



click for updates

Using ParaConc to extract bilingual terminology from parallel corpora: A case of English and Ndebele

Author:

Ketiwe Ndhlovu¹

Affiliation:

¹Department of Linguistics and Modern Languages, University of South Africa, South Africa

Corresponding author:

Ketiwe Ndhlovu,
ndhlok1@unisa.ac.za

Dates:

Received: 28 Jan. 2016

Accepted: 14 June 2016

Published: 26 Oct. 2016

How to cite this article:

Ndhlovu, K., 2016, 'Using ParaConc to extract bilingual terminology from parallel corpora: A case of English and Ndebele', *Literator* 37(2), a1278. <http://dx.doi.org/10.4102/lit.v37i2.1278>

Copyright:

© 2016. The Authors.
Licensee: AOSIS. This work is licensed under the Creative Commons Attribution License.

The development of African languages into languages of science and technology is dependent on action being taken to promote the use of these languages in specialised fields such as technology, commerce, administration, media, law, science and education among others. One possible way of developing African languages is the compilation of specialised dictionaries (Chabata 2013). This article explores how parallel corpora can be interrogated using a bilingual concordancer (ParaConc) to extract bilingual terminology that can be used to create specialised bilingual dictionaries. An English–Ndebele Parallel Corpus was used as a resource and through ParaConc, an alphabetic list was compiled from which headwords and possible translations were sought. These translations provided possible terms for entry in a bilingual dictionary. The frequency feature and 'hot words' tool in ParaConc were used to determine the suitability of terms for inclusion in the dictionary and for identifying possible synonyms, respectively. Since parallel corpora are aligned and data are presented in context (Key Word in Context), it was possible to draw examples showing how headwords are used. Using this approach produced results quickly and accurately, whilst minimising the process of translating terms manually. It was noted that the quality of the dictionary is dependent on the quality of the corpus, hence the need for creating a representative and clean corpus needs to be emphasised. Although technology has multiple benefits in dictionary making, the research underscores the importance of collaboration between lexicographers, translators, subject experts and target communities so that representative dictionaries are created.

Die gebruik van ParaConc om tweetalige terminologie van parallel korpusse te onttrek: 'n Geval van Engels en Ndebele. Die ontwikkeling van die Afrikatale as wetenskap- en tegnologietale hang af van wat gedoen word om die gebruik van hierdie tale in gespesialiseerde domeine soos die tegnologie, handel, administrasie, die media, regte, wetenskap en onderwys te bevorder. Een moontlike manier waarop die Afrikatale ontwikkel kan word, is deur vakwoordeboeke saam te stel (Chabata 2013). Hierdie artikel verken die manier waarop parallelle korpusse met 'n tweetalige konkordanser (ParaConc) deursoek kan word vir tweetalige terme wat dan gebruik kan word om tweetalige vakwoordeboeke saam te stel. 'n Engels-Ndebele Parallelle Korpus het as bron gedien en ParaConc is gebruik om 'n alfabetiese lys saam te stel waarvoor vertalings verskaf is. Hierdie vertalings het moontlike terme verskaf wat in 'n tweetalige woordeboek opgeneem kan word. Die frekwensielys in ParaConc is tesame met sy 'hot words'-instrument aangewend om gepaste terme te bepaal wat in die woordeboek opgeneem kan word en om ook moontlike sinonieme te identifiseer. Aangesien parallelle korpusse bely word, en alle data in konteks (*Key Word in Context*) verskyn, was dit moontlik om voorbeelde uit te lig om aan te toon hoe trefwoorde gebruik kan word. Met hierdie benadering was dit moontlik om resultate baie vinnig en akkuraat te bekom terwyl die vertaling van terme met die hand bykans uitgeskakel word. Daar word aangetoon dat die gehalte van die woordeboek duidelik afhang van die gehalte van die korpus en derhalwe word die behoefte aan 'n verteenwoordigende en skoon korpus beklemtoon. Alhoewel woordeboekmaak op veelvoudige maniere baat vind by moderne tegnologie, beklemtoon die navorsing ook die belangrikheid van samewerking tussen leksikograwe, vertalers, vakkeners en teikengebruikers sodat verteenwoordigende woordeboeke geskep kan word.

Introduction

This study is rooted in the language debate that is currently dominating the African continent regarding the use of African languages in specialised arenas such as technology, commerce, administration, media, law, science and education among others. Generally, it is argued that African languages are not ready to be used in specialised fields owing to their lack of development and

Read online:



Scan this QR code with your smart phone or mobile device to read online.

scarcity of terminology. Chabata (2013:54) succinctly expresses the predominant question: Do the indigenous African languages, in their current state of development, have the necessary capacity required of them to be used in all spheres of life, including those that are highly technical? In response to this question, he says a 'yes' or 'no' does not suffice because African languages have different statuses in different countries and are therefore at different levels of development.

Although some African languages are more developed than others, research has proven over and over again that many African languages struggle to express scientific and technical terms (Gauton & De Schryver 2004; Gauton, Taljard & de Schryver 2013; Kruger 2010; Moropa 2005; Ndhlovu 2014; Trew 1994; Van Huyssteen 1999; Wallmach & Kruger 1999). Unfortunately, technology continues to develop at an unimaginable rate, compounding the problem for many languages that have to develop new terms on a continuous basis to keep up with the technological developments. This state of affairs points to a need to develop African languages to a stage where they are usable in both the private and public spheres. This is because the development of Africa is dependent on its citizens partaking fully in developmental issues and this can happen only if people understand fully what is communicated to them. Mohochi (2006), Prah (2008) and Chabata (2013) just like many other African scholars argue that the development of Africa cannot be achieved without recognising the central role of indigenous African languages in the socioeconomic, political, educational and technological processes. In response to the current debate on the use of African languages in specialised arenas, this study, like many others before, calls for the development and use of African languages in specialised fields. One way of developing African languages is through the compilation of specialised dictionaries (Chabata 2013:51).

This article supports the proposal by Chabata (2013) that specialised dictionaries be compiled to expand the use of African languages into important areas from which they are currently excluded. The study however goes further to practically show how bilingual dictionaries can be created from parallel corpora with the help of bilingual concordancers. In other words, this article suggests that African lexicographers utilise specialised parallel and multilingual corpora that are created by translators as resources to extract bilingual terminology through the use of bilingual and multilingual concordancers. This approach is presented as an alternative method to creating bilingual specialised dictionaries in Africa. Compared with the manual approach of dictionary making, where terms are collected from the public, selected, typed in, tagged and translated among other things, this approach is faster, efficient, cost-effective and produces more accurate results as shall be shown later. In the next section, it is explained why bilingual specialised dictionaries are important in Africa.

Why bilingual specialised dictionaries?

Specialised dictionaries are dictionaries that deal with terminologies in a particular field. The main aim of these dictionaries is to describe vocabularies or specialised or

technical subject fields in an exhaustive manner and to support the needs of lay people, semi-specialists and specialists in the respective fields of knowledge (Chabata 2013:55). Specialised dictionaries therefore act as repositories of specialised language and have a standardising function. That is, the dictionaries have an important normative and standard-setting influence because the users accept and apply the lexicographer's descriptions of word forms and statements about their meanings, standard spelling and pronunciation (Chimhundu 2006:246). The compilation of these dictionaries therefore has the capacity to play a crucial role in the elaboration and intellectualisation of African languages. Through specialised dictionaries, African languages may have a better chance of being scientific and technological languages.

Although specialised dictionaries play an important role in the development of specialised terminology and African languages, this genre unfortunately has received the least attention. Nkomo (2010:372) explains that:

in the indigenous African languages, the LSP (Languages for Specific Purposes) dictionary genre has thus far received the least attention, with lexicographers, linguists and language planners giving more attention to general dictionaries, which are viewed as language standardisation and documentation tools.

Thus more focus has been on the documentation and preservation of the languages rather than on their use. The rationale behind the production of general dictionaries at the expense of specialised dictionaries has been that general dictionaries are usually considered as standardising tools compared with specialised dictionaries because of their inclusive nature and wider usage, but in fact both general and specialised dictionaries contribute to the standardisation of language, though in different spheres, general and specialised, respectively. Still, preservation of language is important, but there is a need for Africa to go beyond preservation to meet the everyday needs of the users. That is, there is a need to develop dictionaries that take into cognisance the needs of African people in a globalised and technological world and bilingual specialised dictionaries can fill that gap. If properly developed, specialised dictionaries can act as reference tools for translators, translation students, language professionals, subject experts and the general public, hence this study.

With regard to bilingual dictionaries, these are important resources because they help users to understand new concepts in a language. They can assist dictionary users by acting as pointers to the link between the older and well-known vocabulary in a European language and the newer and less familiar ones in the African language (Chabata 2013:57). Furthermore, bilingual dictionaries are important as resources for translators when searching for translation terms in various language combinations. They also play a crucial role in making specialised terminology available to language users, and lastly they contribute to the standardisation of specialised terminology in indigenous languages. Nkomo (2010:373) advises that just like parallel corpora, LSP dictionaries in indigenous African languages need to be seen as potentially useful tools and resources that may solve

problems faced in the development and acquisition of specialised languages as well as the translation of specialised texts. Focusing specifically on the Ndebele language (spoken in Zimbabwe, often referred to as Northern Ndebele), which is the language used as a case study, the language has only one bilingual dictionary: *an English-Ndebele bilingual dictionary* which was developed by Pelling in 1971. The dictionary is general and is of little use in specialised fields as it was developed prior to global and technological advancements; hence, the need to compile bilingual specialised dictionaries in this language. The subsequent section shows the development of this methodology.

Extracting bilingual terminology from parallel corpora

The development of this methodology can be traced back to machine translations. Fung (1998) explains that the very trend of using bilingual parallel corpora for machine translations was started by Jelinek's group formerly at IBM. Their work and others that followed are based on the conjecture that there must always exist a parallel corpus between any pair of languages for which *mutual translation is important enough*. From that point of development in the 80s, a sizeable number of research projects were carried out in the field, hence Li and Gaussier (2010:1) state that '... researchers have tried since the end of the 80s, to automatically extract bilingual lexicons from parallel corpora' (see Chen 1993; Kay & Röscheisen 1993; Melamed 1997a, 1997b, for early work). What is notable about the research at the time is that terms could be extracted faster with better accuracy and at a cheaper cost, proving that parallel corpora are good resources for extracting bilingual terminology.

The method of extracting bilingual terms continues to be of use in many Western countries, though the past decade saw some scholars moving beyond parallel corpora to non-parallel corpora and Wikipedia as resources for extracting bilingual terminology. The main reason for this shift towards non-parallel corpora and Wikipedia is the availability of data. Fung (1998:5) explains:

With the advent of internet technology and the World Wide Web, it has become obvious that such type of non-parallel, comparable corpora are more abundant, more up-to-date, more accessible than parallel corpora. Such corpora can be collected easily from downloading electronic copies of newspapers, journals, articles and even books from the World Wide Web. At any given time, there are a lot more comparable corpora available than parallel corpora.

This shift has seen a lot of research being carried out, for example, Guinovart and Simoes (2009) extracted bilingual terminology in two language pairs: English–Galician and English–Portuguese using the *Unesco Corpus – which is part of the CLUVI Parallel corpus*. The results of their study showed a high degree of accuracy of terminology extraction based on probabilistic translation dictionaries complemented by syntactic patterns. Yu and Tsujii (2009b) extracted bilingual

terminology for dictionary making from Wikipedia. Erdmann *et al.* (2010) also extracted bilingual terminology from Wikipedia using a SVM classifier. Their main argument is that Wikipedia and other large multilingual encyclopaedias are good sources for bilingual terminology extraction to complement bilingual dictionaries.

Morin and Prochasson (2011) extracted bilingual lexicon from comparable corpora enhanced with parallel corpora. Their results showed that this simple bilingual lexicon, when combined with a general dictionary, helps to improve accuracy of single word alignments. Li and Gaussier (2010) after noting the importance of developing quality corpora developed a strategy to improve the quality of a comparable corpus so as to improve the quality of the extracted bilingual lexicon. Their study is of significance because it focuses on the importance of developing a high-quality corpus for improved results. Fung (1998) investigated how noisy parallel corpora and non-parallel yet comparable corpora can be used to extract bilingual lexicon. The results showed 55.35% precision from a small corpus and 89.93% precision from a large corpus. From this research, what is notable about extracting terms from parallel or non-parallel corpora is that there is a higher level of efficiency and accuracy, hence this research that aims to show that African languages can benefit from the use of parallel and non-parallel corpora as resources. In other words, African scholars should utilise the advantages brought about by technology to create specialised dictionaries in a faster and more efficient manner.

The main reason for choosing parallel corpora as resources is that a quantifiable number of parallel corpora exist in South Africa; also, expertise to develop specialised parallel corpora is available and the abundance of translations that are produced daily in South Africa ensures that there is raw material available to create parallel corpora.

Methodology

This article relied on an English-Ndebele Parallel Corpus (ENPC) to extract bilingual terminology for the creation of an English–Ndebele medical dictionary. The ENPC is a corpus comprising English-source texts and equivalent Ndebele-target texts. The corpus was created by Ndhlovu (2012) from multiple translations collected from non-governmental and governmental organisations. All the texts were medical texts, which makes the corpus a specialised medical corpus. The ENPC comprises 85 391 words, with 48 452 words in the English file and 36 938 words in the Ndebele file, and it was firstly aligned at word, phrase, sentence and paragraph level to make it easier to identify English-source terms and their Ndebele translations. During the creation of the corpus, the Finnish language was used to represent Ndebele as ParaConc does not cater for all African languages. Thus, some diagrams feature the Finnish language. Some diagrams show the English and Ndebele combination because that is the name used in the workspace.

Prior to usage, the corpus was cleaned for the purpose of this study so as to obtain better results. The cleaning process affected the word count of the original corpus. The approach that is used to extract data can be presented diagrammatically as follows:

As shown in Figure 1 below, term extraction took place at a word and sentence level, from both the English and Ndebele texts using ParaConc. The *search* facility was used to identify source words and possible translations, by typing in the search word and the possible translations being provided in the window below. The Key Word in Context (KWIC) was used to identify contextual usage of words and the hot words list was used to determine term frequency in relation to the source word. Themes were developed based on lexicographic requirements for developing bilingual dictionaries and these include developing an alphabetic list, selecting specialised terms, searching for entry words using the search feature, identifying possible translations, selecting the best term using frequency as a basis and gathering examples via the KWIC feature. In simpler terms, the study showed the process of extracting bilingual data that can be included in a bilingual dictionary. Since this study focuses on term extraction from a parallel corpus using ParaConc, it is important to explain what a parallel corpus and what ParaConc are.

What is a parallel corpus?

A parallel corpus is a body of texts that are available in two (or more) languages, either as an original text with its translation(s) or as texts that deal with the same subject and were created within similar contexts (Kenny 1998:62). In other words, a parallel corpus comprises two texts with the target text being a translation of the source text. Baker (1995:235) and Moropa (2005:27) elaborate that the two components should cover a similar domain, various language and time spans and be of comparable length. Parallel and comparable corpora are important to translation studies because as Aijmer and Altenberg (in McEnery & Xiao 2007:1) state, 'they offer specific uses and possibilities' for contrastive and translation

studies. That is, parallel corpora are a good basis for studying similarities and differences between languages, as well as providing a basis for understanding the possible options of translations for specific terms, phrases and sentences. Parallel corpora are also advantageous in that they can always be updated. Bowker and Pearson (2002:21) explain that their electronic form means that corpora can be larger and more up-to-date than printed resources, and they can be searched more easily. Parallel corpora can also be used for teaching and research in translation, bilingual lexicography and linguistics. It can benefit translation students when searching for possible equivalents. Corpora therefore are invaluable to translation studies as they offer various ways and means of studying and understanding language structure and use in new and innovative ways. Whilst parallel corpora have been mainly used in translation studies to understand the process of others, this study interrogates the ENPC to extract bilingual terminology that can be used to create bilingual dictionaries. The approach is interdisciplinary and can provide a faster and efficient way of creating dictionaries to African lexicographers.

Examples of parallel corpora in South Africa:

- Madiba (2004) developed an English–Venda Parallel Corpus which is a pilot project of the Special language Corpora for African languages (SpeLCAL) of South Africa. The study showed how corpora can be used to develop African languages of South Africa.
- Moropa (2005) compiled an English–Xhosa parallel corpus of financial texts to show the strategies used to translate specialised terminology.
- Ndhlovu (2011) designed an English–Zulu Parallel Corpus of medical terms to explore the accessibility of specialised language to the public.

Beyond parallel corpora, other scholars have created monolingual and multilingual corpora to understand various South African languages better. There is a lot of work in various universities in the field of corpus studies and there is a need to document these and showcase the growth that is taking place in the field. It is necessary to emphasise that in using existent corpora there is a need to re-clean these for better results as these were not created with dictionary making in mind. In addition, most corpora are not comprehensive and representative of the different specialised fields, thus, there is a need to expand these so that they are current, relevant and representative.

What is ParaConc?

ParaConc is a bilingual or multilingual program that was designed for linguists and other researchers who wish to work with translated texts in order to carry out contrastive language studies or to investigate the translation process itself (Barlow 2003:1). The concordance has the following features (Barlow 2003:1–4):

- ability to sort and count words in a variety of ways
- alignment of parallel (translated) texts

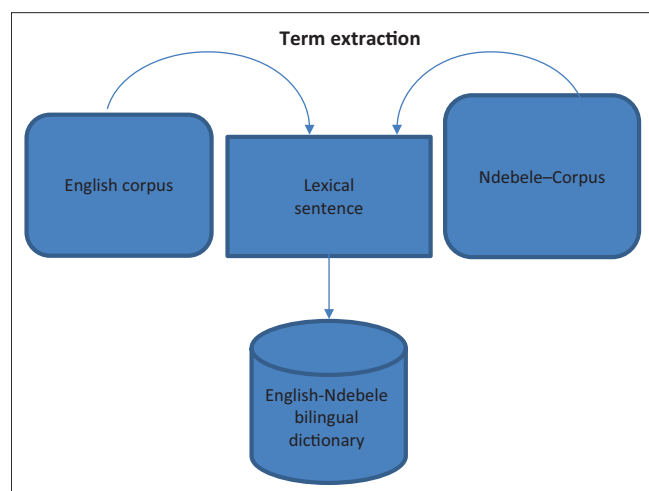


FIGURE 1: Bilingual term extraction.

- finds and displays in an easy-to-read format, in context all occurrences of a particular
- search term (and minor variations thereof)
- identification of translation equivalents
- highlighting potential translations
- a collocation viewer, which allows users to see which words belong together
- frequency lists, etc.

The above-listed features are important in bilingual dictionary making as they can be used to extract headwords and their direct translations, other possible translations that can be used as synonyms, words in context to show how words are used in different contexts and the spelling of words among other things. The next section shows how parallel corpora can be resources for compiling bilingual dictionaries.

Data analysis

ParaConc was selected as a tool to extract bilingual terminology because it has the capacity to align texts semi-automatically, with an option to align manually through the **split & merge** feature. Bowker and Barlow (2008:4) explain the alignment process as follows:

The initial part of the alignment process is carried out in three stages: first the texts are aligned based on headings, if any are present in the texts, then alignment is carried out at the paragraph level, and finally at the sentence level. The software uses the formatting information in the formatting information in files to carry out alignment of headings and paragraphs. Alignment at the sentence level is achieved by applying the Gale-Church algorithm (Gale & Church 1993)... One important thing to note is that the aligned units remain situated within the larger surrounding text. Once the texts are aligned, the translator can consult the corpus.

The ENPC was manually aligned at word, sentence and paragraph level so as to improve the accuracy of the results. Figure 2 below shows an extract of the aligned ENPC.

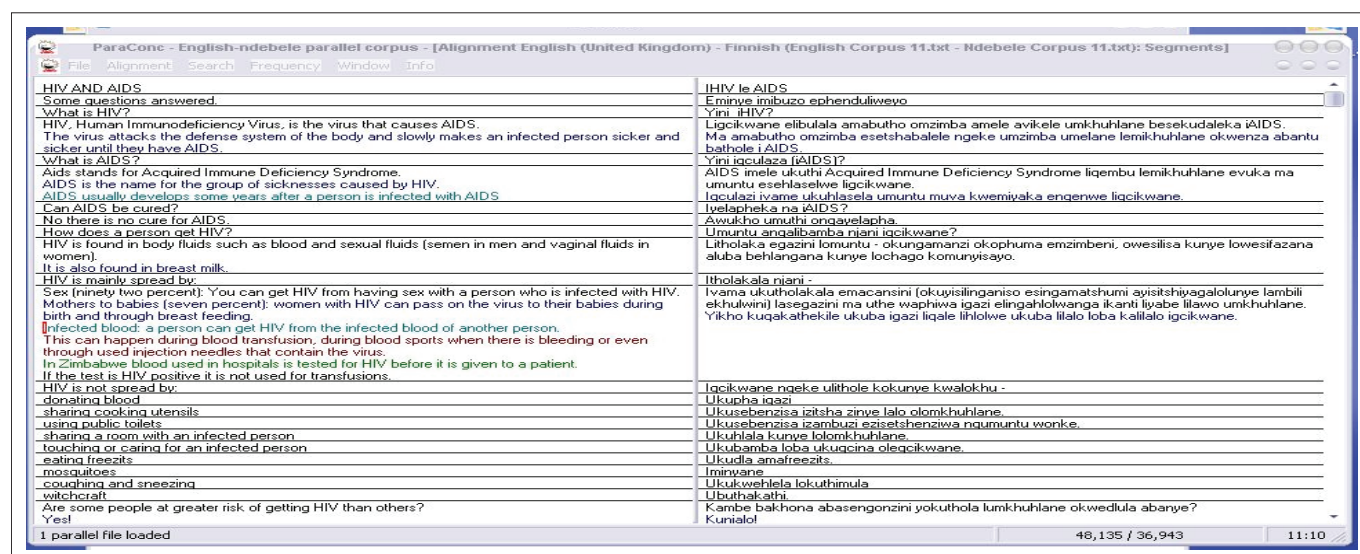


FIGURE 2: Aligned English–Ndebele corpus.

It is important to note that in some instances one sentence is aligned to two or more sentences. This results from differences in language structure and omissions in some instances. Barlow (2001, 2003:13) explains:

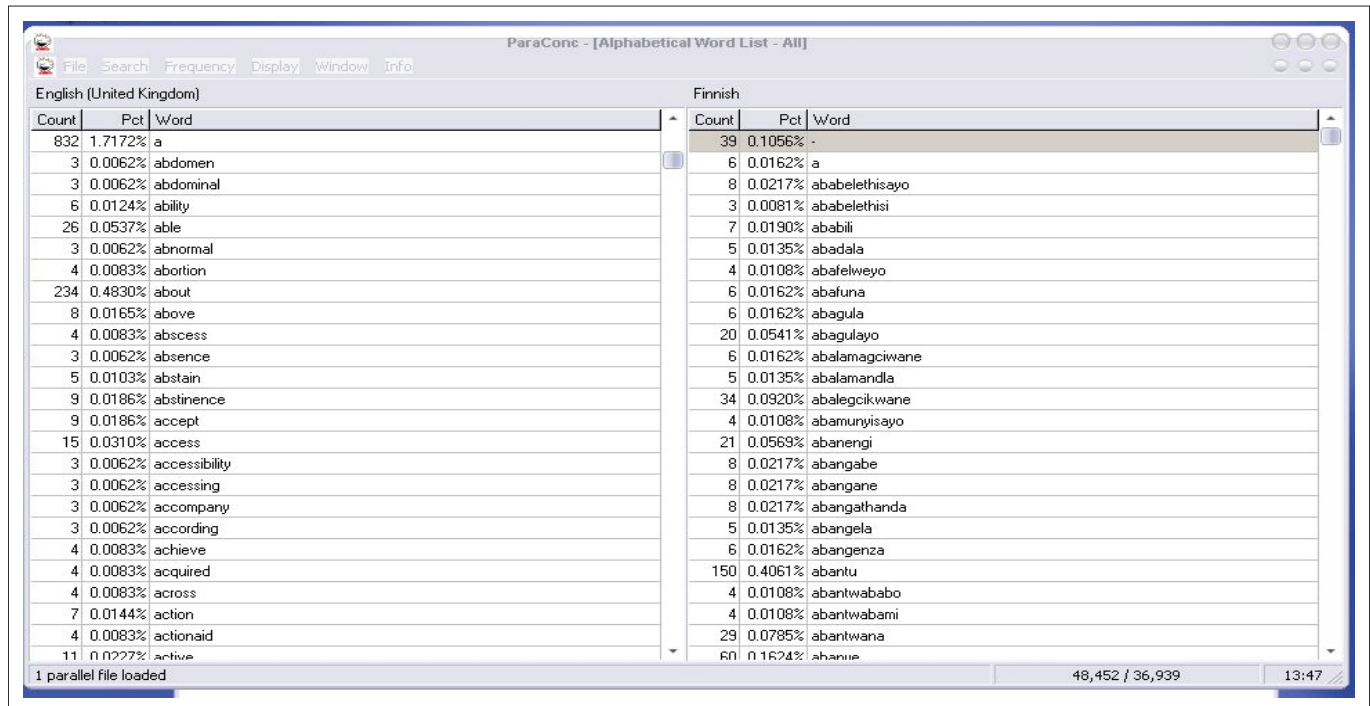
The alignment, an indication of equivalent text segments in the two languages, typically uses the sentence unit as the basic alignment segment, although naturally such an alignment is not one in which each sentence of Language A is always aligned with a sentence of Language B since occasionally a sentence in Language A may be equivalent to two sentences in Language B, or perhaps absent from Language B altogether.

ParaConc reads the information as it is aligned; therefore, it is important to always align the texts properly. Once the texts have been aligned, they can be saved in a workspace, which keeps the save information in a frozen state, thus ensuring that there is consistency in the results. In addition, the researcher can always go back to the same workspace which increases the chances of reliability of results. The information on the corpus is displayed using a format known as Key Word in Context. In a KWIC display, all the occurrences of the search pattern are lined up in the centre of the screen with a certain amount of context showing on either side (Bowker & Pearson 2002:13). For this reason, ParaConc makes it easier to identify words in contexts and word patterns making the analysis more reliable.

Identifying headwords

In order to identify headwords or search words, the alphabetic list plays an important role. Depending on the need, an alphabetic list for each corpus can be drawn, or for both as shown below.

From Figure 3, terms such as abscess and abstinence, and *abalegcikwane* (HIV positive people) and *abagulayo* (patients), can be identified as medical terms. These terms can be entered in the search bar to identify their equivalent translations. Next to the terms are percentages of how frequently the terms are used in the corpus and the number of times each



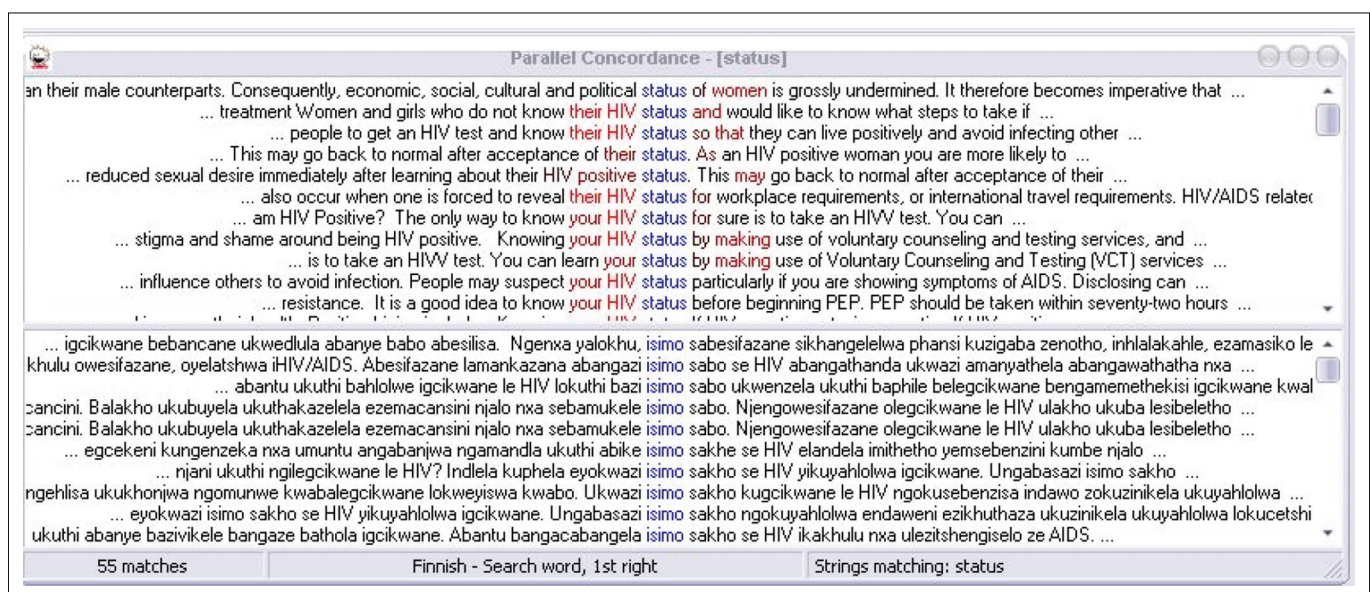
English (United Kingdom)			Finnish		
Count	Pct	Word	Count	Pct	Word
832	1.7122%	a	39	0.1056%	-
3	0.0062%	abdomen	6	0.0162%	a
3	0.0062%	abdominal	8	0.0217%	ababelethisayo
6	0.0124%	ability	3	0.0081%	ababelethisi
26	0.0537%	able	7	0.0190%	ababili
3	0.0062%	abnormal	5	0.0135%	abadala
4	0.0083%	abortion	4	0.0108%	abafelweyo
234	0.4830%	about	6	0.0162%	abafuna
8	0.0165%	above	6	0.0162%	abagula
4	0.0083%	abscess	20	0.0541%	abagulayo
3	0.0062%	absence	6	0.0162%	abalamagciwane
5	0.0103%	abstain	5	0.0135%	abalamandla
9	0.0186%	abstinence	34	0.0920%	abalegcikwane
9	0.0186%	accept	4	0.0108%	abamunyisayo
15	0.0310%	access	21	0.0569%	abanengi
3	0.0062%	accessibility	8	0.0217%	abangabe
3	0.0062%	accessing	8	0.0217%	abangane
3	0.0062%	accompany	8	0.0217%	abangathanda
3	0.0062%	according	5	0.0135%	abangela
4	0.0083%	achieve	6	0.0162%	abangenza
4	0.0083%	acquired	150	0.4061%	abantu
4	0.0083%	across	4	0.0108%	abantwababo
7	0.0144%	action	4	0.0108%	abantwabami
4	0.0083%	actionaid	29	0.0785%	abantwana
11	0.0227%	active	60	0.1624%	abanye

1 parallel file loaded

48,452 / 36,939

13:47

FIGURE 3: Alphabetic list.



an their male counterparts. Consequently, economic, social, cultural and political status of women is grossly undermined. It therefore becomes imperative that ...

... treatment Women and girls who do not know their HIV status and would like to know what steps to take if ...

... people to get an HIV test and know their HIV status so that they can live positively and avoid infecting other ...

... This may go back to normal after acceptance of their status. As an HIV positive woman you are more likely to ...

... reduced sexual desire immediately after learning about their HIV positive status. This may go back to normal after acceptance of their ...

... also occur when one is forced to reveal their HIV status for workplace requirements, or international travel requirements. HIV/AIDS related ...

... am HIV Positive? The only way to know your HIV status for sure is to take an HIV test. You can ...

... stigma and shame around being HIV positive. Knowing your HIV status by making use of voluntary counseling and testing services, and ...

... is to take an HIV test. You can learn your status by making use of Voluntary Counseling and Testing (VCT) services ...

... influence others to avoid infection. People may suspect your HIV status particularly if you are showing symptoms of AIDS. Disclosing can ...

... resistance. It is a good idea to know your HIV status before beginning PEP. PEP should be taken within seventy-two hours ...

... igcikwane bebancane ukwedlula abanye babo abesilisa. Ngenxa yalokhu, isimo sabesifazane sikhangelelwa phansi kuzigaba zenotho, inhlalakahle, ezamasiko le ...

... abantu ukuthi bahlolwe igcikwane le HIV lokuthi bazi isimo sabo ukwenzela ukuthi baphile belegcikwane bengamemethekisi igcikwane kwal ...

cancini. Balakho ukubuyela ukuthakazelela ezemacansini njalo nxa sebamukele isimo sabo. Njengowesifazane olegcikwane le HIV ulakho ukuba lesibeletso ...

cancini. Balakho ukubuyela ukuthakazelela ezemacansini njalo nxa sebamukele isimo sabo. Njengowesifazane olegcikwane le HIV ulakho ukuba lesibeletso ...

... egcekeni kungenzeka nxa umuntu angabaniwa ngamandla ukuthi abike isimo sakhe se HIV elandela imithetho yemsebenzini kumbe njalo ...

... njani ukuthi ngilegcikwane le HIV? Indlela kuphela eyokwazi isimo sakho se HIV yikuyahlolwa igcikwane. Ungabasazi isimo sakho ...

ngehisa ukukhunjwa ngomunwe kwabalegcikwane lokweyiswa kwabo. Ukwazi isimo sakho kugcikwane le HIV ngokusebenzisa indawo zokuzinikela ukuyahlolwa ...

... eyokwazi isimo sakho se HIV yikuyahlolwa igcikwane. Ungabasazi isimo sakho ngokuyahlolwa endaweni ezikhuthaza ukuzinikela ukuyahlolwa lokucetshi ...

ukuthi abanye bazivikele bangaze bathola igcikwane. Abantu bangacabangela isimo sakho se HIV ikakhulu nxa ulezitshengiselo ze AIDS. ...

55 matches

Finnish - Search word, 1st right

Strings matching: status

FIGURE 4: Translation of status.

term appears, which is important in dictionary making. The more frequently a term is used, the higher the chances of it being accepted by the target users of the dictionary as it will have been popularised by the translation. It is important to note that searches can be done in the source language and the target languages depending on the need at a particular point in time. This is important during the process of dictionary making as terms can be selected based on how frequently they are used by different translators.

Performing a search

The search facility is one of the most important features in extracting bilingual terminology. To perform a basic

concordance search, the researcher can click on the search menu; the *Search* option will be selected by clicking on it. After clicking on *Search*, a search box appears where one can enter the term or phrase searched. Bowker and Barlow (2008:4) explain that 'by choosing the basic search command, the translator can retrieve all examples of a word or phrase (or part of a word) from the corpus'. For example, the search word *status* was entered and it was translated as *isimo* by the Ndebele translators (Figure 4 above).

There is a general consensus in the translation of the term *status* in the ENPC as *isimo*. The term *status* appears 55 times in the English corpus, and it collocates with the term

HIV. This shows that ParaConc is able to extract terms and their direct translations thus minimising the process of translating terms during dictionary making. Also, in ParaConc all the matches are displayed at once and the user can peruse them at a glance instead of having to click through them (Bowker & Barlow 2008:11). Table 1 reflects a list of terms that were extracted from the ENPC with their translations and synonyms where applicable. These terms can be included in a dictionary based on the purpose of the dictionary.

As shown in Table 1, in some instances, a word can have more than one translation or can have many synonyms. In a situation where there is more than one translation for a

term, possible translations can be extracted by positioning the cursor in the lower pane and clicking on the right side of the mouse (Bowker & Barlow 2008:5). This will produce a hot words list. According to Barlow (2003:34), hot words are possible translations and associated words (collocates) that are suggested by the program itself. In relation to the search word, the hot word list shows the most commonly used translations and collocates in descending order based on frequency (Figure 5 below).

From Figure 5, the term *AIDS* is ranked highly because it is usually associated (or collocates) with *HIV*. Second on the list is the term *ihiv* which is the most common translation for the English term *HIV*, which is also translated as *igcikwane le*

TABLE 1: Terminology list (words extracted from the English-Ndebele parallel corpus).

Specialised terms	Ndebele translations	Specialised terms	Ndebele translations
Antenatal care	ukuhlolwa uzithwele	Parent-to-child transmission (PTCT)	Ukuthelwa kosane igcikwane ngabazali
AIDS	ingculaza/iAIDS	Pills	Amaphilisi
Catheter	ithumbu lomchamo	Osteoporosis	amathambo angaqinanga
Cervix	umlomo wesibeetho	Opportunistic infections	imikhuhlane ehlasela abaphila le HIV
Condom	ikhondomu/umncwado	Diabetes	umkhuhlane wetshekela
Discrimination	Ubandlululo	Malaria	Uqhuqho
Defence system	amabutho omzimba	TB	Ufuba
Doctor	Udokotela	Ribbon	Umcikiliso
Drugs	Izidakamizwa	Re-infection	ukuthola/ukuthelwa igcikwane njalo
Exclusive breast feeding	ukumunyisa ibele likamama kuphela	Sexually transmitted infections	imikhuhlane yengulamakhwa
Gloves	Amagilavu	Symptoms	izitshengiselo/izibonakaliso
Hospital	Esibhedlela	Status	Isimo
Immune system	amandla okulwisa imikhuhlane	Testing	Ukuhlolwa
Medicines	Imithi	Virus	Igcikwane
Nurse	Umongikazi	Voluntary counselling and testing	uhlelo lokuzikhetela ukuhlolwa lokwelulekwa

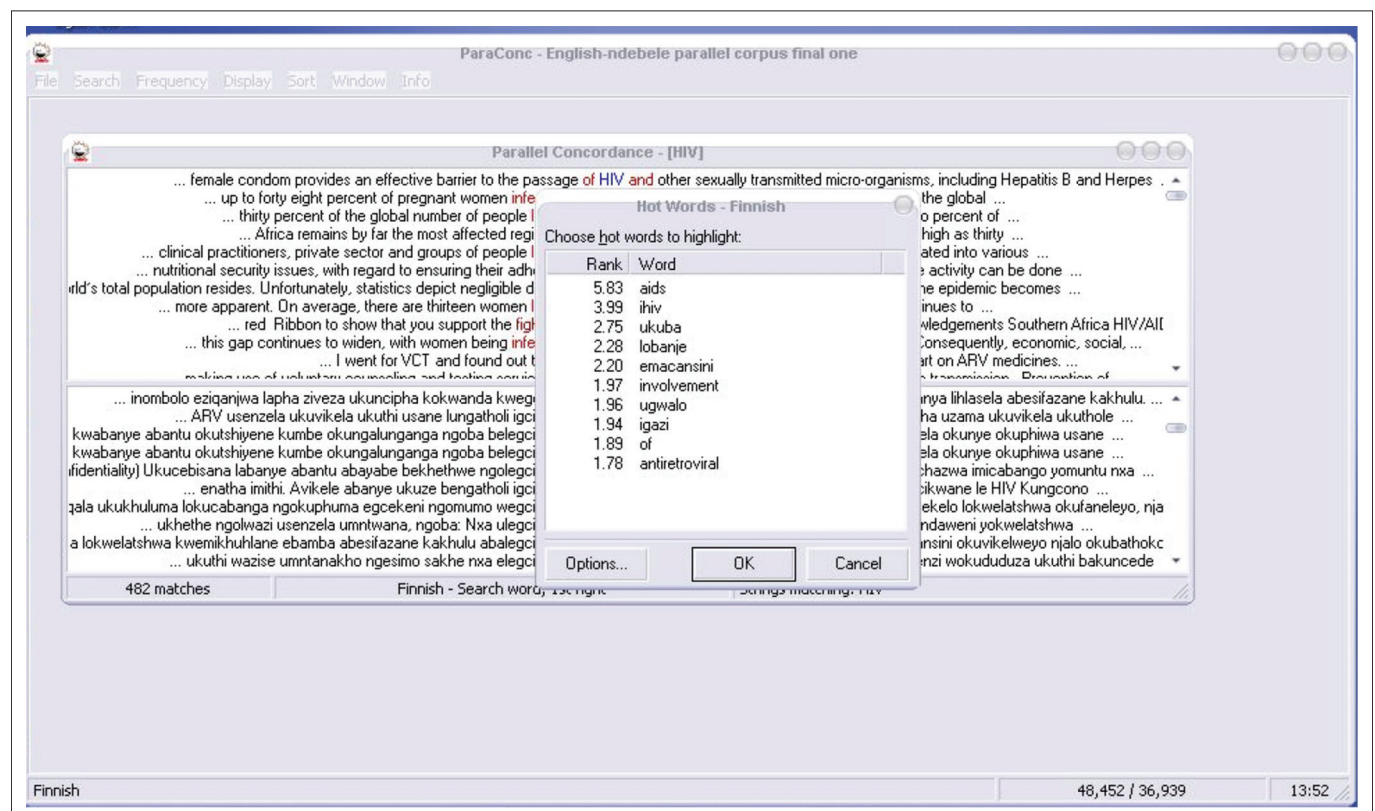


FIGURE 5: Hot words list.

HIV in the ENPC. The term *HIV* is also usually associated with testing of blood, hence the collocate *igazi*. The words are arranged in relation to how 'hot' they are, that is, how frequently they appear in relation to the term. Any word that is searched appears within context, which is important in dictionary making as one word can have more than one meaning. Other examples of words that were extracted from the ENPC with multiple translations are presented below:

- syndrome: *izitshengiselo; izibonakalis*
- counselling: *ukududuzwa; ukududuzwa lokuthola usekelo; ukucetshiswa; ukwelulekwa*
- counsellor: *oduduza losekela abantu, umduduzi; umeluleki*
- stigma: *ukukhonjwa ngomunwe, ukuyangiswa; ihlazo*
- prevention: *ukuvikela, ukuvimba; ukwenqabela*.

Some of the terms can be used as synonyms during dictionary making and some can be discarded if they do not fit the profile. Nevertheless, it is important to note that the multiplicity of terms used to translate medical terms points to the lack of a specific term to translate the term in Ndebele; therefore, translators come up with different terms to explain one concept.

In a situation where there are many synonyms, another method that can be used to determine the suitability of terms is to use the frequency feature to determine the term that is frequently used by translators. To carry out a frequency

search, the researcher selected ALL to create an English–Ndebele frequency list as shown in Figure 6 below.

The list shows that the term 'medicines' appears 327 times in the English corpus and its translation *imithi* appears 386 times in the Ndebele corpus. This shows that many translators translate medicines as *imithi* so that the term can be accepted. However, when the procedure is reversed and *imithi* is entered as a search word, another translation appears – *imithi* as drugs which accounts for more hits of *imithi* (386) in the Ndebele corpus compared with the 327 of the English corpus. This is a common problem in African languages, where one word is used to explain two or more words in the source language. Ndhlovu (2014:332) gave an example of *amaphilisi* which was used in the ENPC to translate English words such as 'tablets', 'dose', 'painkillers' and medicines. These are some of the challenges that lexicographers have to deal with when entering words and their meanings.

The discussion so far has shown that ParaConc can get results that are more accurate, more reliable and faster, and help lexicographers to do away with the labour-intensive process of translations – except in special circumstances, making this a better option of creating bilingual dictionaries than the manual method. Furthermore, through ParaConc, more complex search commands can also be used if desired. Some of the possible advanced search options are: Text search, Regular expression search, Tag (part-of-speech) search, Batch

English (United Kingdom)			Finnish		
Count	Pct	Word	Count	Pct	Word
2045	4.2207%	the	1086	2.9400%	ukuthi
1650	3.4054%	to	522	1.4131%	nxa
1505	3.1062%	and	509	1.3779%	njalo
1047	2.1609%	of	482	1.3049%	kumbe
984	2.0309%	you	386	1.0450%	imithi
832	1.7172%	a	385	1.0423%	le
714	1.4736%	is	335	0.9069%	hiv
706	1.4571%	or	256	0.6930%	loba
583	1.2033%	your	234	0.6335%	lokhu
545	1.1248%	in	195	0.5279%	yama
499	1.0299%	that	192	0.5198%	arv
482	0.9948%	hiv	180	0.4873%	umuntu
478	0.9865%	with	150	0.4061%	abantu
470	0.9700%	for	143	0.3871%	indlela
452	0.9329%	are	135	0.3655%	umzimba
446	0.9205%	be	134	0.3628%	wakho
414	0.8545%	not	130	0.3519%	kuhle
384	0.7925%	i	127	0.3438%	emacansini
374	0.7719%	can	126	0.3411%	kakhulu
365	0.7533%	it	123	0.3330%	aids
339	0.6997%	on	121	0.3276%	ukudla
327	0.6749%	medicines	118	0.3194%	igcikwane
321	0.6625%	have	110	0.2978%	abesifazane
314	0.6481%	if	105	0.2843%	njani
303	0.6254%	as	102	0.2761%	kufanele

1 parallel file loaded 48,452 / 36,939 13:43

FIGURE 6: Frequency list.

search and various heading-sensitive and context-sensitive searches (Bowker & Barlow 2008:5).

Although ParaConc can achieve so much accuracy, it is important to state that the human element cannot be totally eliminated during dictionary making. Some decisions may require further dictionary research to determine the suitability of terms for inclusion in the dictionary. That is, monolingual and bilingual dictionaries can be used to determine the meaning and suitability of terms for selection as headwords. For example, when it comes to a term like *counsellor* which had multiple translations: *ukududuzwa*; *ukududuzwa lokuthola usekelo*; *ukucetshiwa* and *ukwelulekwa*, dictionaries can help define the general meanings of the terms to guide lexicographers on term suitability. The term *counsellor* is defined in the Longman Dictionary of Contemporary English (1995) as 'someone whose job is to help and support people with problems' and in the *South African Concise Oxford Dictionary* (2007), 'as a person trained to give guidance on personal, social, or psychological problems'. The two definitions emphasise the aspect of counselling as a profession, with the second definition underlining the aspect of training. That is, for one to be a 'counsellor' in such a specialised field as HIV/AIDS, one has to undergo training before offering these services. The new phenomenon of specialised counselling led to Ndebele translators translating the term differently with some translating it as *umduduzi*, *umcebisi* and *umeluleki*.

The term *umduduzi* is defined in *Isichazamazwi SesiNdebele* (2001) as *nxa ungumduduzi uyabe uletha intokozo emuntwini owehlelewe yilishwa kumbe odubekileyo* (if you are a comforter, you will be bringing joy to people who will be facing hardships or problems). In other words, *umduduzi* is someone who comforts people during a time of sorrow or tribulations. Comparing this definition, to the corresponding English definitions, it seems this term barely captures the act of counselling, which involves more than the act of comforting the clients. The term *umeluleki* on the other hand is defined in the *Isichazamazwi SesiNdebele* (2001) as *ngumuntu onika abanye amacebo ebatshengisa indlela eqondileyo okufanele bayithathe ekwenzeni ulutho oluthile* (someone who gives other people strategies or ideas, showing them the right way to follow in doing something). The term *umeluleki*, when defined in this way, is a general term that is used to translate a specialised term. However, this term can be extended to mean more than just giving advice or strategies, to include the act of training in order to assist people who are facing problems. Unfortunately, the term *umcebisi*, taken from *ukucebisa/ukucetshiswa*, does not appear in the Ndebele corpus. Looking at these terms, *umeluleki* seems to be a better translation and the frequency list justifies its selection. *Umcebisi* can be used as a synonym and *umduduzi* discarded as it does not capture the essence of counselling. Using dictionaries as supportive material proves that the human element cannot be totally eradicated with the use of technology during dictionary making. That is, technology can quicken the search for terms, their translations, frequencies, synonyms and examples of

usage, but lexicographers have to decide which terms are suitable based on technological result, manual research and field expert advice.

In lexicography, another important element is the contextual usage of terms so that target users can distinguish between the different uses of the terms provided. According to Bowker and Barlow (2008), one important thing to note about ParaConc is that the aligned units remain situated within the larger surrounding text. That is, in the parallel corpus, words appear within context; thus, when a search word is entered, the results show how the word can be used in a sentence and the different uses of a particular word. The following are examples of words in context.

Discrimination: Ubandlululo

Source Text: The purpose of this activity is to help girls and women discuss ways through which they can identify and deal with stigma and *discrimination* ...

Target Text: Injongo yokwenza lokhu yikuncedisa amankazana labesifazane ukuthi bahluze indlela ezingenza baqambe ukukhonjwa ngomunwe *lobandlululo* ...

Back Translation: The aim of doing this is to help girls and women analyse ways of identifying stigma and discrimination ...

Treatment: Ukwelatshwa

Source Text: Gender-based violence associated with *treatment*.

Target Text: Udlakela lwezindlini oluqophe abesifazane kuhambelana njalo *lokwelatshwa*.

Back Translation: Domestic violence that is targeted at women that is related to *treatment*.

In dictionary making, this feature can help provide contextual usage of different terms, thus helping dictionary users to understand how the same word can be used in different contexts. In order to determine if the selected search words are coming from one source or multiple sources so as to eradicate bias during term selection, a distribution profile can be drawn.

File distribution

Distribution of hits helps one to get a sense of where the words are located in the corpus file. To get to the *Distribution* option, the concordance results window must be opened first, and then one can click on *Display* on the main menu. A pop-up box appears with a number of options, and the distribution option can be selected by clicking on it. Figure 7 shows how the term 'HIV' is distributed in the Ndebele corpus. The Finnish language is used to represent Ndebele

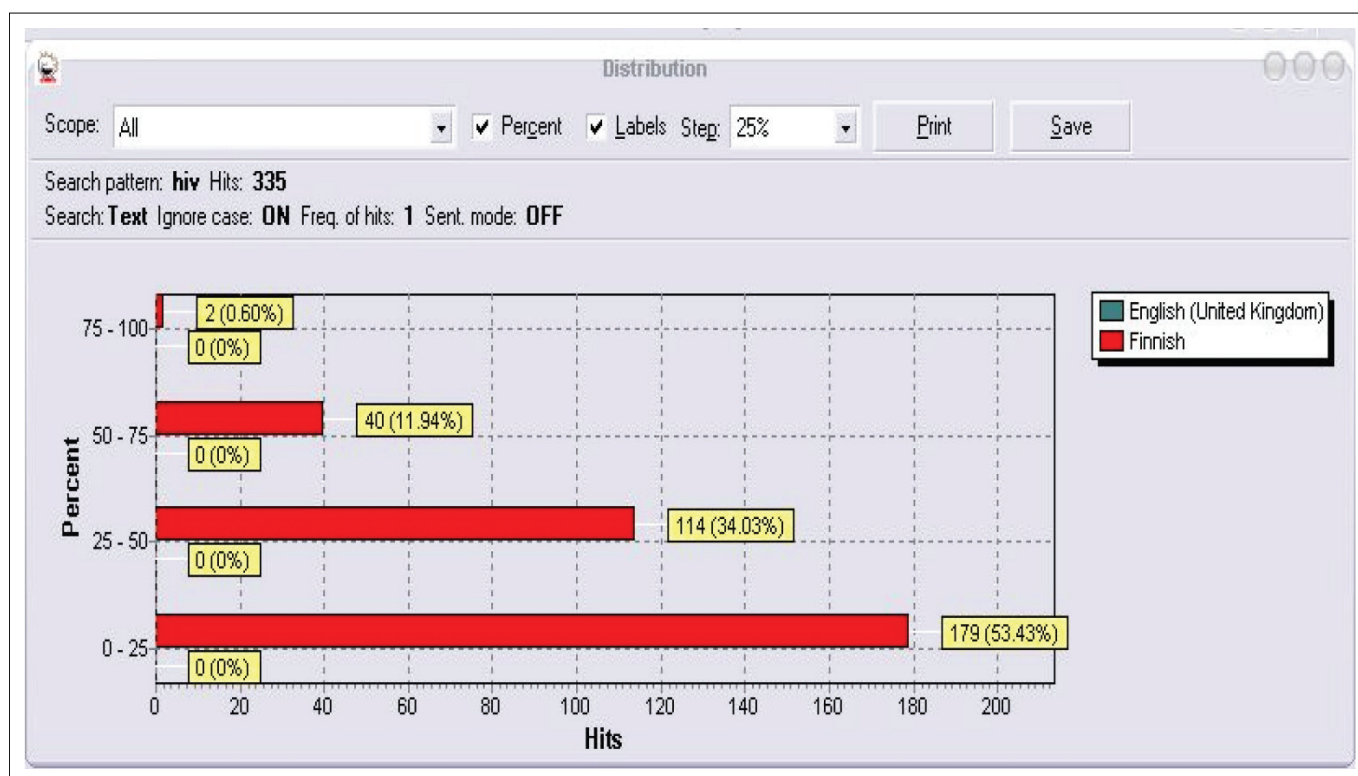


FIGURE 7: Distribution of the term 'HIV' in the Ndebele file.

because ParaConc does not include all African languages in its list of languages.

There are 335 hits for the term 'HIV' in Ndebele. There is a higher concentration of the term 'HIV' in the initial pages of the Ndebele file. To track the hits and explain the differences in the number of hits between the two corpus files, one can click on a sentence with the hit *HIV* in the English window whereupon its translation will appear in the bottom window. The first statement was the heading 'HIV and AIDS' that was translated as 'IHIV le AIDS' in Ndebele. The second hit sentence 'How does a person get HIV?' is translated as *Umntu angalibamba njani igcikwane?* The absence of the term *HIV* in this instance helps to explain the differences between the number of HIV hits in English and Ndebele. Some translators refer to HIV simply as *igcikwane*, though the context always reveals that it is the HIV virus they are referring to. It is possible to get more information about a hit, such as the file name, line and page number among others. This can be done by right clicking on the highlighted hit, whereupon a pop-up box appears with a number of options. Then, select *Display Info* and the information will be furnished.

This article has shown that it is possible to draw terms and their translations, frequencies, collocates, synonyms and words in context. However, in using a parallel corpus, it is important to emphasise that the results drawn are dependent on the data that were entered and how clean the corpus is. For better results, it is vital to clean the corpus ensuring that words are spelt and the texts aligned properly. Bowker and Barlow (2008:5) explain that 'BCs are sometimes criticized because of the nature of the matching process that they use.

By default, these tools basically search through the corpus for occurrences that match the entered search pattern *precisely*'. An example of this phenomenon is that if there are two different spellings or more, the results will show these as different translations. For example, the term virus was translated as *igcikwane* and *igciwane* showing differences in spelling, with one being a different orthographic representation or spelling of the other. Through the frequency list, *igcikwane* was shown as the most commonly used spelling. This means that lexicographers have to be conscious of spelling during the process of cleaning the corpus.

In dictionary making, a contentious issue is the entry of loan words. In this article, it was noted that most Ndebele translators resort to loaning when confronted by new and difficult concepts. This loaning appears in the form of pure loaning, pure loan words preceded by explanations, and indigenised loaning (Ndhlovu 2014). Ndhlovu (2014:333) identified the following words as pure loan words in the ENPC: *ama-gland* (glands), *ama-hormones* (hormones) and *ama-antibodies* (antibodies) among others. She argued that such words limit the accessibility of translated texts and the very same words are bound to cause problems during dictionary making, pointing to a need by lexicographers to create terms that are more accessible to the readers. In fact retaining pure loan words in a dictionary is tantamount to cheating the users of the dictionaries as comprehension maybe diminished, especially in cases where readers are not familiar with the word structure and pronunciation; hence, more strategies should be explored to overcome the challenges of term scarcity.

Another thorny issue in the ENPC is the voluminous existence of indigenised words. Though not the most commonly used strategy compared with pure loaning, the ENPC is replete with indigenised words that are used to translate specialised terms. Words such as *ikhondomu* (condom), *amaphilisi* (pills) and *udokotela* (doctor) were identified by Ndhlovu (2014:332) as indigenised words in the ENPC. These words that have adopted the Ndebele spelling and structure and are commonly used by the Ndebele community show that lexicographers have to have clear principles on which loan words can and which ones cannot be included. It is the researcher's opinion that indigenised words should be included in monolingual and bilingual dictionaries after other methods of term creation have been exhausted and if their inclusion is beneficial to the target community. However, this should be done within set boundaries because large quantities of loan words can defraud the dictionary of its identity. Over-reliance on loaning also inhibits the growth of African languages – hence the need for more robust methods of creating terms in African communities.

One way of creating terms is working hand in hand with field experts and target communities during the process of dictionary making – this is more so in the health sector – because whatever disease affects the people, they are bound to have a term for it. Terms like *ingulamakhwa* (STI), *utshukela* (diabetes), *ufuba* (TB) and *ingculaza* (HIV) among others were coined and accepted by communities and more can still be developed if collaboration between translators, lexicographers, subject specialists and the public among others is maximised. This approach shows that dictionary making is a team effort that involves lexicographers, translators, field experts and the target communities or users.

Another point to note when creating specialised bilingual dictionaries is that it is important to create representative parallel corpora so that as many terms as possible are covered meaning different translations from different translators are represented, as such as many strategies of translation are presented as possible – for better results. The current corpus was not representative as it focused mostly on HIV and/or AIDS and related diseases, and the texts were from 11 translators. In addition, the results produced are dependent on the information entered and the quality of the corpus; it is necessary to check the quality of the corpus used. Lastly, the needs of the users have to be taken into account so that the dictionaries are user-friendly, accessible and acceptable to the target communities.

Conclusion

This article showed that it is possible to successfully extract bilingual terms for an English and Ndebele bilingual dictionary from an ENPC using ParaConc features such the search facility, alphabetic list, frequency list and hot words. English headwords were identified through the alphabetic list and their equivalent Ndebele translations and synonyms were

identified using the search feature. To determine the suitability of a headword, various strategies were suggested such as using the frequency feature, hot words list and also relying on dictionaries, proving that technology is not the beginning and end of all things. Whilst technology can offer a faster and reliable way of identifying terms and their translations, it was noted that the human factor plays an important role in dictionary making. The article also showed that it is possible to extract terms within context which is important in dictionary making, so that target users can understand the different uses of words. The researcher emphasised that dictionary making is a collaborative effort, wherein lexicographers should work with translators, field experts, communities and potential users during the process of creating dictionaries. This approach is emphasised in order to create dictionaries that meet the needs of the target users. The researcher also emphasised the importance of creating a high-quality representative parallel corpus because the results produced are dependent on the data that went in. Like all dictionary projects, a few challenges were identified and these included: (1) ensuring that words are spelt correctly before being uploaded onto ParaConc because any mistake affects the outcome in terms of word count and frequencies; (2) inclusion of loan words – the ENPC was cluttered with pure loan words which required re-translation, and with regard to indigenised loan words, they were recommended for inclusion but only after other strategies were exhausted and based on the level of acceptability among the community members; (3) availability of synonyms – many specialised English terms had a plethora of synonyms pointing to a lack of standardisation in the Ndebele language as different translators translate terms according to their understanding. In spite of these challenges, extracting bilingual terminology from parallel corpora was faster and the results were more accurate. Lastly, dictionaries should be created with the needs of the end user in mind, and they are a necessary resource in our globalised world as standardising tools and repositories of subject knowledge.

Acknowledgements

Competing interests

The author declares that she has no financial or personal relationship which may have inappropriately influenced her in writing this article.

References

- Baker, M., 1995, 'Corpora in translation studies: An overview and some suggestions for future research', *Target* 7(2), 223–243. <http://dx.doi.org/10.1075/target.7.2.03bak>
- Barlow, M., 2001, *ParaConc: A concordance for parallel texts* (draft 3/03), Rice University.
- Barlow, M., 2003, *ParaConc: Concordance software for multilingual parallel corpora*, Rice University, Houston, TX. <http://www.mt-archive.info/LREC-2002-Barlow.pdf>
- Bowker, L. & Barlow, M., 2008, 'Bilingual concordancers and translation memories: A comparative evaluation', in *Topics in language resources for translation and localisation*, vol. 79, pp. 1–22, John Benjamins Publishing Company, Amsterdam, viewed n.d., from <https://www.ac/web-org/anthology/W/WO4/WO4-1408.pdf>
- Bowker, L. & Pearson, J., 2002, *Working with specialized corpora*, Routledge, London.
- Chabata, E., 2013, 'The language factor in the development of Africa: A case for the compilation of specialised dictionaries in indigenous African languages', *South African Journal of African Languages* 33(1), 51–58. <http://dx.doi.org/10.2989/02572117.2013.793940>

- Chen, S., 1993, Aligning sentences in bilingual corpora using lexical information, In *Proceedings of the 31st ACL*, 91-16, Columbus, Ohio, USA.
- Chimhundu, H., 2006, 'Language, dialect and region: The handling of language variation in Shona dictionaries', in B. Brock-Utne & I. Skattum (eds.), *Languages and education in Africa: A comparative and transdisciplinary analysis*, pp. 237–252, Symposiums Books, Oxford.
- Erdmann, M., Nakayama, K., Takahiro, H. & Nishio, S., 2009, 'Using an SVM classifier to improve the extraction of bilingual terminology from Wikipedia', *ACM Transactions on Multimedia Computing, Communications and Applications* 5(4), Article 31 <http://dx.doi.org/10.1145/1596990.1596995>
- Fung, P., 1998, 'A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora', in D. Farwell, et al. (eds.), *AMTA'98, LNAI 1529*, viewed n.d., from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.54.5787&rep=rep1&type=pdf>
- Gale, W. A. & Church, K. W., 1993, 'A program for aligning sentences in bilingual corpora', *Computational Linguistics*, 19, 75–102.
- Gauton, R. & De Schryver, G.-M., 2004, 'Translating technical texts into Zulu with the aid of multilingual and/or parallel corpora', *Language Matters* 35(1), 133–147. <http://dx.doi.org/10.1080/10228190408566209>
- Gauton, R., Taljard, E. & De Schryver, G.-M., 2003, 'Towards strategies for translating terminology into all South African languages: A corpus-based approach', in G.M. De Schryver (ed.), *TAMA 2003, South Africa: Terminology in advanced management applications: 6th international TAMA conference: Conference proceedings, Language Matters*, UNISA Press, Pretoria(SF) 2, pp. 81–88.
- Guinovart, X.G. & Simoes, A., 2009, 'Parallel corpus-based terminological extraction', viewed n.d., from <http://ceur-ws.org/Vol-578/papers23.pdf>
- Hadebe, S., Dube, K., Matshakayile-Ndlovu, T., Mhlabi, S., Khumalo, L., Maphosa, M., et al., 2001, *IsiChazamazwi SesiNdebele*, College Press & ALRI, Harare.
- Kay, M & Martin, R., 1993, Text-translation alignment. *Computational Linguistics*, 19 (1), 121–142.
- Kenny, D., 1998, 'Creatures of habit? What translators usually do with words', *Meta* 43(4), 515–523. <http://dx.doi.org/10.7202/003302ar>
- Kruger, A., 2010, 'Translating public information text on health issues into languages of limited diffusion in South Africa', in R.A. Valdeón (ed.), *Translating information*, pp. 151–168, Universidad de Oviedo Press, Ediuono.
- Li, B. & Gaussier, E., 2010, 'Improving corpus comparability for bilingual lexicon extraction from comparable corpora', *Proceedings of the 23rd International conference on computational linguistics*, Beijing, pp 644–652, viewed n.d., from <http://anthology.aclweb.org/C/C10-1073.pdf>
- Madiba, M., 2004, 'Parallel corpora as tools for developing the indigenous languages of South Africa, with special reference to Venda', *Language Matters* 35(1), 133–147. <http://dx.doi.org/10.1080/10228190408566208>
- McEnery, A.M. & Xiao, R.Z., 2007, 'Parallel and comparable corpora: What are they up to?', in G. James & G. Anderman (eds.), *Incorporating corpora: Translation and the linguist: Translating Europe*, pp. 1–13, Multilingual Matters, Clevedon, UK.
- Melamed, I. D., 1997a, A portable algorithm for mapping bitext correspondence, In *Proceedings of the 35th ACL and the 8th EACL*, 305–312, Madrid, Spain.
- Melamed, I. D., 1997b, A word-to-word model of translational equivalence, In *Proceedings of the 35th ACL and the 8th EACL*, 490497, Madrid, Spain.
- Mohochi, S., 2006, 'Turning to indigenous languages for increased citizen participation in the African development process', Eagleton University, pp. 1–14, viewed 15 November 2015, from http://www.codesria.org/links/conferences/general_assembly11/ga
- Morin, E. & Prochasson, E., 2011, 'Bilingual extraction from comparable corpora enhanced with parallel corpora', *Proceedings of the 4th workshop and using comparable corpora*, Association for Computational linguistics, OR, pp. 27–34, viewed from <http://www.mtarchive.info/BUCC-2011-Morin.pdf>
- Moropa, K., 2005, 'An investigation of translation of universals in a parallel corpus of English-Xhosa texts', Unpublished DPhil thesis, University of South Africa, Pretoria.
- Ndhlovu, K., 2012, 'An investigation of strategies used by Ndebele translators in Zimbabwe in translating HIV/AIDS texts: A corpus-based approach', Unpublished PhD thesis, Alice University of Fort Hare.
- Ndhlovu, K., 2014, 'Term-creation strategies used by Ndebele translators in Zimbabwe in the health sector', *Stellenbosch Papers in Linguistics Plus* 43, 327–344. <http://dx.doi.org/10.5842/43-0-192>
- Ndlovu, V., 20011, 'The accessibility of translated Zulu health texts: An investigation of strategies', Unpublished PhD thesis, University of South Africa, Pretoria.
- Nkomo D., 2010, 'Affirming a role for specialised dictionaries in indigenous African languages', *Lexikos* 20, 371–389.
- Pearson, P.T.R., 1995, *Longman dictionary of contemporary English*, Longman, London.
- Pelling, J.N., 1971, *A practical Ndebele dictionary*, Longman Rhodesia, Bulawayo.
- Prah, K.K., 2008, 'African languages in a globalising world', *OPENSOURCE: A Digest of the Open Society Initiative for Southern Africa* 2, 19–23.
- Kavanagh, K. (ed.), 2007, *South African Concise Oxford Dictionary*, Oxford University Press, Southern Africa, Cape Town.
- Trew, R., 1994, 'The development of training models for African language translators and interpreters', in A. Kruger (ed.), *New perspectives on teaching translators and interpreters in South Africa*, pp. 73–102, University of South Africa, Pretoria.
- Van Huyssteen, L., 1999, 'Problems regarding term creation in Southern African languages with special reference to Zulu', *South African Journal of African Languages* 19(3), 179–187.
- Wallmach, K. & Kruger, A., 1999, "'Putting a sock on it": A contrastive analysis of problem solving translation strategies between African and European languages', *South African Journal of African Languages* 19(4), 275–289.
- Yu, K. & Tsujii, J., 2009a, 'Extracting bilingual dictionary from comparable corpora with dependency heterogeneity', *Proceedings of NAACL-HLT 2009, Companion Volume: Short Papers*, Boulder, Colorado, pp. 121–124.
- Yu, K. & Tsujii, J., 2009b, 'Bilingual dictionary extraction from Wikipedia, 2009', *Proceedings from Machine Translation Summit XIII*, Ottawa, Canada <http://www.mt-archive.info/MTS-2009-Yu.pdf>