Power of Language: From Postcolonial to Knowledge Society in Africa*. Cambridge: CUP.

Keet, C.M. & L. Khumalo 2017. Grammar Rules for the isiZulu Complex Verb. Southern African Linguistics and Applied Language Studies 35,2: 183 - 200. https://doi.org/10.2989/16073614.2017.1358097

Keet, C.M. & L. Khumalo 2017. Toward Verbalizing Ontologies in isiZulu. In Davies, B., L. Kaljurand & D. Kuhn (eds.): *Controlled Natural Language Proceedings.* Switzerland: Springer.

Kilgarriff, A. 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10,2: 135 - 155. https://doi.org/10.1093/ijl/10.2.135

Kituku, B., L. Muchemi & W. Nganga. 2016. Framework for Many to One Machine Translation. *International Journal of Advanced Research in Computer Science and Software Engineering* 6,5: 103 - 110.

Language Policy of the University of KwaZulu-Natal 2006. Available at: http://utlo.ukzn.ac.za/Libraries/November_2011_Conferences/APPENDIX_D4_Language_Policy_-_Council_approved_010906.sflb.ashx

Language Policy of the University of KwaZulu-Natal 2014. Ref: CO/02/0109/06. Unpublished.

Leech, G.N. 1992. Corpora and Theories of Linguistic Performance. In Svartvik, J. (ed.): *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm*, 4 - 8 August. Berlin/New York: Mouton de Gruyter.

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

McEnery, T., R. Xiao & Y. Tono. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London/New York: Routledge.

Meyer, C.F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511606311

Pak, A. & P. Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Paper presented at the LREC. 17-23 May, Valletta, Malta.

Pretorius, L. & S.E. Bosch. 2003. Finite State Computational Morphology: An Analyzer Prototype for Zulu. *Machine Translation* 18: 195 - 216. https://doi.org/10.1007/s10590-004-2477-4

South Africa's People: Languages. Available at: https://www.gov.za/about-sa/south-africas-people#languages

Summers, D. 1996. Computer Lexicography: The Importance of Representativeness in Relation to Frequency. In Thomas, J. & M. Short (eds.): *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech.* London/New York: Longman.

Taljard, E. 2012. Corpus-based Language Teaching: An African Language

Perspective. *Southern African Linguistics and Applied Language Studies* 30,3: 377 - 393. https://doi.org/10.2989/16073614.2012.739318
Teubert, W. 2005. My Version of Corpus Linguistics. *International Journal of Corpus Linguistics* 10,1: 1 - 13. https://doi.org/10.1075/ijcl.10.1.01teu
Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam/ Philadelphia: John Benjamins. https://doi.org/10.1075/scl.6

Langa Khumalo
Director
University Language Planning and Development Office
University of KwaZulu-Natal
Durban
khumalol@ukzn.ac.za