Disregarding the Corpus: Headword and Sense Treatment in Shona Monolingual Lexicography*

Webster M. Mavhu, *African Languages Research Institute (ALRI), University of Zimbabwe, Harare, Zimbabwe (vhezh2000@yahoo.com)*

Abstract: With specific reference to Shona monolingual lexicography, this article discusses how corpus-based lexicographers might, in some instances, decide not strictly to adhere to the corpus when it comes to headword and sense treatment. The writer is a member of the African Languages Research Institute (ALRI), formerly known as the African Languages Lexical (ALLEX) Project. ALRI is a nonfaculty interdisciplinary unit dedicated to research on and the development of African languages in Zimbabwe. The writer is part of the six-member team that compiled the now published Shona monolingual, synchronic, medium-sized and general-purpose dictionary *Duramazwi Guru ReChiShona* (2001). The article originates from the writer's experience of working on this dictionary. The article highlights the fact that being corpus-based does not necessarily imply being corpus-bound.

Keywords: CORPUS, CORPUS-BASED, FREQUENCY, HEADWORD, LEXICOGRAPHY, SENSE, SHONA, SLANG, SYNONYMS

Opsomming: Verontagsaming van die korpus: Trefwoord- en betekenisbehandeling in die Sjona- eentalige leksikografie. Met spesifieke verwysing na die Sjona- eentalige leksikografie bespreek hierdie artikel hoe korpusgebaseerde leksikograwe in sommige gevalle kan besluit om nie streng by die korpus te bly wanneer dit kom by trefwoord- en betekenisbehandeling nie. Die skrywer is 'n lid van die African Language Research Institute (ALRI), vroeër bekend as die African Languages Lexical (ALLEX) Project. ALRI is 'n niefakulteits-interdissiplinêre eenheid wat hom beywer vir navorsing oor en die ontwikkeling van die Afrikatale in Zimbabwe. Die skrywer is deel van 'n span van ses lede wat die reeds gepubliseerde Sjona- eentalige, sinchroniese, middelgroot en meerdoelige woordeboek *Duramazwi Guru ReChiShona* (2001) saamgestel het. Die artikel het uit die skrywer se ervaring van werk aan hierdie woordeboek ontstaan. Die artikel belig die feit dat korpusgebaseerdheid nie noodwendig korpusgebondenheid impliseer nie.

Sleutelwoorde: BETEKENIS, FREKWENSIE, KORPUS, KORPUSGEBASEER, LEKSIKOGRAFIE, SINONIEME, SJONA, SLENG, TREFWOORD

* This article is based on a paper presented at the Seventh International Conference of the African Association for Lexicography, organized by the Dictionary Unit for South African English, Rhodes University, Grahamstown, 8–10 July 2002. Information on frequency counts appearing in this article was provided by Daniel Ridings.

1. Introduction

Corpora may be compiled (and used) for many different purposes in language research, including their lexicographic use. The majority of ALRI's research activities are either corpus-based or corpus-aided. In fact, research in corpus work is one of ALRI's basic and essential research areas (Chimhundu 2000: 5). The ALRI team's research activities have so far culminated in the development of corpora for two of Zimbabwe's main languages, Shona and Ndebele. Work is currently under way to develop corpora for four of ALRI's prioritised and Zimbabwe's officially recognised 'minority' languages, Kalanga, Nambya, Tonga and Shangani.

2. A Brief Discussion of the Shona Corpus

The contents of the Shona corpus came from oral and written data. For oral data collection, undergraduate Shona students were sent out to tape-record interviews on almost all aspects of life, in all Shona-speaking districts of Zimbabwe and from males and females of different age groups. In the process of systematically collecting this oral material, details on the context of the discourse, date of interview, physical location, topic, setting and other relevant details were recorded. Extra-linguistic features such as hesitations, coughs and pauses were also recorded and marked. Some written data from Shona texts was also introduced into the corpus. The material then underwent the processes of transcription, encoding, proofreading, tagging and parsing. These processes are the main stages of corpus design. Transcription is the process of reducing an oral text to writing. Encoding is the keying in of data into a computer. Scanning refers to the process of electronically recognising written material that appear as hard copies and saving them as soft copies. Tagging is the process of assigning a specific code to each word in a text. Parsing involves checking tagging errors.

A discussion of how the above-mentioned processes were employed to produce the Shona corpus must be left to a more detailed report. Suffice it to say that at this point, oral material constitute seventy percent of the 2 600 000 running words that are in the current Shona corpus and written material thirty percent (Chabata 2000: 79). It should be noted that the Shona corpus could be viewed as a monitor corpus, since it is open-ended. Texts are continuously being added to it so that it gets larger and larger as more samples are added. A monitor corpus is important for ALRI, which specialises in dictionary making. In fact, monitor corpora, according to McEnery and Wilson (2001: 30), 'are primarily important in lexicographic work for they enable lexicographers to trawl a stream of new texts looking for occurrence of new words or for changing meanings of old words'.

The Shona corpus was utilised in the production of two Shona dictionaries: *Duramazwi ReChiShona* (DRC) (1996) and *Duramazwi Guru ReChiShona* (DGC) (2001). Whilst the compilation of DRC was corpus-aided, that is, its compilation was assisted with material from the Shona corpus, that of DGC was corpus-based.

3. The Implications of Being Corpus-Based

Before discussing the degree to which DGC was corpus-based, it is perhaps necessary to survey the debates that have been conducted with regard to the idea of relying on a corpus in linguistic research. Reliance on a corpus would be biased towards an empiricist approach to the study of language that is dominated by the observation of naturally occurring data, typically through the medium of a corpus. Rationalists (notably Chomsky) have maintained that this approach has its limitations. Their main argument is that no one corpus can ever be regarded as a significant record of any language. Perhaps such an argument used to make sense at a time when texts were put on slips of paper and where relevant information could only be accessed manually. Then, there were only very small corpora. However, this is no longer the case. As McEnery and Wilson (2001: 31) put it, 'nowadays, the term "corpus" almost always implies the additional feature, machine-readable'. At present, researchers are coming up with machine-readable corpora that contain several billions of running words that can easily be searched and manipulated.

A corpus has the advantage that corpus-based observations are intrinsically more verifiable than introspectively based judgements. Empiricists observed that the type of sentence typically analysed by the introspective linguist is far removed from the type of evidence we typically tend to see occurring in the corpus. Empiricism maintains that the corpus does not only seem to be a more reliable source of frequency-based data but also provides the basis for a much more systematic approach to the analysis of language. There is, therefore, no doubt that a corpus is an essential linguistic tool. Since DGC was intended to be corpus-based, it meant that all headwords, senses, citations and other relevant linguistic information that would be required in the compilation of the dictionary would come from the Shona corpus. Whilst the Shona corpus was heavily relied upon for the majority of these items, there were instances when the editors of DGC had to disregard this corpus as shall be illustrated in the following sections.

4. Disregarding the Corpus in the Treatment of Headwords

At times the editors of DGC disregarded the Shona corpus in their treatment of the words that they selected as headwords for the dictionary. This was particularly the case in two areas:

4.1 Headword Selection

Headword selection is one of the most crucial stages in compiling dictionaries because it is during this stage that the contents of a dictionary are determined. Comprehensive criteria defining the process of headword selection has to be set up and should be detailed in the style manual that guides the compilation of any dictionary. If headword selection is corpus-based, as was intended in the compilation of DGC, lexicographers have to rely heavily on frequency, that is, the number of times a word appears in the corpus. Thus, the most frequent words should be selected first, then the less frequent and ultimately the least frequent ones. Since it is not practically possible to include all the words of a language in a dictionary, it follows that some words have to be left out. DGC was intended to contain approximately 50 000 words. It was not possible to go beyond this number to prevent the dictionary from becoming too voluminous in size, too expensive to produce and also too highly priced for its target users.

It was, however, difficult solely to rely on the corpus when deciding on which words to include in or to exclude from the dictionary. The following are the 20 most frequent words in the Shona corpus, listed in descending order according to their frequency.

46 021 *kuti* (that, so that, in order that)
25 272 *kana* (when, although, even, or, if, whether)
10 505 *asi* (but, except)
9 197 *zvino* (now)
8 460 *munhu* (person)
8 259 *saka* (hence, consequently, therefore, for this reason)
7 840 *here?* (is that so?)
7 064 *vanhu* (people)
5 916 *chete* (only)
5 781 *mwana* (child)
5 766 *uyu* (this one)
5 110 *ari* (who is)
5 018 *ini* (me)
4 660 *nokuti* (because)
4 401 *iri* (this one)
4 280 *iyi* (this one)
4 137 *sei?* (how?/why?)
4 093 *izvi* (these)
4 011 *vana* (children)
3 997 *iye* (him/her)

The words in the above frequency list are of not much value to a Shona lexicographer, especially a monolingual one. Neither are they of much value to the target audience of DGC that happens to be mother-tongue speaker-writers of the Shona language. The reason is that they are mostly function words. In fact,

the most frequent word in the Shona corpus, *kuti*, which occurs more than 46 000 times, is a conjunctive. No verbs are found in the list and only a couple of nouns such as *munhu*, *vanhu*, *mwana* and *vana*.

As far as headword selection is concerned, it would not make much sense, at least in monolingual Shona lexicography, to prioritise the most frequent word *kuti* over say, for example, either *rufu* (death) which occurs 258 times or *ivhu* (soil) which occurs 254 times in the Shona corpus. Thus, by prioritising certain less frequent lexical items over those that were the most frequent, but were suppletive and function forms, the Shona corpus was disregarded.

4.2 Presentation of Synonyms

Another instance where the editors of DGC did not strictly adhere to what features in the Shona corpus, is in the presentation of synonyms. According to Jackson (1988: 65), two words are said to be synonyms if they have the same meaning. He also notes that since the term 'meaning' can only be understood contextually, synonymy also needs to be defined in terms of contexts of use. He then proceeds to give a rather revised definition of the term 'synonym'. He maintains that two words are synonyms if they can be used interchangeably in all sentence contexts (Jackson 1988: 65). Examples of Shona synonyms would be *-mhanya* and *-rumba* both of which mean 'run'.

As a way of saving space, it had been decided that synonyms were to be defined only when it was deemed necessary. Otherwise, the more commonly used form would carry the definition and the less commonly used one(s) would be cross-referred to the commonly used form. Where in doubt, the strength of the corpus would help to determine the main headword (Mawema 2000: 218). This would be through the use of the frequency counts that have already been mentioned. The frequency counts were, however, disregarded in some cases, for example, when an indigenous word competed with an adoptive.

The general desire of the editors of DGC was to promote indigenous words as much as possible. However, at times indigenous words appear less frequently than adoptives in the corpus as can be seen from the following example. The English noun 'nurse' is rendered by two equivalents in the Shona language: *mukoti* and *nesi*. The former is indigenous whilst the latter is borrowed. Following the principle of prioritising indigenous words over adopted ones, the editors chose *mukoti* to carry the definition whilst they cross-referred *nesi* to *mukoti*. If one looks at the frequency counts in the Shona corpus, one finds that *nesi* appears more frequently (61 times) than *mukoti* (51 times). This example shows that the editors of DGC disregarded the Shona corpus in the presentation of some synonyms.

5. Disregarding the Corpus in the Treatment of Sense

The editors of DGC at times disregarded the Shona corpus in their treatment of

sense. This is particularly noticeable in two areas:

5.1 Sense Selection

As has already been noted, a corpus is useful in dictionary making since it provides certain senses of words that lexicographers might not think of among themselves. In this regard the Shona corpus was quite useful to the editors of DGC. There were, however, instances when some senses that appear in the Shona corpus were deliberately omitted despite their occurrence in it. This was particularly so with some terms or senses that can be regarded as slang. According to Flexner and Wentworth (1975: vii), 'slang is an ever changing set of colloquial words and phrases that speakers use so as to establish group identity and solidarity'. It was noted earlier on that the Shona corpus comprises oral material that came from different groups of Shona speakers and that focus on various aspects of life. Among these groups of people were youths using Shona slang. Hence some slang found its way into the Shona corpus.

Editors of DGC were quite cautious when dealing with Shona slang. They decided to enter into the dictionary only slang that has become an integral part of the Shona language. They resolved to omit slang that was considered ephemeral. Thus, some senses that can be regarded as Shona slang, and were frequent in the corpus but were considered to be of ephemeral use, were omitted. An example is the term *chitunha* which in typical Shona refers to a corpse, the body of a dead human being. Shona slang extends the term to refer to a slaughtered chicken, the result of metonomy, a type of semantic transfer whereby one entity is taken to stand for another on the basis of some contextual relationship (Bonvillain 1993: 75). Although the 'second' sense appears in the Shona corpus, it was omitted in DGC for fear that it will be short-lived. A practice such as this disregards the Shona corpus.

5.2 Ordering of Senses

During the defining process, in cases where there were two or more meanings for a headword, senses were to be ranked, with the basic meaning appearing first. Where the basic meaning could not be ascertained, usage would determine the ranking of definitions. The literal sense would precede the metaphorical, idiomatic and proverbial senses. Frequency of occurrence would be considered with the aid of the corpus. The corpus was, however, only useful when there were two or three senses being dealt with and when all the senses could be found in it. In the case of some verbs, for example, the senses would sometimes be so many that it was difficult and problematic to handle them.

An example of such a problematic lexical item is the verb *-bata* (lit. touch, hold, catch). In addition to its basic senses, the verb has several other metaphorical and idiomatic ones. In DGC, the senses of the verb are listed as follows:

1. to hold/touch
2. to catch
3. to work somewhere
4. to do your work wholeheartedly
5. to be firm (as in a planted seedling)
6. to attack (as in disease)
7. to be tight (as in small clothing)
8. to arrest (as in arresting by the police)
9. to discover someone doing something bad
10. to understand something
11. to pin (as in pinning a shirt)
12. to be firm and strong (as in something being made/being constructed)
13. to catch (as in catching a bus)
14. to treat (as in treating a subordinate)
15. to face a hindrance
16. to have a lot of money
17. to catch up
18. to be dense (as in a forest)

Altogether, there are eighteen senses listed. Whilst some of them occur in the corpus several times, some do not. However, some of those that occur nil times are also listed in the dictionary and, more so, even before some of those that occur several times. This is because they were found to be more important and closer to the primary meaning of the verb. This example of sense treatment also shows an instance where the corpus was disregarded.

6. Conclusion

This article has shown that although the corpus is a very useful tool, especially in aiding some lexicographic decisions in corpus-based lexicography, there are times when lexicographers have to disregard it during the compiling process. It has highlighted the fact that being corpus-based does not necessarily have to imply being corpus-bound. This has been shown through focusing on headword and sense selection with specific reference to corpus-based monolingual Shona lexicography. Most of the considerations in this article could, however, be true of the modus operandi in corpus-based lexicographic projects of other languages of the world.

References

Bonvillain, N. 1993[2]. *Language, Culture and Communication.* Englewood Cliffs, New Jersey: Prentice-Hall.

Chabata, E. 2000. The Shona Corpus and the Problem of Tagging. *Lexikos* 10: 75-85

Chimhundu, H. 2000. *The Agenda for ALRI*. Harare: University of Zimbabwe.

Chimhundu, H. (Ed.). 1996. *Duramazwi ReChiShona*. Harare: College Press.

Chimhundu, H. (Ed.). 2001. *Duramazwi Guru ReChiShona.* Harare: College Press.

Flexner, S. and H. Wentworth. 1975. *Dictionary of American Slang*. New York: Crowell.

Jackson, H. 1988. *Words and their Meaning*. London/New York: Longman.

Mawema, M.B. 2000. Challenges Encountered in the Compilation of an Advanced Shona Dictionary. *Lexikos* 10: 209-224.

McEnery, T. and A. Wilson (Eds.). 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.