Challenges to Issues of Balance and Representativeness in African Lexicography*

Thapelo Joseph Otlogetswe, *Information Technology Research Institute, University of Brighton, Brighton, United Kingdom and Department of English, University of Botswana, Gaborone, Botswana (otlogetswe@mopipi.ub.bw)*

Abstract: Modern dictionaries depend on corpora of different sizes and types for frequency listings, concordances and collocations, illustrative sentences and grammatical information. With the help of computer software, retrieving such information has increasingly become relatively easy. However, the quality of retrieved information for lexicographic purposes depends on the information input at the stage of corpus construction. If corpora are not representative of the different language usages of a speech community, they may prove to be unreliable sources of lexicographic information. There are, however, issues in African languages which make many African corpora questionable. These issues include a lack of texts of different genres, the unavailability of balanced and representative written texts, a complete absence of spoken texts as well as literacy problems in African societies. This article therefore explores the different challenges to the construction of reliable corpora in African languages. It argues that African languages face peculiar challenges and corpus research may require a different treatment compared to European and American corpus research. It finally concludes that issues of balance and representativeness appear theoretically impossible when looking at the results of sociolinguistic research on the different existing language varieties which are difficult to represent accurately in a corpus.

Keywords: AFRICAN LANGUAGES, BALANCE, BANK OF ENGLISH, BORROWING, BRITISH NATIONAL CORPUS, COBUILD, CODE-SWITCHING, COMPUTERS, CORPORA, DIALECT, DICTIONARIES, FREQUENCY, LANGUAGE VARIETY, REPRESENTATIVENESS, SETSWANA, SOCIOLINGUISTICS, SPEECH, TEXT

Opsomming: Uitdagings betreffende kwessies van balans en verteenwoordigendheid in Afrikaleksikografie. Moderne woordeboeke steun op korpusse van verskillende groottes en soorte vir frekwensielyste, konkordansies en kollokasies, voorbeeldsinne en taalkundige inligting. Met die hulp van rekenaarprogrammatuur het die herwinning van sulke inligting toenemend redelik maklik geword. Die gehalte van herwonne inligting vir leksikografiese doeleindes steun egter op die inligtingsinset by die korpusboufase. Indien korpusse nie verteenwoordigend is van die verskillende taalgebruike van 'n spraakgemeenskap nie, mag hulle blyk

* This article is a revised version of a paper presented at the Eighth International Conference of the African Association for Lexicography, organised by the Department of German and Romance Languages, University of Namibia, Windhoek, Namibia, 7–9 July 2003.

onbetroubare bronne van leksikografiese inligting te wees. Daar is egter kwessies in Afrikatale wat baie Afrikakorpusse problematies maak. Hierdie kwessies sluit in die tekort aan tekste van verskillende genres, die niebeskikbaarheid van gebalanseerde en verteenwoordigende geskrewe tekste, die volkome afwesigheid van gesproke tekste asook geletterdheidsprobleme in Afrikagemeenskappe. Hierdie artikel ondersoek derhalwe die verskillende uitdagings betreffende die bou van betroubare Afrikataalkorpusse. Dit voer aan dat Afrikatale teenoor besondere uitdagings staan en korpusnavorsing 'n verskillende behandeling mag vereis in vergelyking met Europese en Amerikaanse korpusnavorsing. Ten slotte kom dit tot die gevolgtrekking dat kwessies van balans en verteenwoordigendheid teoreties onmoontlik lyk wanneer gekyk word na die resultate van sosiolinguistiese navorsing oor die verskillende bestaande taalvariëteite wat moeilik is om presies in 'n korpus te verteenwoordig.

Sleutelwoorde: AFRIKATALE, BALANS, BANK OF ENGLISH, BRITISH NATIONAL CORPUS, COBUILD, DIALEK, FREKWENSIE, KODEWISSELING, KORPUSSE, ONTLENING, REKENAARS, SETSWANA, SOSIOLINGUISTIEK, SPRAAK, TAALVERSKEIDENHEID, TEKS, VERTEENWOORDIGENDHEID, WOORDEBOEKE

Introduction

More and more lexicographers realise the inevitability of using a corpus or corpora in the compilation of dictionaries. Leech (1991: 8) defines a corpus as "a sufficiently large body of naturally occurring data of the language to be investigated". Renouf (1987: 1) refers to the use of computers in the storing and analysis of corpora in his definition: "a collection of texts, of written or spoken words, which is stored and processed on computer for the purpose of linguistic research". McEnery and Wilson (1996: 24) similarly mention a reliance on computers in their definition of a corpus as "a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration". Leech (1991: 5), however, insists that a corpus has to be differentiated from an "archive",

> the latter being a repository of available language materials, and the former being a systematic collection of material for given purposes. A corpus draws upon the resources of an archive and therefore both are important. The systematic compilation of a structured corpus however is the primary objective.

Leech points to the systematicity of the collection of material as an important characteristic of a corpus. In this regard he does not conflate the substance for study with the tools used for its analysis and storage. However, whether the insistence on systematicity is crucial to the definition of a corpus may be subject to debate. Maybe "corpus" should be seen as textual data collected for linguistic research, usually stored in computers for quick analysis. But the fact that it is machine-readable, although important for its analysis, does not make it a corpus, for long before the introduction of computers there was much robust corpus research as exemplified by Kading's 1897 German corpus of some

11 million words for collating the frequency distribution of letters and sequences of letters.

For ages, lexicographers contended with ways and means of producing authentic and reliable reflections of the lexicon. Most of these lexicographers depended on their ability to remember words existing in the languages under study, something that De Schryver and Prinsloo (2000: 219) call "the random approach" and Kilgarriff (2000: 109) "the lexicographer's intuition". Others again, in the Oxford tradition, depended on readers, who searched texts for occurrences of words and submitted these for lemmatisation in the dictionary. For many years, these readers' contribution made the *Oxford English Dictionary* (OED) the unparalleled authority on the English language. More than any other English dictionary existing at the time, it included words from different genres and stylistic and regional varieties with reliable etymological information. Later developments in lexicography proved that readers were not very reliable sources of dictionary material since not only was their processing of data too slow, but it was also impossible for them to authoritatively deliver information on matters of frequency across texts and genres (see the *Longman Dictionary of Contemporary English* (1995[3]), the *Collins COBUILD English Dictionary* (1995[2]) or Kilgarriff (1997: 1)).

Over the past 20 years, a rapid growth of corpus lexicography has been witnessed, which was championed and popularised, more than by any other group, by the COBUILD (Collins Birmingham University International Language Database) group in Birmingham, led by John Sinclair. The earlier Birmingham school of corpus lexicography adhered strictly to the corpus as a source of dictionary evidence (Sinclair 1987). It was argued that corpora were the sole source of lemmatisation, frequency information and word lists. If a word was not in a corpus, it was not recognised as legitimate dictionary material. However, as corpus lexicography develops, there is a greater focus on its composition. Issues of balance and representativeness are continuously engaging theoretical and practical lexicographers. Researchers want to know the kinds of texts forming corpora and in what percentage they exist. These questions and concerns are not trivial since they put the credence and reputation of a dependency on corpus lexicography in question. Therefore the greatest challenge lies not so much in what can be obtained from a corpus, but rather in its construction.

Against this background, this article attempts to investigate the problems associated with the construction of corpora for dictionary making, particularly in many African contexts. It argues that some of the challenges facing the construction of robust corpora to be used in language research are the poverty of data, that is, the lack of texts to construct corpora representative of the different instances of language usage in a specific speech community. High illiteracy levels in African countries too pose great challenges to researchers hoping to collect written texts read by specific populations. Added to this, is the fact that, even where levels of literacy have increased, the literate members of a society

read and write texts written in English or French and not in their native languages. Even where such texts could be found in African languages, they mostly belong to a certain genre, like novels, plays and poetry, to the exclusion of other genres, like newspapers and academic texts. Even if the use of such data is attempted, the contention would still be with "sanitised" data, purified by the editorial policies and stylistic dictates of many publishing houses and newspaper offices, calling into question its authenticity as original and credible texts. The problem of representing speech still stands as one of the great challenges not only to African lexicographic research but also to research in many Western countries. At first, balance and representativeness must be investigated.

Balance and Representativeness

Most of the latest corpus-based lexicography researches consider issues of representativeness and balance (Ooi 1998) as marking standards of authenticity and robustness in corpus construction. A language corpus must be balanced and representative of the language from which it is extracted. By representativeness is meant "the extent to which a sample [text] includes the full range of variability in a population" (Biber 1993: 243), and as Summers (1993: 186) stresses "unless the corpus is representative, it is *ipso facto* unreliable as a means of acquiring lexical knowledge". Therefore, for a corpus to be representative, it must reflect the typical cross-spectrum of language use of a defined language community or period (see Ooi 1998: 49). But Summers's (1993) claim will be returned to since it raises considerable difficulties, particularly for corpus building in many African contexts and for certain linguistic theories.

A balanced corpus is one that includes proportions of a range of different text types of a language as they are reflected in the language studied.

The problem of what constitutes balanced and representative corpora still remains controversial. The selection of language from different genres to include in the language database is largely unresolved. The compilation of text must finally capture language from a specified population from which a sample is taken, which reflects how that particular language community uses language. This is significant since, as Summers (1993: 186, 190) points out, the results of corpora analysis must be generalised to the language community from which the samples were abstracted. Kennedy (1998: 94) argues for a pedagogical purpose to corpus research by noting that "high frequency of occurrence as determined by the analysis of texts should be a major determinant of lexical content of language instruction".

In a way, it is clear that issues of balance and representativeness of corpora are related. A representative corpus must reflect a representation of different genres of language use in a language community, while a balanced corpus should attempt to capture those different percentage levels or ratios in the way they occur in the specified language community. This obviously is difficult

to achieve, mainly because it is difficult to precisely know all the text types and their proportions of use in a population with its ever-changing dimensions. The difficulties are compounded when the building of a corpus of spoken language is attempted. As Kilgarriff (1997: 137) points out, dialectal varieties stand at different ratios to one another and should be represented within a corpus that attempts to accurately capture the language characteristics as a whole. There must also be contended with whether spoken texts can be accurately sampled and represented along the same lines as written texts. How many words are being looked for and what percentage of the spoken language do such words constitute? Whether spoken texts can be sampled in a representative manner is greatly questionable. Although a sample of Sengwaketse, Sekgatla, Sekwena and Sengwato can establish an acceptable representative percentage of the spoken form of these Setswana dialects, speech is a flood that refuses to be adequately accounted for numerically, for even when an attempt is made to quantify it, more of it is produced. It is Kennedy (1998: 62) who casts doubt on whether the representativeness of a corpus can confidently be argued for:

> In light of the perspectives on variation offered by several decades of research in discourse analysis and sociolinguistics, it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre or subject field or topic.

By "perspectives on variation" Kennedy refers to different speech varieties existing in a speech community. Problems are faced with sampling the standard against non-standard varieties, various sociolects covering status, gender, ethnicity, age, occupation, and others, different regional varieties, like Sengwaketse, Sekgatla, Sekwena, and Sengwato in the case of Botswana, and different registers like casual, formal, technical and others. Such variations are difficult to represent in a corpus. By noting this difficulty, Kennedy does not imply that representativeness should not be attempted, but that perhaps theoretically an attempt at representativeness may not conclusively capture the nuances of existing varieties as outlined by linguistic research.

Because of practical constraints, such as a shortage of time and money, the unavailability of machine-readable text, and copyright restrictions, it is not always possible to assemble the representative and balanced corpus ideally wanted. It is precisely these problems that stand out as some of the major stumbling blocks particularly in the African context of corpus construction.

Two English Corpora

This section will bring to the fore the composition of more influential corpora which have been considered by many lexicographers and numerous language researchers as examples of "good" corpora. What should particularly be noted is the percentage of spoken text against written text since it is central to subsequent arguments made in this article.

In 1991, COBUILD launched the Bank of English (BoE), which currently has over 450 million words and continues to grow as more material is published and deposited into it. It forms the basis for the compilation of the COBUILD dictionaries (Sinclair 1991). The BoE does not claim any balance or representativeness of usage, but it does claim to provide evidence of the way everyday English is used. The spoken word is represented by transcriptions of everyday casual conversation, radio broadcasts, meetings, interviews, discussions, etc. However, even with the seemingly impressive 450 million words, the BoE is only a small sample of human speech produced on a daily basis.

The other corpus that has extensively been used is the British National Corpus (BNC) which has "a 100 million collection of samples of written and spoken British English of the late twentieth century from a wide range of sources designed to represent a wide cross-section of current British English both written and spoken" (BNC website). Ninety per cent of its composition consists of written texts including amongst other kinds of texts, extracts from regional and national newspapers, academic books and popular fiction, essays and letters (75% from informative writing such as fields of applied science and commerce and finance; 25% from imaginative, i.e. literary and creative, works). Spoken texts, which include unscripted informal conversation, government meetings and radio shows, constitute only 10%. The corpus has 4 124 texts, of which 863 are transcribed from spoken conversation and monologues. It was developed by the Oxford University Press, the Longman Group Ltd, Chambers Harrap, the Unit for Computer Research on the English Language (Lancaster University), the Oxford University Computing Services, and the British Library Research and Development Department. It has been used for a wide variety of research in language, including lexicography, as in the making of the third edition of the *Longman Dictionary of Contemporary English*.

The Primacy of Speech

It is a widely held fact that children speak before they write and that speech is primary to human communication (Aitchison 1998). It is also generally agreed that in a speech community the spoken word exists in abundance compared to written texts. Taking these linguistic arguments as base and applying them by implication to issues of balance and representativeness, it can be concluded that if corpus construction has to reflect the different ratios between spoken and written texts, different text genres and various dialectal varieties, then the percentage of spoken language has to be much greater than that of written language in a corpus. Such a greater occurrence of spoken over written texts would approximate the ratios of written and spoken texts in the real world and would be likely to produce corpora that accurately represent language as used in a speech community. However, in none of the corpora discussed in the previous section the percentage of spoken texts exceed that of written texts. Ten per cent of the data of the BNC consists of spoken texts. Leech et al. (2001: 1)

recognise the inadequacy of speech in the BNC which contains about 90 per cent written data and 10 per cent spoken data:

> Although spoken language, as the primary channel of communication, should by rights be given more prominence than this, in practice this has not been possible, since it is a skilled and very time-consuming task to transcribe speech into the computer-readable orthographic text that can be processed to extract linguistic information. In view of this problem, these proportions were chosen as realistic targets which, given the size of the BNC, are also sufficiently large to be broadly representative.

According to Leech et al., the percentage of the speech text in the BNC was reached by determining what was possible to the compilers and not by making allowance for the proportion of speech to written language in a speech community. If corpora do not reflect in their composition that the spoken word is more common in real life than the written text, it calls the power and authority of corpora as sources of evidence for linguistic research in question and opens them to possible doubt.

A Newspaper versus the Purchase of a Pair of Shoes

While Kennedy (1998: 63) acknowledges the common occurrence of speech in daily discourse, he argues against it by noting:

> No one knows what proportion of the words produced in a language on any given day are spoken or written. Individually speech makes up a greater proportion than does writing of the language most of us receive or produce on a typical day. However, a written text (say in a newspaper article) may be read by 10 million people, whereas a spoken dialogue involving the purchase of a pair of shoes may never be heard by any person other than the two original interlocutors.

Kennedy introduces a dimension to corpus creation that raises great controversy. It is true that a newspaper is likely to be read by many people and that its circulation can be obtained from reliable sources. However, it is not true that newspaper buyers equally read different sections of a newspaper. Some readers pass over the business section, classifieds, cartoons, letters to the editor and many other sections. Although circulation numbers might be available to assist corpus builders sample newspaper text, they are heavily unreliable because though a newspaper might be selling 40 000 copies, those copies might be read by over 100 000 individuals while others might be bought and never be read!

A similar point may be made that although lots of corpora depend on published texts, there is indeed no guarantee that such texts are widely read (or read at all). This is particularly so in the Setswana language situation where the majority of Batswana do not read Setswana texts, except at elementary school. Kennedy (1998: 52) suggests that to fix this problem "best seller lists, library

lending, statistics and periodical circulation figures can only partially reflect receptive use and influence". For many readers of texts in African languages "best seller lists, library lending, statistics and periodical circulation figures" are foreign concepts unheard of in African literature. Kennedy's use of "partially" is an indication of the immensity of problems surrounding attempts to construct corpora on the basis of common and influential texts. If "receptive use and influence" are taken as determinants of text inclusion in a corpus, varying *degrees* of such use and influence will have to be contended with. School textbooks and creative texts read by thousands of students across the country would be in use more than a library text which is rarely read. How would such a distinction be represented in a corpus? Is it not the case that textbooks would have been read more widely and therefore their texts should somehow reflect the fact that they have been seen more than other texts? This argument can be pursued further. This would mean that a sign reading "Welcome to Gaborone" would make "welcome" "to" and "Gaborone" very high in a frequency list since they have been seen many times by many people entering the city. Words like "stop", used on traffic signs and seen again and again, would be amongst some of the most common terms. Such conclusions would certainly distort the way language is used since the word "stop" does not occur frequently in daily discourse. The problem of how its commonality is represented in a corpus therefore remains.

It would appear that Kennedy's argument against spoken texts on the basis that they are private while written texts are in the public domain, is not very convincing but rather raises new problems and challenges. Spoken texts are as important as written texts in corpus creation and attempts should be made to reflect approximate ratios between written and spoken texts, ratios which are problematic to establish.

Can Anything Good Come out of Spoken Texts?

Much would be lost if a corpus does not reflect spoken texts in their right ratios. One such loss would be instances of borrowing common in written texts but censured by editors and publishers in communities where there is much code-switching, language contact and borrowing, particularly in many African countries where both native languages and former colonial languages like English or French are used. An observation of spoken Setswana texts will show a high degree of borrowing from English and Afrikaans. Borrowing is here used in Nevejina's (1998) sense of "the element of an alien language which is carried from one language to another as a result of language contact". The documentation of this phenomenon in Setswana is not recent. Cole (1955) noted words like *beke (week)* "week", *baki (baadjie)* "jacket", *gouta (goud)* "gold", *heke (hek)* "gate", *hempe (hemp)* "shirt", *kofi (koffie)* "coffee", *pena(e) (pen)* "pen", *peipe (pyp)* "pipe", *sukiri (suiker)* "sugar" from Afrikaans and *baesekele* "bicycle", *buka* "book", *ofisi* "office", *šeleng* "shilling" from English. There are other more recent borrowings

which reveal a certain layering in the nature of what is considered borrowed words. For instance, many Setswana speakers are not aware that *baki* and *heke* are borrowed from Afrikaans, while *jakete* "jacket" and *geiti*(?)† "gate" are recognised as borrowings from English. The result is that *baki* and *heke* are considered by some as "good" established Setswana, while the more recent borrowings *jakete* and *geiti* are condemned. Spoken Setswana is interspersed with instances of code-switching and borrowing in sentences such as the following:

Go shapo! (Good-bye!)
O tsile in the afternoon. (He came in the afternoon.)
Ke bra/sistere ya gagwe. (It is his brother/sister.)
O apere jase. (He is wearing his coat.)

Greater levels of code-switching and borrowing are also evident in naming the days of the week and the months of the year, and in naming the numerals. For instance, many Setswana speakers would say *Monday* or *Mantaga* (from Afrikaans *Maandag), Tuesday, Wednesday ... Saturday* or *Sateretaga* (from Afrikaans *Saterdag*) and *Sunday* or *Sontaga* (from Afrikaans *Sondag*). Reference to the months by Setswana speakers is also usually in English, and most would have difficulties in saying them in Setswana. In many instances Batswana speakers use the English instead of the Setswana names for the numerals. Many speakers would find it difficult saying 1 567 in Setswana since numbers are generally expressed in English. It is common for Batswana to use one, two, three, fifteen, two thousand, or one million in their speech instead of the Setswana terms.

These are some of the problems a Setswana lexicographer would have to face if he/she depends on a corpus with greater levels of spoken data rather than a corpus with written data or with smaller levels of spoken text. The lexicographer would grapple with decisions on the kind of borrowed words that should be lemmatised and the kind of stylistic information that should be derived from borrowed words. Obviously the kind of dictionary being compiled would influence such decisions: whether it is monolingual or bilingual, for learner's or general use, of table or pocket size, etc.

Dealing with borrowings and code-switching in lexicography is not a new phenomenon. Lichtenberk (2003) considers the question of which borrowed words qualify as belonging to the borrowing language and therefore deserving inclusion in a dictionary. In his report of the dictionary of Toqabiqita, an Austronesian language spoken in the Solomon Islands, he points out that the central point in determining the wordlist of a dictionary is "the prospective audience", that is, the intended users of a dictionary, and "its expectations", that is, the purposes the dictionary will be expected to serve in the society. This view is shared by Zgusta who contends that decisions of what to include are determined by "fundamental decisions concerning the type of dictionary which is to be prepared" (Zgusta 1971: 243). For instance, if the dictionary is intended to contribute to historical and comparative studies, it may list archaic and obsolete words while the inclusion of loanwords may prove to be of interest to pho-

nologists. But the greater part of Lichtenberk's article is devoted to a discussion of the inclusion or not of loanwords in the dictionary of Toqabaqita. There are comparisons which may be drawn between Setswana and Toqabaqita. Lichtenberk is confronted with a language situation where he has to decide whether to include Pijin words in the dictionary of Toqabaqita since some of them fit the phonological and phonotactic constraints of Toqabaqita while others do not. A similar challenge faces Setswana: whether to include borrowings from English or Afrikaans. Like Setswana, Toqabaqita does not permit consonantal clusters or syllable-final consonants and has a simple syllable structure of CV and V. This characteristic of Toqabaqita guides Lichtenberk (2003: 395) in deciding what to include:

> Pijin words used in Toqabaqita are listed provided they fit the phonological and phonotactic patterns of Toqabaqita, either because they fit them already in Pijin or because they have been accommodated to them. Words which do not fit the patterns are not listed.

According to this principle, certain words in common use are excluded, because they are, in Lichtenberk's view, instances of code-mixing. Not satisfying the phonotactic constraints of Toqabaqita, they are not listed in the dictionary. Similar to the Setswana situation, code-mixing in Toqabaqita is common. Lichtenberk (2003: 396) argues:

> Considering such words to be part of Toqabaqita lexicon would amount to claiming that the phonological inventory and the phonotactic patterns of the language have undergone some major changes.

Therefore Lichtenberk decided to restrict the matter of code-mixing to the front matter where the common but non-accommodated words would be listed. There are also problems concerning pairs of words which, though accommodated from Pijin, have variants which do not conform to the phonotactics of Toqabaqita. In these instances, the variant that does not conform to the phonotactic constraints is not listed. But it gets more complicated when the non-accommodated variant is more common than the accommodated one. In such cases, Lichtenberk ignores the most frequently used word, since it violates the phonotactic constraints of the language, and instead chooses to enter the less common one on the principle that the non-accommodated variant, though frequent, is an instance of code-mixing.

Lichtenberk (2003: 396) develops further principles which determine what to list. These are:

1. "Words that belong in well-circumscribed and relatively small sets are not listed if some other members of the same set do not occur in an accommodated form and so are not listed."

2. "A Pijin word that has been encountered only once is not listed even if it fits the phonological and phonotactic pattern of Toqabaqita."

The question of what has to be listed in the dictionary raises an issue of what are the boundaries of the lexicon of a language. Lichtenberk therefore divides Toqabaqita words into three categories: (a) native Toqabaqita words, (b) accommodated borrowings from Pijin, and (c) Pijin words used without being accommodated. Lichtenberk (2003: 397) concludes that:

> Only the first two types are to be listed in the dictionary, which amounts to saying that only those words are part of the Toqabaqita lexicon, while the non-accommodated words are not.

He gives proper criticism to his approach when he says:

> The principle, while explicit and applicable in a straightforward way, is nevertheless arbitrary. It gives priority to the phonological and phonotactic patterns of Toqabaqita over usage. Pijin words that are not accommodated are, by fiat, placed outside the circumference of the Toqabaqita lexicon, although by virtue of their usage they could be inside.

Some of Lichtenberk's principles are better not followed, particularly the preference of phonology over usage. Take for instance his first principle for listing sets of words. Such sets include the names of numerals, the days of the week and the months of the year. This principle creates problems for accounting for the class days of the week in Setswana.

Days such as *Sateretaga, Sontaga* and *Mantaga* are colloquial and more common in spoken than in written language, while *Matlhatso, Tshipi* and *Mosupologo* are common in written texts and formal addresses. This stylistic information is significant, particularly in dictionaries which attempt to achieve a broader coverage and a fuller understanding of a word's meaning and usage. When both formal and informal terms are lemmatised, they may provide, except stylistic information, significant information for future research on when a word has entered the language or changed its meaning.

Additionally, cases where certain terms, although known in the native language, are rarely used in speech, but are replaced by borrowings and code-switchings, cannot be ignored. This is particularly true of numerals where sentences such as *O rekisitse dinamune di le ten* "He has sold ten oranges" and *Mmiting o ka thene kamoso* "The meeting is at ten tomorrow" are found. In these examples, the speaker has chosen the English word *ten*, instead of the Setswana term *lesome/some*. The transcription of the term *ten* as either *ten* or *thene*, as in the above examples, is based on the theoretical question of whether such a term has gained currency as an instance of borrowing or of code-switching. Are lexicographers to assume that such language usages do not exist in the language and that they do not have any relevance to dictionary compilation? Any answer to these questions would lead to disagreements among lexicographers.

It is important to note that although *lesome* and *ten* refer to the same number, they usually have different usages. *Lesome* would be more common amongst the elderly, in written texts and in very formal "tribal" meetings. *Lesome* is also used to refer to P1 (one Pula). *Ten* is much more common in colloquial exchanges, spoken language and amongst the educated.

This hopefully shows the importance of including greater occurrences of spoken text in a corpus since the spoken word occupies a greater level of language usage in human communication. Next the lack of data and the available data for lexicographic research will be considered.

The Poverty of Data

While Western lexicographers enjoy an abundance of data for the construction of huge corpora running into millions of texts of different genres covering newspapers, magazines, novels, academic texts, parliamentary pronouncements, and legal texts, African lexicographers work under great constraints because of the lack of data. Unlike their Western counterparts, they usually do not possess the luxury to be discriminative and selective of texts in electronic form since in the first place such texts are nonexistant. Many African countries do not use their indigenous languages in parliamentary debates, the publication of laws, instruction at schools and journalistic publications. This is certainly the situation in Botswana where there exist very little text in Setswana. In comparison with English, there are very few Setswana novels and plays. There is also little instructional material in Setswana for lower primary school levels and virtually none for higher education. The only newspaper which wrote exclusively in Setswana, *Mokgosi*, closed down in 2005 because of poor advertising and sales. Another, *Mmegi*, which had a three and a half page Setswana insert, called "Naledi", also no longer publishes these pages. These low levels of written text give an idea of the seriousness of the problem confronting African lexicographers if they were to adopt the Western approach to corpus creation. They face practical constraints similar to those outlined above, such as a shortage of time and money, the unavailability of machine-readable text, and copyright restrictions.

Although there are few written texts in African languages, their existence does not guarantee that they are accessible to both native speakers and corpus researchers, or that the literate native speakers of the language read them. Many literate Africans rarely read texts in their own languages, although they may communicate extensively in them. The reason is not only because there is not enough written material in the African languages, but also because there is no culture of reading African literature in many African communities. African lexicographers therefore face great hurdles in attempting to access both written and spoken texts for corpus construction. In cases where they have access to

written texts, they run the risk of basing their research on the shaky foundations of the attitudes of language purists and prescriptivists who remain wedded to a linguistic world that has never existed.

This leads to the question of whether many corpora created for lexicographic research in Africa could be considered balanced and representative to the extent that they could be taken as bases for generalisations about the general language. This is greatly doubtful since most African corpora are biased towards one language variety as African languages are generally not used to render a variety of social contexts like the writing of laws, medical texts, government or official communications, academic books and business texts. Although these languages may not be used for *writing* about these topics and areas, in many occasions they are used to *speak* about them. A corpus of an African language constructed on a dependency on spoken texts is, however, likely to cover a rather restricted scope of language usage partly because of the unavailability of machine-readable data (MRD). It is also a well-known fact in natural language processing and computational linguistics that the transcription of spoken text is time-consuming and expensive, and cannot be afforded by many researchers, both Western and African. This further narrows the amount of text that could be included in many African languages corpora.

The Sanitised Data

Still on issues of written text, consideration need to be given to the involvement of publishers and editors and the power of stylebooks on the written word, resulting in what can be called "sanitised data". Many publishers and editors have very rigid principles of which words should be used in their publications. They are heavily prescriptive, as in the newspaper *Mokgosi*, for example, where the rare Setswana words *Mosupologo* (Monday), *Tshipi* (Sunday), *dira* (work, v.), and *kgwele* (ball) were preferred to the much more common *Mantaga*, *Sontaga, bereka,* and *bolo* respectively. Such preferences show the biased prescriptive stance adopted by numerous publishers and editors who believe that borrowed language is not authentic and not part of the language. Their control of language does not reflect how the people use language, but rather reflects *how they wish it to be used.* A dependency on such language for the construction of corpora brings serious questions to the kind of corpora whose results have to be generalised to the entire language. This is especially so since corpora provide information about what to include and exclude, guide the lexicographer towards sharper sense distinction, and assist in selecting corpus-based examples. While "sanitised data" may be unavoidable, it is greatly unsatisfactory for dictionary research where generalisations about language use must be made. Instead, it should be considered together with spoken texts to obtain a clearer picture of the language use of a speech community.

Conclusion

In this article, an attempt has been made to show that, while corpus research remains one of the most useful approaches to language research in that it can speedily offer information for addressing language-related issues and problems, a critical look at the process of corpus construction and inclusion would help determine if generalisations drawn from its results can be trusted as a true reflection of language use. The bias against spoken texts, for whatever reason, results in the greatest weakness of many corpora. The African context is unique in that, unlike Western communities, many African countries do not use their languages for academic purposes, in the media, and for governmental and official communication, making MRD difficult to access. Slow developments in computer software automatically changing spoken text into written text means that approaches to building corpora of spoken texts may remain challenged for a long time to come.

The future of a rigorous corpus research in Africa appears to be to approach issues of representativeness and balance with great caution. Kilgarriff and Grefenstette (2003: 334, 340), echoing Kennedy (1998: 62), state that " 'representativeness' begs the question, 'representative of what?' " since, as they point out, "a corpus comprising the complete published works of Jane Austen is not a sample, nor is it representative of anything else". Although considered a language event, it is still unclear whether it is a matter of language production or of language reception. With the uncertainty surrounding matters of representativeness and balance, and with no convincing research of what precisely constitutes corpus material, it can be concluded with Kilgarriff and Grefenstette's (2003: 343) sentiments on web language that:

> The Web is not representative of anything else. But nor other corpora, in any well-understood sense. Picking away at the question exposes how primitive our understanding of the topic is and leads inexorably to larger and altogether more interesting questions about the nature of language, and how it may be modeled.

For many African lexicographic projects there is a need to build organic corpora along the lines of the Bank of English (that currently has over 450 million words and continues to grow), which, in spite of attempts to update the corpus frequently to maintain a balance between written and spoken forms, does not claim to be balanced and representative. Such an approach would be sensitive to the current situation of many African languages that require a certain systematicity in their study, but would also recognise the fact that certain demands and expectations common to Western lexicography cannot be met in the African context. What goes into the compilation of a corpus must also be accounted for as much as what is extracted from it. In addition to pursuing corpus research, there is also a need for African lexicographers to look towards old and new approaches within theories of word meaning and analysis that would assist them in the collection and classification of words. A case in point is

WordNet, a University of Princeton's systematic analysis of words, whose design and execution were inspired by psycholinguistic theories of human lexical memory. It is crucial that lexicographers should not lose direction of what they want to achieve by sacrificing it to the quest of theoretical substantiality. The aim is to achieve the knowledge base of the lexical system of a language.

Note

† A question mark is put after "geiti", borrowed from the English "gate", since Setswana does not have the voiced, velar plosive as part of its sound system, which in this instance occupies the initial word position in "geiti". There is therefore no agreed orthographic representation of such a sound in Setswana.

References

Aitchison, J. 1998[4]. *The Articulate Mammal: An Introduction to Psycholinguistics*. London: Routledge.

Biber, D. 1993 Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics* 19(2): 219-241.

Cole, D.T. 1955[9]. *An Introduction to Tswana Grammar*. Cape Town: Longman

De Schryver, G.-M. and D.J. Prinsloo. 2000. Towards a Sound Lemmatisation Strategy for the Bantu Verb through the Use of *Frequency-based Tail Slots* — with Special Reference to Cilubà, Sepedi and Kiswahili. Mdee, J.S. and H.J.M. Mwansoko (Eds.). 2001. *Makala ya Kongamano la Kimataifa Kiswahili 2000: Proceedings*: 216-242, 372. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam. Also available at: <http://tshwanedje.com/publications/kiswahili2000fbts.pdf>.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

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

Kilgarriff, A. 2000. Business Models for Dictionaries and NLP. *International Journal of Lexicography* 13(2): 107-118.

Kilgarriff, A. and G. Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29(3): 333-347.

Leech, G. 1991. The State of the Art in Corpus Linguistics. Aijmer, K. and B. Altenberg (Eds.). 1991. *English Corpus Linguistics: Essays in Honour of Jan Svartvik*: 8-29. London: Longman.

Leech, G., P. Rayson and A. Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Pearson Education.

Lichtenberk, F. 2003. To List or Not to List: Writing a Dictionary of a Language Undergoing Rapid and Extensive Lexical Changes. *International Journal of Lexicography* 16(4): 387-401.

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

Moe, R. 2003. Compiling Dictionaries Using Semantic Domains. *Lexikos* 13: 215-223.

Nevegina, S.B. 1998. Some Problems of Borrowing in the Russian Language. *Vestnik Omskogo Universiteta* 1: 72-75.

Ooi, V.B.Y. 1998. *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Renouf, A. 1987. Corpus Development. Sinclair, J.M. (Ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*: 1-40. London: Collins ELT.

Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J.M. (Ed.). 1995[2]. *Collins COBUILD English Dictionary*. London: HarperCollins.

Summers, D. 1993. Longman/Lancaster English Language Corpus — Criteria and Design. *International Journal of Lexicography* 6(3): 181-208.

Summers, D. (Ed.). 1995[3]. *Longman Dictionary of Contemporary English*. Harlow: Longman.

Websites

Bank of English: <http://www.titania.bham.ac.uk/docs/svenguide.html>.

The British National Corpus: <http://www.natcorp.ox.ac.uk/what/index.html>