The Role of Spoken Language Corpora in the Intellectualisation of Indigenous Languages in South Africa

Mtholeni N. Ngcobo
Nozibele Nomdebevana

Abstract

South Africa is a multilingual country with eleven official languages. Nine of these official languages are indigenous languages and they are not considered as developed compared to English and Afrikaans. The main cited problem is the lack of appropriate terminology that will enable these languages to be used in modern domains of language use such as science and technology. Term creation for the purpose of modernising the indigenous languages of South Africa falls under a process called intellectualisation. This is a process of accelerating the growth and development of languages. In this article we argue that we need to use spoken corpora of these indigenous languages as a strategy in their intellectualisation. Building large spoken language corpora is regarded as the most important first step in the development of adequate language material and other applications. Apart from being rich sources for varied and significant empirical research in linguistic, cultural and natural language processing domains, language corpora have globally become extremely important data base resources for a variety of linguistic, socio-economic, cultural, educational, inter-lingual or inter-communicational and language technological applications and developments (such as in machine translation and human-machine interactions). The spoken corpus approach allowed us to extract and to observe the nature of borrowed terms used by

people in natural communication. We extracted a representative sample from Xhosa and Zulu transcribed spoken corpora and the results showed that there is a significant number of new terms in spoken communication. These terms can be used as the basis for an objective strategy in the planned process of accelerating the growth and development of indigenous languages.

Keywords: spoken language corpus, language development, indigenous languages, intellectualisation, African Renaissance, South Africa.

Introduction

South Africa has a multilingual language policy that accords eleven languages an official status. It is, however, noticeable that only two of these official languages (English and Afrikaans) are considered to be more developed in the country. The other nine, which are all indigenous languages, are not developed sufficiently to cope with modern society. One of the requirements for the development of a language is the planning of its corpus. This refers to the standardisation and intellectualisation of a language. Wardhaugh (2006:33) defines language standardisation as

> the process by which a language has been codified in some way. That process usually involves the development of such things as grammars, spelling books, and dictionaries, and possibly a literature.

Language intellectualisation is the process of language elaboration where new terms are coined and the language is modernised to be used for specialised communication (Mwansoko 2004). It may be argued that while most of the indigenous official languages in South Africa are standardised, progress in their intellectualisation is very slow. According to Garvin (1973), intellectualisation is an important dimension of language development. He (Garvin 1973:43) argues that in the developing languages intellectualisation is a way of providing

> more accurate and detailed means of expression, especially in the domains of modern life, that is, in the spheres of science and technology, of government and politics, of higher education, of

contemporary culture, etc.

The process of intellectualisation of the African languages is directly linked to the notion of African Renaissance (cf. Alexander 2003). The concept of Renaissance encompasses a resurgence of learning based on classical sources. It also puts emphasis on rebirth and reform. Reforming indigenous languages in South Africa will require (among other things) the use of modern tactics including technology (cf. Crystal 2000; Ngcobo 2009). It will also require taking cognisance of the fact that language adapts to a changing world as it is used in specialised domains of language use.

Finlayson and Madiba (2002) addressed the intellectualisation of the South Africa's indigenous languages. Although they argue that intellectualisation can be deliberately planned, they also note that it can occur naturally (Finlayson & Madiba 2002:40). They also assert that there is a lack of strategies as to how intellectualisation can be done (Finlayson & Madiba 2002:46). Similarly, since there is a wide variety of methods used in the standardisation of South Africa's indigenous languages, there is an urgent need for compiling the spoken language corpora of the nine indigenous languages in according to a standardised morphology (Allwood & Hendrikse 2003). This can then serve as basis for specialised communication.

In this article we argue that spoken language corpus is the most relevant method that can be used effectively in the intellectualisation of indigenous languages in South Africa. A centralised approach to terminology development can benefit from decentralised approach which will take into consideration language as it is used in interaction (cf. Tollefson 1981:183). As Cooper (1989) notes, languages develop through use and the form always follows the function. Terms may be extracted from the day-to-day interactive natural speech rather than by relying on the subjective term creation done by language experts in isolation. By following a corpus-based approach, we use examples from the spoken corpora of Zulu and Xhosa languages compiled under the Southern African Spoken and Signed Language Corpora Project (SASSLC) at the University of South Africa. These two Nguni[1] languages are selected in this study mainly because they have the largest number of

[1] Nguni is a group of languages spoken in South Africa including isiZulu, isiXhosa, isiNdebele and siSwati.

mother tongue speakers in the country compared with other indigenous languages (cf. Stats SA 2001). Another reason is that they have representative spoken corpus data in the SASSLC project. In the first part of this article we provide a background and motivation by explaining the rationale of this study. We then proceed to a discussion of the importance of spoken language corpora in the development of languages. We also briefly present the current status of the spoken language corpus project by describing what has been achieved so far and explaining the various problems that have been encountered in developing corpora for indigenous languages in South Africa. After describing the method followed in this study, we then demonstrate how spoken language corpora can become the basis for extracting terms that can be used in the intellectualisation process.

Background and Motivation

The reason that we decided to embark on this study is the desire for people and language scholars to intervene in the issue of language and change. Such zeal has allowed space for a conversation on language reform and development in South Africa. According to Finlayson and Madiba (2002:53), in South Africa,

> not much research has been done so far to establish guiding principles and procedures for the development of new terminology for the African languages.

They argue further that 'such principles should give guidance on the choice of the word-formation patterns and also on their actual use' (Finlayson & Madiba 2002:53). There is also dissatisfaction about the spelling systems and a lack of suitable terminology for specialised communication in the indigenous languages (cf. Finlayson & Madiba 2002, LANGTAG 1996). In this article we concentrate on the issue of terminology in the intellectualisation of the indigenous languages. The issue of terminology in the development of indigenous languages has often been pursued from a pessimist perspective (Van Huyssteen 1999), and in the absence of formal motivation most new terms that exist in spoken interaction have not been

taken into consideration for language intellectualisation. The other issue has been the tendency of the purists to refuse to accept that indigenous languages have to adapt to a new environment of change in order for these languages to develop into languages of specialised communication (Zuckermann 2009).

Although there are efforts aimed at developing indigenous languages in South Africa, these activities have ignored terms coined during interaction in spoken language corpora. As a result we see an attempt to coin terms that are totally foreign to the speakers of indigenous languages as they are created by a few individuals in isolation. In this article we seek to present a persuasive objective argument concerning terminology development which is based on language use. This can be seen as a proposal put forward for a better approach to intellectualising indigenous languages. Fishman (1977:37) argues that when a language moves into functions for which it was not previously employed, its modernisation becomes necessary if it is to fulfil its new roles. In the case of the indigenous languages of South Africa we can argue that teaching in the medium of these languages in schools and in tertiary education can necessitate the introduction of many technical terms into these languages to express the concepts that were not originally available in them. Terms extracted from spoken interaction can eventually be standardised and used in learning and teaching material.

We have already stated in the introduction to this article that intellectualising indigenous languages may be considered as part of the African Renaissance. The latter is about rebirth and reform. The tools of language reform are either simplification or purification (cf. Lewis 2002). Simplification is about regularising vocabulary and grammar while purification on the other hand is about conforming to a version of a language known as pure. The latter entails the opposition to any changes of a given language and it is a desire to get rid of those changes. Purification may not necessarily help in terms of language intellectualisation, as it does not allow adoptives that arise from multilingual situations. In this article we focus on simplification as a progressive strategy in language intellectualisation (Zuckermann 2009). We follow Allwood (2001) in that terminology development based on the spoken language corpora which reflect language used in real life can do more to help than to hinder education and culture as highlighted in the following section.

The Importance of Spoken Language Corpora in Language Development

Spoken language corpora can allow for the use of statistical performance measures and observation of language used in real life (cf. Biber *et al.* 1998; Biber & Finegan 1991; Allwood 2001). Written language corpora have been already collected by other institutions for a few indigenous languages. The most noticeable corpora of written texts are the Zulu corpus and Sepedi corpus based at the University of Pretoria. However, written corpora do not necessarily represent how language is used in real life since it is based on fiction and other formal written texts such as the newspapers, school books, etc. The focus, therefore, must be on corpora of spoken varieties of these languages, since 'spoken language varies considerably in different social activities with regard to pronunciation, vocabulary, grammar and communicative function' (Allwood *et al.* 2001).

Allwood and Hagman (1994), on the other hand, assert that spoken language is a fundamental trait of the human species. They further argue that

> spoken language is fundamental also from a social point of view since it is integrated not only with the human brain but also with human society in various not yet totally understood ways (Allwood & Hagman 1994:1).

Previously, linguistic research focused on written language and as a result this has limited our knowledge of spoken language (cf. Linell 1982; Harris 1980). However, according to Allwood and Hagman (1994:1), the progress in audio, video and computer technology enables us 'to record and analyse spoken language without having to rely on either memory or written language'. It is due to such technological advancement that during the last two decades there has been a strong focus on the corpus linguistic approach.

The Spoken Language Corpora of Indigenous Languages in South Africa

The first task in corpus development is the compilation of a body of texts called a corpus. The rise of computer knowledge and the capacity of

computers to store up large quantities of information have greatly enhanced corpus studies. While English and other European languages are currently far advanced in this development, there are still no comprehensive spoken corpora available for the indigenous languages in South Africa. Therefore, the research and the development of these languages is hampered due to a shortage of data. A corpus represents an enormous research potential, which entails linguistic, social, cultural, educational, inter-lingual and inter-communicational and technological aspects.

In South Africa spoken language corpora for the official indigenous languages are being developed by the University of South Africa (UNISA) as an open-ended corpus project. The first goal of the spoken language corpus project is to produce an average of one million words per language (Allwood & Hendrikse 2003: 194). It is envisaged that the collected corpus data will be used to accomplish the following aims:

- to establish a spoken language corpus research centre
- to adapt and develop computational linguistic software suitable for the agglutinating languages of South Africa
- to develop the relevant African languages of South Africa, and
- to understand the role of language and communication in real life situations.

The intention is to include spoken language from as many social activities as possible in order to get a more complete understanding of the role of language in society. In the meantime, recordings for a pilot study have been collected and a description of the planning phase has been published in some articles (cf. Allwood *et al.* 2003; Allwood & Hendrikse 2003).

Progress and Problems

Out of nine official indigenous languages of South Africa it is only Xhosa that has been able to show greater progress in corpus development. This is because Xhosa was used for piloting the project and it has a consistent transcriber. Zulu is lagging behind with less than 25,000 transcribed tokens so far. The Xhosa corpus includes a total of 45,723 tokens from audio recording and 201,292 tokens from video recordings.

It is discouraging, however, to note that little or nothing is currently happening in the development of the spoken language corpora for the remaining official languages. This is due to insufficient tools and a lack of sufficient financial and human resources. Even in the languages that have shown some progress in corpus development, there is no appropriate monitoring as some of the video recordings get damaged and some of the digitized recordings have been lost. Some of the recordings are of such a poor quality that no transcriptions can be made based on them. Besides these problems there are many uncoordinated individual activities that are going on in corpus development. The other problem is the word-hood in the agglutinating languages. While it is easy to identify words in isolating languages, in agglutinating languages it is almost impossible to find words. The so-called words in agglutinating languages are complex and morpho-syntactic. For example, the Zulu construct '*yayingakaguqulwa*' means 'it had not been changed' and it has a lexico-grammatical structure as it can also be analysed at the syntactic level.

Methodology

General Procedures

In order to see the nature of words that can be considered as borrowed in the two Nguni languages, we extracted from the transcribed spoken Zulu and spoken Xhosa corpora approximately 20, 000 tokens per language. We wanted to use a method which would identify the frequency of occurrence of borrowed words in normal spoken language. Once the frequency of occurrence of borrowed words in spoken Zulu and Xhosa corpora can be determined this can be analysed and be contrasted with original Zulu and Xhosa words that are found in spoken corpora. A spoken corpus-linguistics approach was selected as the most appropriate method in making observations about the nature of intellectualisation that is already taking place in spoken languages.

Data Analysis

We applied *WordSmith Tools* to isolate each token that is found in transcribed spoken corpora. *WordSmith Tools* is a form of lexical analysis

software that was developed by Mike Scott. The application of *WordSmith Tools* to a corpus results in isolating each token that is used in the corpus. Our task was to extract borrowed words and concepts from the list and note their frequency of use in transcribed spoken Zulu and spoken Xhosa corpora. We also wanted to look at the nature and the formation of these words in order to support the hypothesis that the spoken corpus data has a significant number of borrowed lexical items that can be used as the basis for the intellectualisation of indigenous languages. The tools gave us access to the ranked frequency of all words in the sample corpora of both languages. By using the token lists we were able to obtain a sample of single words extracted from the transcribed spoken corpus data. The list called *A* provided us with alphabetically arranged words and the frequency of their occurrence.

Results and Discussion

In the transcribed spoken Zulu corpus data of exactly 18954 tokens and 5584 word types we found 285 tokens representing borrowed words and code switches and mixes (see Appendix). That counted for 114 word types in the total sample corpus. In the transcribed spoken Xhosa corpus of exactly 19238 tokens and 5143 word types we found 430 tokens representing borrowed words and code switches and mixes. This counted for 151 word types in this sample corpus.

The observation here is that while Zulu has a significant number of borrowed words, Xhosa has more of these words than its counterpart. Another observation is that while Zulu classifies borrowed words in almost every class gender, Xhosa only classifies these words in specific classes (see Appendix). It was also observed that these borrowed words can be categorised into various groups according to how they are borrowed and adapted in indigenous languages.

Adoptives

In this process the borrowing is integrated into the linguistic structure of the borrowing language. This includes new words that are formed by adapting foreign words into the phonological and lexical system of the borrowing language as reflected in the following examples:

Zulu	**Gloss**	**Xhosa**	**Gloss**
Udokotela	Doctor	*Usisi*	Sister
Ijele	Jail	*Izinki*	Zinc
Isikole	School	*Iyunivesithi*	University
Imali	Money	*Ikoloni*	Colony

Sounds that are not in the borrowing language are replaced with a similar sound such as the case of /r/ in the word 'doctor'. The spelling is adapted and the word such as 'zinc' is pronounced as '*uzinki*'. Morphologically these words adopt particular class genders in South African indigenous languages. For example, classes 5 and 9 become the default class genders where speakers of Nguni languages automatically classify borrowed words. Some exceptions in the classification of borrowed words are due to factors such as morphology and semantics. For example, '*isikole*' is classified under class 7 because morphologically the first consonant of the word 'school' has become a basic prefix in the adapted form. Some words such as '*unobhala*' (secretary) are classified in class 1(a) because their semantic properties are associated with humans.

Loan Translations or Calques

Loan translation is a process of translating morphologically complex foreign expressions by means of combinations of native elements that match the meaning and the structure of the foreign expression and their component parts. The following examples depict this phenomenon:

Zulu	**Gloss**	**Xhosa**	**Gloss**
Umengameli (derived from the verb e*ngamela* - 'preside')	President	*Unobhala* (from the act of writing - *bhala* - 'write')	Secretary
Umcwaningi (from the verb *cwaninga* - 'analyze')	Researcher	*Umqhubi* (from the act of driving - *qhuba* - 'drive')	Driver

Umnotho (from *notha* - 'rich')	Economy	*Umabonakude* (from to see and distance - bona kude - 'see far')	Television
Isichazamazwi (something that explains words - *chaza amazwi* - 'explain words')	Dictionary	*Isithuthi* (that which moves from one place to another - *thutha* - carry away)	Transport

Coinages

Coinage refers to the invention of new words. A new word that does not exist in a language is formed from a native perspective of what the borrowed object or phenomenon might mean. For example, the word '*ingculaza*' (HIV/Aids) is a coined word and it has the invented meaning to native speakers of the language. Although this word has never existed before in the borrowing language, native speakers can easily understand it because it describes the painful act that comes as a result of this particular new disease.

Loan Shift

This takes place when a word that already exists in a borrowing language is used to refer to a borrowed concept which is similar in function or meaning to that which exists in the borrowing language. This is sometimes referred to as the semantic shift. For example *ingwazi* is a word that was used to refer to a person who was a hero in the war, but now the term refers to someone who is a champion or prolific in his or her work.

Codemixes and Codeswitches

It is also noticeable in the words extracted from the corpus that there is significant number of words that retain their original form or some part of

their form from the donor language. In this case we have found a lot of English words in Xhosa and Zulu corpora. Xhosa has more of these words compared to Zulu. However, the use of English in Zulu and Xhosa spoken text may depend on whether people are bilingual or not. The educated will most likely use English forms instead of adapted forms, as English is the lingua franca in South Africa and it is regarded as a language with high status.

Synonyms

There is also an abundance of synonyms as a range of new terms are sometimes used to refer to a single entity. For example *ingculazi*, *ugawulayo*, *i-Aids* refer to HIV/Aids respectively. People understand all these words and they use each word depending on context. One would think that it is only the educated people who can refer to HIV/Aids as *i-Aids*. But even people who do not speak English understand the latter term since this word has been used repeatedly because of the severity and the scourge of this pandemic. According to Finlayson and Madiba (2002:47), we will need to define distinctions between words that may have been synonyms. The use of the statistical method based on spoken corpora can allow us to see which words are commonly used in spoken interaction.

Some Observations

From the foregoing, it is clear that success in the intellectualisation of indigenous languages in South Africa will depend on the strategies used in documenting new terminology for specialised communication. Ngcobo (2009) argues that we need to use modern strategies such as communication and technology in promoting the use and development of indigenous languages in South Africa. If indigenous languages are used in modern and specialised domains such as science and technology, the communicative imperative will result in new intellectualisation terms being developed in speech rather than in isolation. Finlayson and Madiba (2002) argue that the failure to use African languages as media of instruction will surely impede the intellectualisation of these languages. They argue further that,

> the use of a language as a medium of instruction … creates the demand for new terms and further provides the opportunity to use them in meaningful contexts, e.g. in textbooks, in teaching, for scientific experiments and so forth (Finlayson & Madiba 2002:46).

The use of methods such as borrowing can contribute in the intellectualisation of indigenous languages. In a multilingual situation it is almost impossible to exclude new words that come as a result of interaction with other languages. It is also an observable fact that since English is prominent in modern domains we can expect that it will influence the intellectualisation of indigenous languages in great measure. For example, in a bilingual environment code switches or mixes are used to deliver the message. This strategy provides a clear sense of optimism for the development of indigenous languages as media of instruction.

Conclusion

In this study we have revealed that we need to use spoken language data in the intellectualisation of indigenous languages. Corpora may be used as the basis for generating new term lists. This is the strategy that can help in providing objective term creation and the acceptance of new terms by a wider community. The use of spoken corpus data for the two Nguni languages (Xhosa and Zulu) shows that there is a significant number of raw terms coined by people in interaction. These terms may be used as the basis for generating other terms as the need arises. This is a significant step forward towards the intellectualisation of the indigenous languages. Indigenous languages' morphology is very flexible and therefore allows a variety of word formation patterns with ease.

References

Alexander, N 2003. *The African Renaissance and the Use of African Languages in Tertiary Education.* Cape Town: PRAESA. (PRAESA Occasional Papers No. 13.)

Allwood, J 2001. Capturing Differences between Social Activities in Spoken Language. In Kenesei, I & RM Harnish (eds): *Perspectives on Semantics, Pragmatics and Discourse*. Amsterdam: John Benjamins.

Allwood, J, L Grönqvist, E Ahlsén & M Gunnarsson 2002. Annotations and Tools for an Activity Based Spoken Language Corpus. *Procedures of the Second SUIGdial Workshop on Discourse and Dialogue,* Alborg, Denmark.

Allwood, J, L Grönqvist & AP Hendrikse 2003. Developing a Tag Set and Tagger for the African Languages of South Africa with Special Reference to Xhosa. *Southern African Linguistics and Applied Language Studies* 21,4: 221-235.

Allwood, J & J Hagman 1994. Some Simple Automatic Measures of Spoken Interaction. *Procedures of the 14th Scandinavian Conference of Linguistics & 8th Conference of Nordic and General Linguistics 72,* University of Göteborg.

Allwood, J & AP Hendrikse 2003. Spoken Language Corpora for the Nine Official Languages of South Africa. *Southern African Linguistics and Applied Language Studies* 21,4: 187-199.

Biber, D, S Condrad & R Repen 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D & E Finegan 1991. On the Exploitation of Computerised Corpora in Variation Studies. In Aimer K & B Altenberg (eds): *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Cooper, RL 1989. *Language Planning and Social Change*. Cambridge: Cambridge University Press.

Crystal, D 2000. *Language Death*. Cambridge: Cambridge University Press.

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

Fishman, JA 1977. Comparative Study of Language Planning: Introductory Survey. In Rubin J *et al.* (eds): *Language Planning Processes*. The Hague: Mouton.

Garvin, PL 1973. Some Comments on Language Planning. In Rubin J & R Shuy (eds): *Language Planning: Current Issues and Research*. Wahsington, D.C.: Georgetown University.

Harris, R 1980. *The Language Makers*. Duckworth: London.

LANGTAG 1996. *Towards a National Language Plan for South Africa. Language Plan Task Group Report*. Pretoria: Ministry of Arts, Culture, Science and Technology.

Lewis, G 2002. *The Turkish Language Reform: A Catastrophic Success*. New York: Oxford University Press.

Linell, P 1982. *The Written Language Bias in Linguistics, SIC 2*. Linköping: University of Linköping. (Tema Kommunukation.)

Mwansoko, HJM 2004. Kiswahili Intellectualization Efforts in Tanzania. *Journal of Asian and African Studies* 67: 151-161.

Ngcobo, MN 2009. A Strategic Promotion of Language Use in Multilingual South Africa: Information and Communication. *Southern African Linguistics and Applied Language Studies* 27,1: 113-120.

Stats SA 2001. Census 2001.

Tollefson, TW 1981. Centralized and Decentralized Language Planning. *Language Problems and Language Planning* 5,2: 175-188.

Van Huyssteen, L 1999. Problems Regarding Term Creation in the South African Languages, with Specific Reference to Zulu. *South African Journal of African Languages* 19, 3:179-187.

Wardhaugh, R 2006. *An Introduction to Sociolinguistics*. Malden, MA: Blackwell Publishing.

Zuckermann, G 2009. Hybridity versus Revivability: Multiple Causation, Reform and Patterns. *Journal of Language Contact, Varia* 2: 40-67.

Appendix

New Terms Attested in Zulu and Xhosa Corpora

Zulu Nouns	English gloss	Cases	Xhosa Nouns	English gloss	Cases
Gender 1a/2a					
U-mongameli	President	1	Ubhuti	Brother	21
U-dokotela	Doctor	21	Unongxowa	Treasurer	48
U-Jesu	Jesus	1	Utata	Father	6
U-sihlalo	Chairperson	3	Ugqirha	Doctor	6
U-solwazi	Professor	2	Umam	Madam	1
U-chwepheshe	Expert	2	Udisemba	December	1
U-mhleli	Editor	3	Urhulumente	Government	8
U-thisha	Teacher	1	Unobhala	Chairperson	10
U-Mrs	Mrs	1	Uthixo	God	4
			Unesi	Nurse	1
			usisi	Sister	1
Gender 1/2					
Um-cwaningi	Researcher	3	Umlungu	White person	9
Um-lungu	White person	1	Umqhubi	Driver	4
Um-fundisi	Reverend	1	umthwebuli	Photographer	1
Gender 3/4					
U-hulumeni	Government	1	Umkhenkce	Ice	1
U-beetroot	Beetroot	1	Umzuzu	Minute	1
U-gawulayo	HIV/Aids	1	Umabonakude	Television	6
Um-khandlu	Council	3	U-sugar	Sugar	1
Um-gwaqo	Road	1	Umbane	Electricity	1
Um-khakha	Field (of study)	1	Uduladula	Bus	1
Um-notho	Economy	1			
Gender 5/6					
Ilokishi	Township	3	Ityotyombe	Slum	1
I-dolobha	Town	1	Izinki	Zinc	1
I-researcher	Researcher	1	Ikhumsha	Modernised per	1
Ijele	Jail	1	Inesi	Nurse	1
I-shalofu	Shelf	1	Ipasi	Identity doc	2

I-bhokisi	Box	1	Ikwayikwayi	High hilled shoe	1
I-bhuku	Book	1	Ihotela	Hotel	2
I-khafula	Kaffir	1	Ikwerekwere	Foreigner	1
I-ndiya	Indian	1			
I-philisi	Tablet/pill	5			
I-phepha	Paper	1			
I-holidi	Holiday	1			
I-phoyisa	Police man	1			
I-sosha	Soldier	2			
I-bhadi	Bad luck	2			
I-bhayibheli	Bible	1			
I-bhanoyi	Aeroplane	2			
I-viki	Week	1			
I-khemisi	Pharmacy	2			
I-ngisi	English person	1			
Gender 7/8					
Isi-bhedlela	Hospital	6	Isibhedlele	Hospital	15
Isi-kole	School	8	Iziyobisi	Drugs	4
Isi-poti	Drinking area	1	Isikolo	School	7
Is-ayoni	Zion	2	Isihlangu	Shoe	5
Isi-chazamazwi	Dictionary	12	Isibhamu	Gun	3
Isi-funda	Region	1	Isilungu	English	2
Isi-fundazwe	Province	2	Isimomondiya	Pretty girl	1
Isi-hlalo	Chair	1	Isibhulu	Afrikaans	2
Isi-kali	Scale	3	Isithuthi	Transport	1
Isi-khungo	Institution	12	Isipaji	Wallet	1
Isi-lungu	English	5	Isingesi	English	1
Isi-ngisi	English	5	Isithuthuthu	Motor-bike	4
Is-ono	Sin	1	Isithixo	a stature/god	2
Is-onto	Church	4	isitulo	Chair	1
Isi-klabhu	Sheep	2			
Isi-gaba	Section/ paragraph	1			
Gender 9/10					
Im-ali	Money	8	Incwadi	Book	9
I-genge	Gang	6	i-Afrika	Africa	3
Im-oto	Motor car	7	Imali	Money	21

In-gculazi	HIV/Aids	34	i-Medunsa	Medunsa	2
Ibhodi	Board	1	Imelika	America	1
i-computer	Computer	1	Inkonzo	Service	5
I-geji	Gauge	1	Iyunivesithi	University	1
i-guiter	Guitar	1	I-ajenda	Agenda	1
In-cwadi	Book	2	Idrinki	Drink	1
In-diza	Aeroplane	1	Ilali	Rural area	2
In-hlosongqangi	Main aim	1	Imoto	Car	4
In-tatheli	Journalist	2	Impi	War	2
i-punchbag	Punch-bag	1	Ingqesho	Employment	2
i-support	Support	1	Irabhi	Rugby	1
I-tekisi	Taxi	2	Iphanado	Panado	1
In-gwazi	Prolific person	1	Isheleni	Ten cents	1
In-konzo	Church service	1	Itoliki	Interpreter	2
			Itranskayi	Transkei	2
			Iveni	Van	1
			Ikoloni	Colony	2
			Ikota	Quarter	1
			i-dean	Dean	1
			itishala	Teacher	2
			Ingculazi	HIV/Aids	1
			Icawa	Church/Sunday	1
			Iswelikle	Sugar	4
			Ifama	Farm	1
			itshomi	Friend	1
			Ibhasi	Bus	1
			Idolophu	Town	2
			Ikaka	Faeces	5
			Ihagu	Pig	1
			Ibhola	Ball	3
			Igusha	Sheep	1
			Ikofu	Coffee	1
			Iholide	Holiday	1
			Ipali	Pole	1
			Ikati	Cat	1
			Iti	Tea	1
			Itreyi	Tray	1

			Iveki	Week	1
			Ipasika	Easter	1
			Ijele	Jail	1
Gender 11/10					
U-hlamvu	Consonant	2			
U-cingo	Phone/wire	1			
U-cwaningo	Research	1			
U-shanguzo	ARVs	1			
Gender 14					
Ubu-chwepheshe	Technology	1			
Ubu-krestu	Christianity	1			
Ubu-dokotela	Medical profess.	1			
Gender 15					
Uku-cubungula	To analyse	5			
Uku-preventa	To prevent	1			
Uku-shicilela	To publish	1			
Zulu verbs/adjectives			**Xhosa verbs/ adjectives**		
jova	Inoculate	1	fona	Phone	1
ngcwele	holy	1	owebala	Coloured	2
jima	Gym exercise	1	tolika	Interprete	1
cwaninga	Research	2	bhala	Write	11
qopha	Record	1	bhalisa	Register	2
bhala	Write	3	driva	Drive	1
tolika	Interpret	2	khumusha	Speak English	3
Shutha	Shoot	1	bhatala	Pay	9
cubungula	Analyse	1	thenga	Buy	1
shayela	Drive	1	tshintsha	Change	2
shintsha	Change	2	qhomfa	Abort	1
			thwebula	Take photo	1
			qhuba	Drive	3
			thengisa	Sell	1
			tyikitya	Sign	1
English word			**English words**		
right		1	About		1

Africa	2	Admires	1
Aids	2	Adoptive	1
as	4	Advertise	1
behalf	1	And	1
but	3	As	1
buses	1	But	2
coming	1	Code	8
community	1	dragging	1
discussion	1	Drive	1
English	1	English	8
first	1	Ever	1
hand	8	Example	1
handshake	2	Flat	1
HIV	1	For	1
introduction	1	Head	1
just	1	His	1
mayor	2	ID	2
MP	1	Is	2
problem	1	Know	5
question	1	Laughter	1
		Loan	1
		Man	8
		Minutes	1
		Mix	5
		Money	1
		Next	2
		Not	1
		Numbers	1
		Ok	1
		Shake	1
		Should	2
		Shouldn’t	1
		Sixteen	1
		So	2
		Sound	1
		Storytelling	1
		Switch	3
		Talk	4

		Teen	2
		Teenage	2
		Themselves	2
		Time	2
		TV	1
		Twelve	1
		Unwittingly	1
		USA	1
		US	1
		Vir	1
		Word	1
		You	7
TOTAL	**285**		**430**

Mtholeni N. Ngcobo
Department of Linguistics
University of South Africa
ngcobmn@unisa.ac.za

Nozibele Nomdebevana
Department of Linguistics
University of South Africa
nomden@unisa.ac.za