# Semi-automatic Term Extraction for an isiZulu Linguistic Terms Dictionary[*]

Langa Khumalo, *Linguistics Program, School of Arts, University of KwaZulu-Natal, South Africa (khumalol@ukzn.ac.za)*

**Abstract:** The University of KwaZulu-Natal (UKZN) is compiling a series of Language for Special Purposes (LSP) dictionaries for various specialized subject domains in line with its language policy and plan. The focus in this paper is the term extraction for words in the linguistics subject domain. This paper advances the use of frequency analysis and the keyword analysis as strategies to extract terms for the compilation of the dictionary of isiZulu linguistic terms. The study uses the isiZulu National Corpus (INC) of about 1,2 million tokens as a reference corpus as well as an LSP corpus of about 100,000 tokens as a study corpus. The study is analyzed through the use of a software tool called WordSmith Tools (version 6). WordSmith Tools (hence forth WS Tools) is an integrated suite of three main programs, which include the WordList, Concord and Keywords, used in analysing words and word patterns in any given text. Using the WS Tools software a lot of qualitative and quantitative research can be done in the language. Central to this study is a computational determination of which words are typical of the linguistic domain in isiZulu and therefore stand out as preferred candidates for headword selection. Thus the study uses the corpus linguistics method as a basis for theoretical analysis. The advantage of such a theoretical approach is that a corpus is stored and queried by means of computer and computer software, which makes it easy to find, sort and count items, either as a basis for linguistic description or for addressing language-related issues and problems. Using the WS Tools software, the study shows that term extraction for the isiZulu dictionary of linguistic terms is done following reliable computational techniques in corpus lexicography.

**Keywords:** TERM EXTRACTION, LGP CORPUS, LSP CORPUS, WORDSMITH TOOLS, FREQUENCY, WORDLIST, CONCORD, KEYNESS, LEXICOGRAPHY, CORPUS LEXICOGRAPHY, HEADWORD SELECTION, LSP DICTIONARY

**Opsomming: Semi-outomatiese term-onttrekking vir 'n isiZulu taalkundige termwoordeboek.** Die Universiteit van KwaZulu-Natal (UKZN) is besig met die samestelling van 'n reeks Taal vir Spesiale Doeleindes (TSD)-woordeboeke vir verskeie gespesialiseerde vakge- biede wat strook met hul taalbeleid en -plan. Die fokus van hierdie artikel is die termontrekking vir woorde in die vakgebied taalkunde. Die gebruik van frekwensieanalise en sleutelwoordanalise as strategieë in die samestelling van die isiZulu taalkundige termwoordeboek word bevorder. Die studie gebruik die isiZulu National Corpus (INC) van ongeveer 1,2 miljoen items as 'n verwysings- korpus asook 'n TSD-korpus van ongeveer 100,000 items as 'n studiekorpus. Die studie is ontleed

---

met behulp van 'n sagteware nutsprogam, WordSmith Tools (weergawe 6). WordSmith Tools (voortaan WS Tools) is 'n geïntegreerde programsuite bestaande uit drie hoofprogramme, wat WordList, Concord en Keywords insluit, en wat gebruik word in die analise van woorde en woord-patrone in enige gegewe teks. Met behulp van die WS Tools-sagteware kan baie kwalitatiewe en kwantitatiewe navorsing in die taal gedoen word. Sentraal in hierdie studie is 'n rekenaarmatige bepaling van watter woorde verteenwoordigend is van die isiZulu-taalkundige domein en daarom voorkeur geniet by trefwoordseleksie. Sodoende word die korpuslinguistiekmetode as basis vir teoretiese analise gebruik. Die voordeel verbonde aan so 'n teoretiese benadering is dat 'n korpus gestoor en geraadpleeg word deur middel van 'n rekenaar en rekenaarsagteware, wat dit maklik maak om items te vind, te sorteer en te tel, óf as basis vir taalkundige beskrywing óf om taalkundig verwante kwessies en probleme aan te spreek. Deur gebruik te maak van WS Tools-sagteware, toon die studie dat term-ontrekking vir die isiZulu taalkundige termwoordeboek gedoen word deur betroubare rekenaarmatige tegnieke in korpusleksikografie te volg.

**Sleutelwoorde:** TERM-ONTTREKKING, TAD-KORPUS, TSD-KORPUS, WORDSMITH TOOLS, FREKWENSIE, WOORDELYS, KONGRUENSIE, SLEUTELSTATUS, LEKSIKOGRAFIE, KORPUSLEKSIKOGRAFIE, TREFWOORDSELEKSIE, TSD-WOORDEBOEK

## 1.    Introduction

The University of KwaZulu-Natal (UKZN) is compiling a series of Language for Special Purposes dictionaries for various specialized subject domains in line with its language policy and plan (Khumalo 2014: 1). The Language Policy and Plan of the University of KwaZulu-Natal (UKZN) is wholly informed by the country's widely acclaimed constitution, which enshrines multilingualism and provides that every official language must enjoy parity of esteem and must be treated equitably. In line with the provisions enshrined in the South African constitution section 6 (subsection 2 and 4), the Language in Education Policy of 1997, and consistent with the framework as set out in the Language Policy for Higher Education of 2002, and congruent with the Use of Official Languages Act of 2012, UKZN identifies with the goals of South Africa's multilingual language policy and seeks to be a key player in the successful implementation of this policy. Consequent to these statutory provisions UKZN has articulated this commitment through its Language Policy and Plan, which was first approved by Senate on the 2nd of August 2006. The Language Policy and Plan was recently revised and approved by Senate in November 2014.

UKZN has further taken a conscious and practical decision to develop isiZulu through its framework of functional bilingualism. Through this framework it recognizes English as the primary language of its academic program, and commits itself to the development and intellectualization of isiZulu to be a language of administration, teaching and learning, innovation and science. To this end, a detailed Language Plan monitored and evaluated by the University Language Board (ULB) is in place, and a practical Language Program has been set in motion by the University Language Planning and Development Office (ULPDO) in order to fully operationalize the University's Language Policy.

One of the major aims of the UKZN language policy is to achieve for isiZulu the institutional and academic status of English through providing facilities to enable the use of isiZulu as a language of learning, instruction, research and administration in the long term. As a result of these and other language policy objectives there has been a massive language development program, which is isiZulu corpus building and isiZulu terminology development, which are germane in the intellectualization of isiZulu. Work on the building of the isiZulu National Corpus (INC) started in the last quarter of 2014. The INC was piloted in November 2014 at 1, 1 million tokens and now stands at just under 2 million. Terminology development has taken place through arduous resource intensive statutory processes of consultation, verification, authentication and standardization. The terminology that has been standardized and approved by the isiZulu National Language Body include terminology for architecture, anatomy, computer science, corporate relations, environmental science, law, and nursing. A total of 1863 terms are now in the isiZulu Term Bank. The imperative to provide teaching and learning tool in the form of discipline specific dictionaries has thus been voiced. These will enhance cognitive capacity of both the staff and students in accessing otherwise complex scientific phenomenon, which hitherto have been contributing to the negative student performance. Specialized dictionaries are the ones that cover a relatively restricted set of phenomena. This type of dictionary covers the terminology of a particular subject field or discipline. It is also known as an LSP dictionary, which is short for Language for Special Purposes. In this paper we discuss term extraction for an isiZulu linguistic terms dictionary using a corpus linguistics method.

## 2.    Corpus linguistics method

The study uses the corpus linguistics method as a basis for theoretical analysis. According to Sinclair (2005) a corpus is "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." The advantage of such a theoretical approach is that "[…] a corpus [is] stored in a computer, it is easy to find, sort and count items, either as a basis for linguistic description or for addressing language-related issues and problems" (Kennedy 1998: 11). A corpus is thus a collection of naturally occurring texts derived from real life language use in either written or spoken form, which is then processed, stored and accessed by means of computers. Such a corpus is then useful as a basis for investigating language use and for developing dictionaries, spell checkers and other human language technologies (HLTs).

The approach we espouse in this study is a corpus linguistic one. We use a language for general purposes corpus (aka LGP) as a *reference corpus* (RC) and a language for special purposes (aka LSP) as an *analysis corpus* (AC). The RC is a non-technical corpus while the AC is a domain-specific, technical corpus. The

LSP corpus used in this study comprises of the two main isiZulu grammar textbooks *Uhlelo lwesiZulu*, and *Izikhali zabaqeqeshi nabafundi*, a collection of isiZulu grammar lecture notes from academics in the School of Arts and the School of Education at UKZN, and online linguistic documents in isiZulu. Using these two corpora that are quite different in terms of content, we compare the behavior of lexical units and identify lexical units that are specific to the AC.

In order to explicate the LSP corpus further, Lynne Bowker (2002: 45) states that the LSP corpus is one that "focuses on a particular aspect of a language. It could be restricted to the LSP of a particular subject field, to a specific text type, to a particular language variety or to the language used by members of a certain demographic group (e.g. teenagers). Because of its specialized nature, such a corpus cannot be used to make observations about language in general. However, general reference corpora and special purpose corpora can be used in a comparative fashion to identify those features of a specialized language that differ from general language …"

The advantages of LSP corpora are that they contain a wealth of authentic usage information. Since LSP corpora comprise of texts that have been written by subject field experts, the researchers have before them a body of evidence pertaining to the function and usage of words and expressions in the LSP of the field. With the help of corpus analysis tools, it becomes possible to sort these contexts so that meaningful patterns are revealed. An LSP corpus basically contains thousands of words that have been written by subject field experts and, as such, it can be seen to represent distilled expert knowledge.

The RC used in this study is an LGP corpus with 1 186 675 running words. The size of the RC, although still modest, can guarantee that the articles cover a wide range of subjects and that their content is heterogeneous. In contrast the AC is discipline-specific with an exclusive focus on linguistics. The AC is an LSP corpus with 111 922 running words, which comprises two isiZulu basic grammar textbooks and a collection of notes on the teaching of isiZulu grammatical structure.

Our study is analyzed through the use of a software tool called Word-Smith Tools (version 6). WordSmith Tools (hence forth WS Tools) is an integrated suite of three main programs, which include the WordList, Concord and Keywords, used in analysing words and word patterns in any given text. WS Tools was developed by Mike Scott, who had earlier worked with Tim Jones to develop MicroConcord. WS Tools was first released in 1996 and the current version (version 6.0) was released in 2011. The Wordlist tool can be used to produce wordlists or word-cluster lists from a text and render the results alphabetically or by frequency order. It can also calculate word spread across a variety of texts. The Concord tool can give any word or phrase in context – so that one can study its co-text, i.e. to see what other words occur in its vicinity. The Keywords tool calculates words which are key in a text, i.e., used much more frequently or much less frequently in a given corpus (e.g. the LSP corpus) than expected in terms of a general corpus of the language (e.g. the INC). Using

the WS Tools software a lot of qualitative and quantitative research can be done in the language. Table 1 below shows the main features of the WS Tools as described above.

**Table 1:**     Wordsmith Tools (version 6)
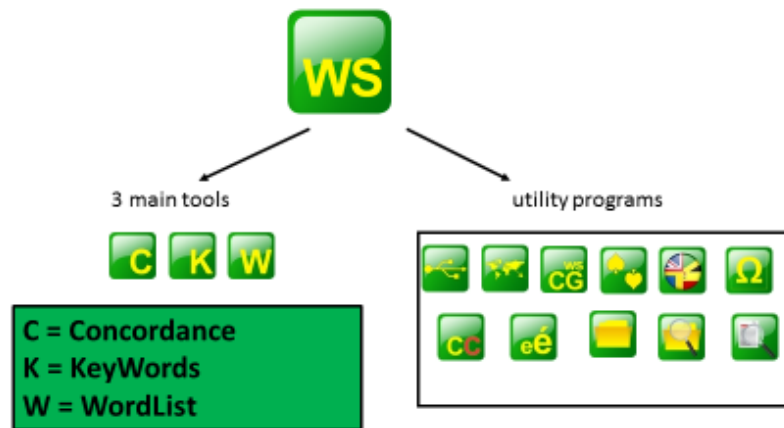


Central to this study is thus a computational determination of which words are typical of the linguistic domain in isiZulu and therefore stand out as preferred candidates for headword selection. Using the WS Tools software, the study will perform the following. The author will run a frequency list to determine the most frequent words in the LSP corpus. A frequency list provides an array of different types of words, tokens, or forms which make up a corpus. These can be listed from the most frequent token to *hapax legomena* (i.e. those forms that occur only once in a given corpus) or vice versa. Frequency lists are a powerful tool in corpus lexicography. They guide lexicographers on which words to include in a dictionary. Frequency lists also provide developers of second language teaching material with the most relevant words, phrases, and expressions to teach. In this study a frequency list sheds 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."

### 3.     Term extraction

The focus in this study is the term extraction for words in the linguistics subject domain. Term extraction means the automatic mining or retrieval of relevant terms from a given corpus. Term extraction remains a challenge to anyone interested in domain-specific information retrieval (Jacquemin 2001); (Bouri-

gault et al. 2001); (Drouin n.d.). The goal in this study is to extract words that are typical for the isiZulu linguistic domain. We use the keyword tool in WS version 6 to extract linguistic terms. The main goal is to reduce (not eliminate) the amount of noise in the list of candidate terms.

## 4.    Frequency analysis

It is crucial to affirm the observation by Summers (1996: 261) that "all aspects of lexicography are influenced by frequency." This is particularly crucial in selecting word candidates for inclusion in a dictionary. Headword selection becomes informed by the frequency through a statistical analysis. We premise our analysis on the most frequent 100 words on the assumption that this would be the most typically used words. The word list flows from the most frequent word to the least frequent in a descending order. The most frequent words in the AC are given in Table 2. *N* stands for the number the word occupies in the list of words in the word list, and *Freq.* is the number of times a word occurs in the corpus.

**Table 2:**    Most frequent 100 tokens

| N | Word | Freq. | N | Word | Freq. |
|---|------|-------|---|------|-------|
| 1 | ukuthi | 861 | 51 | bona | 67 |
| 2 | noma | 812 | 52 | emva | 67 |
| 3 | bese | 512 | 53 | mina | 66 |
| 4 | kodwa | 481 | 54 | kubo | 64 |
| 5 | lapho | 421 | 55 | ziye | 63 |
| 6 | futhi | 419 | 56 | indawo | 62 |
| 7 | ngoba | 409 | 57 | kule | 62 |
| 8 | nje | 353 | 58 | kwezinye | 62 |
| 9 | ke | 342 | 59 | nayo | 62 |
| 10 | ukuba | 296 | 60 | kusho | 59 |
| 11 | lokhu | 279 | 61 | ngenhla | 59 |
| 12 | khona | 262 | 62 | nokuthi | 59 |
| 13 | phela | 255 | 63 | yini | 59 |
| 14 | naye | 236 | 64 | ala | 58 |
| 15 | ngo | 236 | 65 | izakhi | 58 |
| 16 | kanti | 231 | 66 | nazo | 58 |
| 17 | kanye | 213 | 67 | wena | 57 |
| 18 | ngaye | 190 | 68 | leli | 56 |
| 19 | lapha | 189 | 69 | isimo | 55 |
| 20 | kahle | 187 | 70 | lesi | 54 |
| 21 | no | 178 | 71 | laba | 53 |
| 22 | zonke | 157 | 72 | zona | 53 |
| 23 | njengoba | 152 | 73 | ngazo | 52 |

| 24 | ake | 148 | | 74 | uhlelo | 52 |
|----|-----|-----|-|----|--------|----|
| 25 | sithi | 148 | | 75 | wonke | 52 |
| 26 | kuye | 147 | | 76 | enye | 51 |
| 27 | na | 138 | | 77 | lezo | 51 |
| 28 | ukuze | 137 | | 78 | zakhe | 51 |
| 29 | lezi | 132 | | 79 | lolu | 50 |
| 30 | kanje | 131 | | 80 | nga | 50 |
| 31 | ngokuthi | 130 | | 81 | thina | 50 |
| 32 | lusizo | 121 | | 82 | yona | 49 |
| 33 | usuke | 117 | | 83 | nazi | 48 |
| 34 | ngayo | 116 | | 84 | ngaso | 48 |
| 35 | kube | 115 | | 85 | ngakho | 46 |
| 36 | kuthi | 110 | | 86 | yena | 45 |
| 37 | ngabe | 89 | | 87 | kuze | 44 |
| 38 | lo | 87 | | 88 | kude | 43 |
| 39 | ngu | 87 | | 89 | kulo | 43 |
| 40 | manje | 85 | | 90 | kuwo | 43 |
| 41 | uye | 82 | | 91 | nabo | 43 |
| 42 | ba | 80 | | 92 | aba | 42 |
| 43 | kanjani | 80 | | 93 | kepha | 41 |
| 44 | lokho | 76 | | 94 | uzobe | 41 |
| 45 | yakhe | 75 | | 95 | konke | 40 |
| 46 | yonke | 73 | | 96 | siye | 40 |
| 47 | njalo | 72 | | 97 | kuzo | 38 |
| 48 | lowo | 71 | | 98 | labo | 38 |
| 49 | bonke | 70 | | 99 | sakhe | 38 |
| 50 | baye | 67 | | 100 | sika | 38 |

Table 2 shows that the ten most frequent words in the AC are *ukuthi*, *noma*, *bese*, *kodwa*, *lapho*, *futhi*, *ngoba*, *nje*, *ke*, and *ukuba*. All these words are function or grammatical words, which belong to a closed word class. The closed word classes include concords, pronouns, numerals, connectives etc. This top ten word list is not unique as function words commonly dominate all frequency lists. It is therefore the case that functional words are normally removed from the word list in order to retain content words. Table 3 below shows the list of the most frequent 100 tokens after excluding the function words.

**Table 3:**    Most frequent 100 tokens excluding function words

| N | Word | Freq. | | N | Word | Freq. |
|---|------|-------|-|---|------|-------|
| 1 | u | 829 | | 51 | lusizo | 121 |
| 2 | e | 550 | | 52 | usuke | 117 |
| 3 | lapho | 421 | | 53 | ngayo | 116 |

| | | | | | | |
|---|---|---|---|---|---|
| 4 | ngoba | 409 | 54 | kube | 115 |
| 5 | isibonelo | 387 | 55 | la | 114 |
| 6 | nje | 353 | 56 | le | 111 |
| 7 | ke | 342 | 57 | onkamisa | 111 |
| 8 | ukuba | 296 | 58 | kuthi | 110 |
| 9 | ulimi | 290 | 59 | isakhi | 104 |
| 10 | lokhu | 279 | 60 | ndlela | 101 |
| 11 | khona | 262 | 61 | umntwana | 101 |
| 12 | amagama | 260 | 62 | izibonelo | 100 |
| 13 | o | 257 | 63 | kolimi | 100 |
| 14 | phela | 255 | 64 | leyo | 100 |
| 15 | naye | 236 | 65 | abanye | 99 |
| 16 | kanye | 213 | 66 | isuke | 99 |
| 17 | indlela | 204 | 67 | kuphela | 99 |
| 18 | umuntu | 201 | 68 | yolimi | 98 |
| 19 | kukhona | 196 | 69 | izenzo | 96 |
| 20 | ubunye | 191 | 70 | izib | 96 |
| 21 | ngaye | 190 | 71 | ezinye | 95 |
| 22 | njll | 190 | 72 | isabizwana | 95 |
| 23 | isigaba | 189 | 73 | ngaphandle | 95 |
| 24 | lapha | 189 | 74 | into | 94 |
| 25 | kahle | 187 | 75 | iziqu | 94 |
| 26 | unkamisa | 180 | 76 | umakoti | 94 |
| 27 | kakhulu | 173 | 77 | zisuke | 90 |
| 28 | abantu | 163 | 78 | ngabe | 89 |
| 29 | zonke | 157 | 79 | abe | 88 |
| 30 | ubuningi | 154 | 80 | umusho | 88 |
| 31 | njengoba | 152 | 81 | lo | 87 |
| 32 | ake | 148 | 82 | ngu | 87 |
| 33 | sithi | 148 | 83 | imisindo | 86 |
| 34 | kuye | 147 | 84 | izintombi | 86 |
| 35 | isenzo | 143 | 85 | ana | 85 |
| 36 | amabizo | 142 | 86 | manje | 85 |
| 37 | kusuke | 142 | 87 | ongwaqa | 85 |
| 38 | phakathi | 139 | 88 | ubaba | 84 |
| 39 | na | 138 | 89 | umoya | 84 |
| 40 | ibhola | 137 | 90 | kuba | 83 |
| 41 | igama | 137 | 91 | kufanele | 83 |
| 42 | ukuze | 137 | 92 | uye | 82 |
| 43 | lezi | 132 | 93 | ekhaya | 81 |
| 44 | kanje | 131 | 94 | eqondisayo | 81 |
| 45 | ibizo | 130 | 95 | ongenazwi | 81 |
| 46 | ngokuthi | 130 | 96 | ba | 80 |

| 47 | umfana | 129 | 97 | kanjani | 80 |
| 48 | ingane | 127 | 98 | ukusetshenziswa | 80 |
| 49 | emshweni | 126 | 99 | izivumelwano | 79 |
| 50 | inkathi | 122 | 100 | isib | 77 |

Table 3 shows the same data as Table 2 with the exclusion of function words. The removal of function words reveals content words that could define the genre. The list of content words reveals clearly the genre of linguistics. For example *u, e, o;* (vowels); *isibonelo* (example); *ulimi* (language), *amabizo* (nouns); *indlela* (mood), *ubunye* (singular) etc. are typical linguistic words. The frequency list has somewhat helped to isolate words that are typical. Other words on the top 100 wordlist are not particular to the discipline. Such words include *ngoba*, *umuntu*, *ngaye* and others. This is not unusual since the top 100 words are not isolated on any measure that isolates words that are typical to a text. In order to achieve this we use the keyword analysis.

## 5.     Keyword analysis

We use the keyword analysis in order to identify words particular to the isiZulu linguistics domain. This is done through the calculation of keyness, which isolates words which are key to the AC. According to Mike Scott (2006: 92) keyness is "calculated by comparing the frequency of each word in the word list of the text under investigation with the frequency of the same word in the reference word list." Calculations are done using the Keyword tool of WS Tools. The output is a list of keywords, or words whose frequencies are higher in the AC than in the RC. Table 4 below shows the top 100 words most typical in the linguistic domain extracted through the Keyness tool.

**Table 4:**     Top 100 linguistic tokens

| N | Keyword | English gloss | Freq. | Keyness |
|---|---|---|---|---|
| 1 | isibonelo | example | 387 | 1515,82 |
| 2 | i | vowel *i* | 1002 | 1424,26 |
| 3 | a | vowel *a* | 1005 | 1172,94 |
| 4 | bese | and | 512 | 875,18 |
| 5 | ulimi | language | 290 | 773,57 |
| 6 | uma | if | 1179 | 659,00 |
| 7 | — | — | — | — |
| 8 | — | — | — | — |
| 9 | unkamisa | vowel | 180 | 557,61 |
| 10 | phela | finish | 255 | 510,56 |
| 11 | e | vowel *e* | 550 | 488,01 |
| 12 | njll | etc. | 190 | 485,03 |
| 13 | u | vowel *u* | 829 | 473,92 |
| 14 | ubunye | singular | 191 | 465,09 |

| 15 | emshweni | in sentence | 126 | 423,19 |
|----|----------|-------------|-----|--------|
| 16 | isigaba | noun class | 189 | 413,95 |
| 17 | kusuke | from | 142 | 400,56 |
| 18 | ongenazwi | voiceless | 81 | 392,36 |
| 19 | ibizo | noun | 130 | 374,68 |
| 20 | amabizo | nouns | 142 | 368,78 |
| 21 | amagama | words | 260 | 365,93 |
| 22 | yolimi | linguistic | 98 | 364,73 |
| 23 | ubuningi | plural | 154 | 361,18 |
| 24 | onkamisa | vowels | 111 | 357,17 |
| 25 | izibonelo | examples | 100 | 356,86 |
| 26 | kolimi | linguistic | 100 | 356,86 |
| 27 | isakhi | morpheme | 104 | 351,84 |
| 28 | zisuke | from | 90 | 350,20 |
| 29 | isuke | from | 99 | 349,82 |
| 30 | umusho | sentence | 88 | 341,03 |
| 31 | usuke | from | 117 | 329,89 |
| 32 | inkathi | tense | 122 | 324,55 |
| 33 | isenzo | verb | 143 | 322,38 |
| 34 | noma | or | 812 | 313,96 |
| 35 | umakoti | bride | 94 | 309,01 |
| 36 | onezwi | voiced | 63 | 303,29 |
| 37 | zenkulumo | of speech | 73 | 299,87 |
| 38 | o | vowel $o$ | 257 | 295,88 |
| 39 | ongwaqa | consonants | 85 | 293,59 |
| 40 | iziqu | stem | 94 | 290,38 |
| 41 | usizo | help | 121 | 281,57 |
| 42 | konkamisa | on vowels | 74 | 280,32 |
| 43 | isabizwana | substantive | 95 | 279,64 |
| 44 | imisindo | sounds | 86 | 273,14 |
| 45 | umkhongi | negotiator | 54 | 268,64 |
| 46 | intombi | girl | 52 | 258,69 |
| 47 | isib | e.g. | 77 | 256,67 |
| 48 | umfana | boy | 129 | 246,03 |
| 49 | ngaye | through him | 190 | 239,20 |
| 50 | abantu | people | 48 | 238,79 |
| 51 | iqhikiza | full-grown girl | 53 | 235,40 |
| 52 | izib. | e.gs | 96 | 230,55 |
| 53 | eqondisayo | inductive mood | 81 | 225,76 |
| 54 | ukusetshenziswa | used | 80 | 223,45 |
| 55 | izakhi | morphemes | 58 | 223,20 |
| 56 | basuke | left | 76 | 222,65 |
| 57 | izib | e.gs | 93 | 220,98 |
| 58 | inkomo | cows | 70 | 220,66 |
| 59 | izivumelwano | agreements | 79 | 219,54 |
| 60 | unsinini | alveolar | 46 | 219,34 |
| 61 | sokukhomba | demonstrative | 69 | 218,52 |
| 62 | yenkulumo | of speech | 68 | 218,48 |
| 63 | isibanjalo | copulative | 68 | 212,35 |
| 64 | ana | reciprocal suffix | 85 | 212,06 |
| 65 | izintombi | girls | 86 | 211,83 |

| 66 | ziye | gone | 63 | 201,67 |
|---|---|---|---|---|
| 67 | ingane | child | 127 | 201,46 |
| 68 | ungwaqabathwa | click sounds | 42 | 199,62 |
| 69 | zamabizo | nominal | 57 | 196,35 |
| 70 | isandiso | locative | 64 | 195,85 |
| 71 | imisho | sentences | 63 | 195,60 |
| 72 | sithi | we say | 148 | 190,00 |
| 73 | qaphela | note | 65 | 189,02 |
| 74 | isiqalo | prefix | 63 | 188,05 |
| 75 | zesenzo | of verbs | 48 | 187,20 |
| 76 | isiqu | stem | 66 | 184,66 |
| 77 | indlela | mood | 204 | 179,69 |
| 78 | onguputshu | plosive | 36 | 179,09 |
| 79 | ngonkamisa | are vowels | 62 | 178,77 |
| 80 | umgudu | cavity | 54 | 176,52 |
| 81 | ukwakhiwa | morphology | 61 | 171,50 |
| 82 | ukulandula | negation | 58 | 171,44 |
| 83 | izenzo | verbs | 96 | 170,55 |
| 84 | izilimi | languages | 71 | 165,12 |
| 85 | umkhwenyana | bridegroom | 42 | 162,97 |
| 86 | udwendwe | que | 34 | 154,04 |
| 87 | iphimbo | tone | 56 | 153,57 |
| 88 | sesenzo | verbal | 48 | 153,35 |
| 89 | izibanjalo | copulatives | 47 | 151,25 |
| 90 | zabomdabu | of tradition | 33 | 144,03 |
| 91 | baye | gone | 67 | 142,51 |
| 92 | ibhola | ball | 137 | 141,52 |
| 93 | emabizweni | in nouns | 44 | 140,74 |
| 94 | izingcezu | morphemes | 44 | 140,74 |
| 95 | sebizo | nominal | 45 | 138,87 |
| 96 | senhloko | subjectival | 49 | 135,74 |
| 97 | zezenzo | verbal | 48 | 135,02 |
| 98 | ndlela | mood | 101 | 134,62 |
| 99 | intombazane | girl | 27 | 134,32 |
| 100 | esuke | from | 39 | 132,81 |

## 6.     Discussion

The 100 keywords in Table 4 are a more typical reflection of the linguistics discipline when juxtaposed with those in Table 3. The keyness tool has successfully extracted terms which are key to the domain of linguistics from the corpus. The list includes the vowels *a, e, i, o, u,* **(3, 11, 2, 38, 13)**; language *ulimi* **(5)**; vowel *unkamisa* **(9)**; singular *ubunye* **(14)**, in a sentence *emshweni* **(15)**; noun class *isigaba* **(16)**, voiceless *ongenazwi* **(18)**; noun *ibizo* **(19)** nouns *amabizo* **(20)**; consonants *ongwaqa* **(39)**; indicative mood *eqondisayo* **(53)**; agreements *izivumelwano* **(59)**; copulative *isibanjalo* **(63)** click sound *ungwaqabathwa* **(68)**; cavity *umgudu* **(80)**; tone *iphimbo* **(87)**; subjectival *senhloko* **(96)**; **etc.**

The top 100 wordlist suggests that the keyness analysis is crucial in iso-

lating data that is domain specific. The results of these experiments are useful as potential candidates for headword selection are highlighted. The study has shown that term extraction for the isiZulu dictionary of linguistic terms is done following reliable computational techniques in corpus lexicography.

## 7.    Conclusion

We explored frequency and keyword analysis in generating domain specific candidates for headword selection. Using such statistical approach is faster, reliable and free from human error or bias. It is clear from the study that corpora are useful in enhancing the dictionary microstructure and the keyness list will form the basis for headword selection for the isiZulu linguistics terms dictionary. Term extraction thus reduces the amount of noise in the list of candidate terms. Native speaker intuition is used to compliment this vital computational resource.

## References

**Bourigault, D. et al.** 2001. *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins.

**Jacquemin, C.** 2001. Spotting and Discovering Terms through Natural Language Processing. Cambridge, MA: MIT Press.

**Kennedy, G.D.** 1998. *An Introduction to Corpus Linguistics*. London/New York: Longman.

**Khumalo, L.** 2014. *Developing an isiZulu Dictionary of Linguistics Terms: Challenges and Prospects.* Unpublished paper presented at the Nineteenth Annual International Conference of the African Association for Lexicography (AFRILEX), which was hosted by the Research Unit for Language and Literature in the SA Context, North-West University, Potchefstroom Campus, Potchefstroom, South Africa, 1–3 July 2014.

**Kilgarriff, A.** 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10(2): 135-155.

**Scott, M.** 2004–2006. *Oxford WordSmith Tools Version 4.* Oxford: Oxford University Press.

**Sinclair, J.** 2005. Corpus and Text: Basic Principles. Wynne, M. (Ed.). *Developing Linguistic Corpora: A Guide to Good Practice*: 1-16. Oxford: Oxbow Books. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 20 October 2005].

**Summers, D.** 1996. Computer Lexicography: The Importance of Representativeness in Relation to Frequency. Thomas, J. and M. Short (Eds.). 1996. *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech*: 260-266. London/New York: Longman.