Wordsmith Tools as an Enabler for Text Analysis

Rooweither Mabuya

South African Centre for Digital Language Resources
North West University

Roo.Mabuya@nwu.ac.za

Abstract

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

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

1 Introduction

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

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

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

2 Corpus Building

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

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

It is important to note that many corpora of different sizes and typologies have been developed for different languages that are spoken globally. Some of these are available through a repository called the Sketch Engine platform. Sketch Engine is a platform for corpora management and analysis and creation of corpora that is accessed through payment of a subscription. It has corpora in ninety (90) languages, with some like English having multiple corpora. African languages on Sketch Engine are Afrikaans, Arabic, Igbo, Setswana, Swahili and Yoruba.

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

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

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

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

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

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

2.1 Medium

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

2.2 Design method

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

2.3 Size

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

2.4 Language variables

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

2.5 Mark-up and annotation

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

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

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

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

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

3 Wordsmith Tools

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

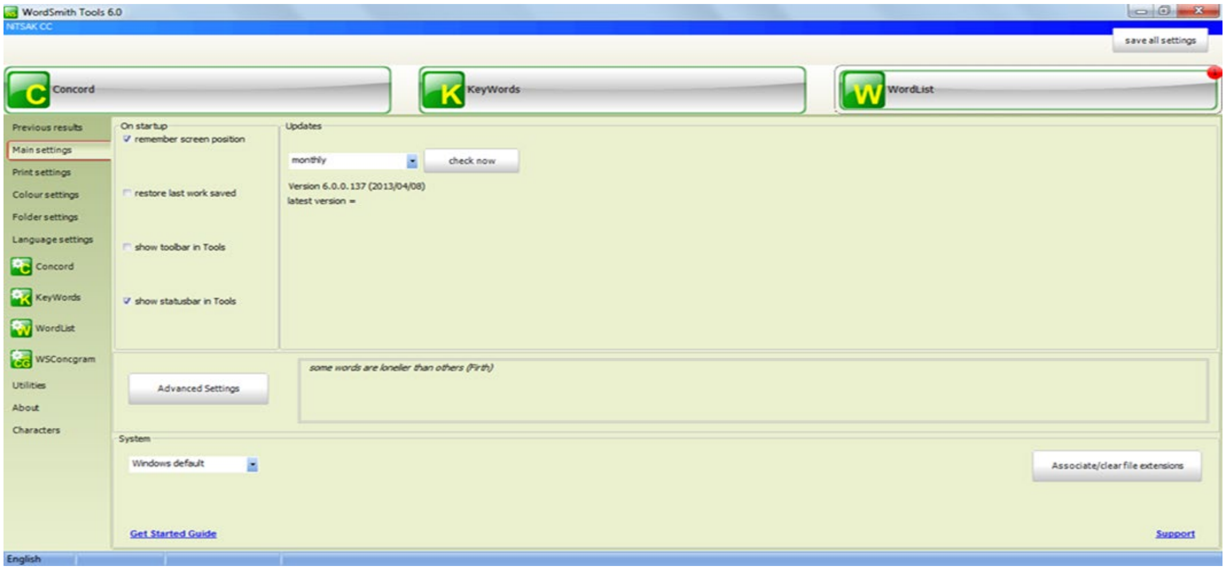
Corpus managers may be web or desktop-based, commercially marketed or open-source, and designed for specific corpora or generic.

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

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

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

Figure 1: Interface of Concordance, Keyword and Wordlist functions



On startup

☑ remember screen position

□ restore last work saved

□ show toolbar in Tools

☑ show statusbar in Tools

Updates

monthly

Version 6.0.0.137 (2013/04/08)
latest version =

some words are lonelier than others (Firth)

System

Windows default

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

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

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

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

Table 2: Acronyms used in the Wordsmith Tools program

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

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

4 Collections

The current study comprised of a small, targeted LSP corpus collected from four isiZulu literature text books, namely, Insila KaShaka by JL Dube (second male author in the history of isiZulu writing), Bafa Baphela by Joyce Jessie Gwayi (first female author in the history of isiZulu writing), Umdonsiswano by Condy Nxaba (contemporary male author), and Amandl’ Esambane by Babhekile Ngcobo (contemporary female author). P Lamula's UZulu kaMalandela was the first piece of fiction written by a Zulu person in 1924 and JL Dube's Insila kaShaka was the second in 1930. The first book by a Zulu author was Magema Fuze's Abantu abamnyama lapho abavela khona in 1922 which was not a literary work. This required that the researcher use JL Dube’s book for this study as it was easy to access. Each of these books was used as an AC. An LSP corpus is a language specific corpus, which looks into language that is restricted to a particular domain (Bowker and Pearson, 2002). The advantage of using LSP corpora is that it is easier to identify specialized terms, thus providing a large amount of information with regard to the words, frequency, structure and style in the specialised language. Language for Specific Purposes is also useful in picking up the difference between standard language and specific registers as well as neologisms (new terms) that emerge in the language (Weisser, 2016). The AC is a technical corpus, meaning that it is domain specific. The motivation for choosing these texts was basing the analysis on traditional/old Zulu and contemporary Zulu. Insila kaShaka with 19 425 tokens and Bafa Baphela with 22 103 tokens fall under the traditional era of writing whilst Umdonsiswano with 49 341 tokens and Amandl’ Esambane at 31 251 tokens follow the contemporary way of writing. The LSP corpora were used to compare gender sensitivity in traditional literature to that in contemporary literature to investigate any language shift and ensure a balance when investigating the corpus. An RC is a non-technical corpus that is designed according to Oostdijk et al., (2021) to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference material. The model for selection usually defines a number of parameters that provide for the inclusion of as many sociolinguistic variables as possible and prescribes the proportions of each text types that is selected (Oostdijk et al., 2013).

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

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

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

5 Querying

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

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

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

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

6 Discussion

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

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

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

Table 3: Concordance List of AC 2

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

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

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

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

Table 4: Concordance List of AC 3

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

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

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

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

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

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

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

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

7 Conclusion

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

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

8 References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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