Divergent Approaches to Corpus Processing: The Need for Standardisation*

Esau Mangoya, *African Languages Research Institute, University of Zimbabwe, Harare, Zimbabwe (emangoya@arts.uz.ac.zw)*

Abstract: This article discusses some problems encountered in the processing of the Shona corpus. Most of the problems deal with the handling of adoptives, punctuation and individuals' idiolects. It also discusses the problem ensuing from an attempt to standardise the formats used in the handling of the corpus. The way a corpus is processed is critical in determining its quality. This article aims to show how the different linguistic backgrounds of the processors affect the appreciation of some vital aspects of the corpus. One of the acclaimed advantages of a corpus is that it allows research to be done on natural language. An ideal corpus should be a body of texts combined in a principled way to become a reliable language bank from which researchers retrieve data for various research purposes. With a good corpus, data can be provided giving an authoritative body of linguistic evidence which can support generalisations and against which hypotheses can be tested. As this proves the invaluable status of a corpus, the article assesses the processing of the Shona corpus and discusses how some aspects of the processing may impact negatively on its quality.

Keywords: CORPUS, STANDARDISATION, DATA, HEADWORD, LEXICOGRAPHY, RETRIEVE, TRANSCRIBE, TAGGING, ENCODING, PARSING, DIVERGENT

Opsomming: Uiteenlopende benaderings tot korpusverwerking: Die behoefte aan standaardisasie. Hierdie artikel bespreek 'n aantal probleme wat teëgekom is by die verwerking van die Sjonakorpus. Die meeste van die probleme handel oor die hantering van leenwoorde, punktuasie en die idiolekte van individue. Dit bespreek ook die probleem wat voortvloei uit 'n poging om die formate gebruik vir die hantering van 'n korpus te standaardiseer. Die manier waarop 'n korpus verwerk word, is krities vir die bepaling van sy gehalte. Hierdie artikel wil toon hoe die verskillende taalkundige agtergronde van die verwerkers die beoordeling van sommige van die wesenlike aspekte van die korpus beïnvloed. Een van die geloofde voordele van 'n korpus is dat dit toelaat dat navorsing oor die natuurlike taal gedoen word. 'n Ideale korpus behoort 'n geheel van tekste te wees wat op 'n geordende manier saamgestel is om 'n betroubare taalbank te wees waaruit navorsers inligting vir verskillende navorsingsdoeleindes kan verkry. Met 'n goeie korpus kan gegewens verskaf word om 'n betroubare geheel van taalkundige bewyse te gee wat veralgemenings kan bevestig en waarteen hipotesisse getoets kan word. Deurdat dit die

[*] This article is a revised version of a paper presented at the Seventh International Conference of the African Association for Lexicography, organised by the Dictionary Unit for South African English, Rhodes University, Grahamstown, Republic of South Africa, 8–10 July 2002.

waardevolle status van 'n korpus bewys, beoordeel die artikel die verwerking van die Sjonakorpus en bespreek hoe sommige aspekte van die verwerking die gehalte negatief mag beïnvloed.

Sleutelwoorde: KORPUS, STANDAARDISERING, DATA, TREFWOORD, LEKSIKOGRAFIE, VERKRY, TRANSKRIBEER, ETIKETTERING, ENKODERING, WOORDBENOEMING, AFWYKEND

Introduction

According to McEnery and Wilson (1996: 29), the Latin term *corpus* means 'body'. In linguistic studies, the meaning has been extended to indicate any 'body of texts'. It may basically be referred to as a collection of texts combined as a databank which can be used as a tool for any linguistic related research. Because the possible utterances in a language variety are infinite, the corpus can be considered as a sample. This sample is meant to be maximally representative of that language variety, presenting a clear picture of possible tendencies and proportions of linguistic elements (McEnery and Wilson 1996: 30). As a result, different varieties of texts and genres are to be included in the corpus, for instance novels, poetry, drama, newspapers and recorded interviews. For serious and comprehensive long-term use, the corpus has to be continuously developed with more and new texts being added. Through the continuous addition of new data, the changing store of texts can always reflect previous and current linguistic behaviour (Aarts and Meijs 1984: 4). The overall purpose of a corpus is to have raw data for use in various kinds of research. This is one of the newer approaches which has helped in the differentiation of linguistic studies. Hence, today, in linguistic study disciplines, corpus-based syntax and semantics are contrasted with non-corpus-based syntax and semantics. Different corpora may be compiled for various purposes and interests. Over the years, the processing of the corpus has evolved in such a way that today the mentioning of the term *corpus* automatically implies machine-readable data. In the past, the corpus could be in printed form. A corpus remains a basic data reference collection on which various forms of research can be carried out. McEnery and Wilson (1996: 32) stress the importance of a corpus: 'As a standard corpus also means that a continuous base of data is being used and thus variation between studies may be less likely to be attributed to differences to the data being used and more to the adequacy of the assumptions and methodologies in the study.'

In corpus-based lexicography, the corpus is used for headword selection, defining and providing examples. The use of a corpus in dictionary compilation is of particular interest to this article. Six corpora are currently being compiled by the African Languages Research Institute (ALRI) at the University of Zimbabwe, Harare, with the aim of using them in the compilation of monolingual and bilingual dictionaries. The corpora being processed are Shona, Ndebele, Kalanga, Nambya, Tonga and Shangaan.

While the Shona and Ndebele corpora are at an advanced stage of processing, the other four are in their initial stages. The Shona corpus stands at close to five million running words, and the Ndebele corpus at around three million running words. In this article, the focus is on the processing of the Shona corpus. The texts being included in the corpus consist of written material and oral interviews. The written material constitute about a quarter of the total corpus. This article aims to discuss some of the problems encountered during the processing of oral material.

To be a research tool, the corpus has to be qualitative. When there are inconsistencies in the handling of the corpus, some of the poorly processed materials will find their way into the language bank. A false impression on the size and quality will then be created. The poorly processed corpus may have retrieval limitations. Therefore, whatever is to constitute the language bank has to be processed adequately if it is to remain useful, accessible and relevant to research.

Except that the processing of a corpus takes a considerable time, in most cases it usually involves a chain of different individuals for the different stages, as shown here:

text creation → transcribing → encoding → tagging → parsing

Text creation focuses on the interviewee in the case of oral material or the author in the case of written material. However, the focus of this article is the processing of oral material. Transcribing involves the scripting of information from a cassette or recorder onto paper. Encoding is the keying in of the text from the paper into the computer for electronic storage. Tagging is the marking of the documents for purposes of retrieval from the corpus. Parsing is the application of the proof-reading programme to check for consistency of the tagging. All these stages may involve different individuals who have a different conceptualisation of certain aspects of the words in a language.

At times, this led to different approaches by individuals working on the corpus. The differences emanated from the nature of the treated oral texts. Some of the problematic material came from language contact areas.

Influence of language contact

One of the problems that lead to inconsistencies in the processing of the Shona corpus is caused by borrowed words, particularly from the Nguni languages and English with which Shona is in contact. The problematic words from Nguni mainly come from three fronts. There are words from Ndebele, a language spoken in the south-western parts of Zimbabwe. Ndebele is one of the Nguni group of languages also including Xhosa, Zulu and Swati which are spoken in the Republic of South Africa. The Shona and Ndebele groups came in contact during the second quarter of the 19th century in the time which has come to be known as the Mfecane (Sibanda 1989: 25). This was a period of political instability among the Nguni which caused Mzilikazi and his followers to cross the Limpopo into the present-day Zimbabwe. A language contact zone was created in the areas stretching from Kwekwe down to Chirumhanzu south of Chivi to Mberengwa and across the Lowveld to Matibi.

These language contact areas have played a big role in the borrowing of words from Ndebele into Shona. Also at a later stage, around the 1890s during the time of the British colonisation, the African groups accompanying the British as guides were mainly from the Nguni group. Since these people spoke different languages, they communicated mainly through a pidgin which has come to be known as Fanagalo. This Fanagalo was composed mainly of words from Afrikaans, English and the Nguni languages. So, as the pioneer column moved up into Mashonaland, they created a second form of contact, thereby reinforcing the initial language contact as more words came into Shona through this new encounter. As a result many words from Nguni were introduced and accepted into Shona. Examples of such words are given under (1).

(1) *-funda* (learn)
mufundisi (reverend/teacher)
-zama (try/attempt)
-bopa (inspan)

These words, originally from Nguni, have been introduced and accepted into the Shona corpus as borrowed words. Their adoption is no longer questionable since their introduction and reinforcement through the second form of contact also coincided with the introduction of institutions in which they have been extensively used. The words *-funda* and *mufundisi* are widely used in educational and religious circles, and *-zama* and *-bopa* are quite prevalent in agricultural and industrial sectors.

Adoption is a vital linguistic phenomenon which cannot be ignored, particularly when the corpus has to be used for the compilation of dictionaries. The reason why it was not crucial to consider the etymology of these words at this stage was that, until then, the focus had been on the compilation of general dictionaries, *Duramazwi reChiShona* and *Duramazwi Guru reChiShona*, published in 1996 and 2001 respectively, the first being a general medium-sized dictionary and the second an advanced dictionary.

However, there is a second group of words about whose status there were divergent views. These are words of common language usage, having a high frequency in Shona. They are words that feature much in social conversations. Some team members believed that they had to be considered foreign while others felt the opposite. The reason is that some of these words were more acceptable as already adopted compared to others. Examples of such words about which there were divergent views are given under (2).

(2) *mnandi* (delicious/sweet)
mgane (friend)

-sakaza (speak)
-saba (be afraid of)

Part of the reason why team members felt that these could be acceptable Shona adoptives is that Shona orthography can easily make provision for them, representing them as *munandi*, *mugane*, *-sakaza* and *-saba* respectively. A lack of clear guidelines leaves the processors making subjective decisions about the categorisation of words in the corpus. Resultant discussions led to the conclusion that *mnandi* and *-sakaza*, when their frequent featuring in various literary genres is considered, could be regarded as adoptives as long as they were made to conform to Shona orthography.

There is a high level of subjectivity about which words should be considered adoptives. This subjectivity is also influenced by the orthographic closeness between the source language of the word and the adopting language. Therefore, some Nguni words found in the oral texts were unanimously considered foreign, basically owing to the difference in orthography between the two languages. Despite the long years of language contact, some of the words have not been accepted into Shona. The team was agreed that these were to be clearly indicated and rightly tagged as foreign in the corpus as shown by the examples under (3).

(3) <foreign> *khombisa* </foreign> (seek for love)
<foreign> *thanda* </foreign> (love)
<foreign> *nkosi* </foreign> (king)

The guideline formats for corpus processing assumes that it is clear what constitutes a foreign and what an adopted word. It does not consider that there are different levels of acceptance by different individuals. Guideline formats should be designed in such a way that at different levels of processing, it becomes clear how to handle different words from other languages.

As pointed out, the main problem that led to contradictions was differences in spelling where cluster combinations in Ndebele such as *kh*, *th* and *nk* are unacceptable in Shona. There were situations where the words under (3) would be adapted to Shona orthography. These would appear as *kombisa*, *tanda* and *ngosi* respectively. Despite being written according to Shona orthography, however, they were tagged as foreign, seemingly under the influence of the way they are spelt in the source language.

There was no immediate solution to the problem of handling words from other languages. It was suggested that all the words individuals came across could be submitted to the team panel to consider their status. The question of subjectivity still remained central though. It depends on the individual to choose words that need to be discussed by the panel. It also requires knowledge to recognise when the status of a word is not clear, selecting it for consideration by the team. Because corpus building has to be a continuous process, panel meetings also have to become routine.

On the other hand, the Mfecane created another form of language contact in the eastern part of Zimbabwe. A further group, the Shangaan, which was linguistically closely related to the Nguni group, also crossed the Limpopo just about that time. The Shangaans did not settle as a bigger group like the Ndebeles. After having crossed the Limpopo, small groups remained behind while others proceeded into Zambia and finally into Malawi. Linguistically, an interesting scenario was created in the south-east of Zimbabwe stretching into Mozambique where Ndau, one of the dialects of Shona, is spoken. The migrating groups actually fused with the Shona groups they found in those areas. Over the years, because of this language contact situation, the Shona spoken in these areas adopted some lexical items. This resulted in the creation of a peak dialect, Ndau, spoken in and around the Chipinge district. There now exist linguistic elements that are problematic when they appear in the corpus. Ndau has many salient linguistic elements when compared to the rest of the Shona dialects.

Inconsistencies occurred in the handling of texts from the rest of the Shona-speaking areas and those from the Ndau dialect. In the texts from the rest of Mashonaland, some of the words from the Nguni group were clearly marked foreign while those very words are actually accepted as part of the vocabulary of Ndau. Examples from the Ndau vocabulary that would automatically be marked as foreign if they appear in texts from other dialect regions are given under (4).

(4) *-tshaya* (beat)
nqondo (brain)
-gqoka (put on)
-qonda (go straight)

This historical background information on the language situation is not provided to the user of the corpus. When these words feature in any of the Shona dialects except Ndau they are treated as foreign. At the same time, these have become natural Ndau words which are recognised as such. Once it is realised that it is a Ndau text, they are not treated and marked as foreign. However, the overall analysis of the whole corpus may give the impression that it has been poorly processed.

Part of the problem is that the Ndau dialect was not greatly taken into consideration during the standardisation of the Shona dialects. Doke (1931), who played a pivotal role in the standardisation of the Shona dialects, recommended that the words from Ndau should be used sparingly. There are new challenges now as the corpus has to reflect the living language of the people. If a representative corpus is to be produced, Ndau should be considered as any other Shona dialect. What is evident here is that the problems arise as a result of neglecting Ndau during standardisation. In some instances, the distinctive spelling of the Ndau words in the corpus are not scripted but are deliberately removed and substituted with those allowed in general Shona orthography.

This makes it very difficult to recognise them in the new form. Examples of a group which involves the adapted words are shown under (5).

(5)	**Original Ndau**	**Adapted form**
	nqondo	*n'ondo* (brain)
	-gqoka	*-goga* (put on)
	-qonda	*-konda* (go straight)

The other adapted group consists of words which involve consonantal combination *hl*. The standard orthography does not allow for the spelling Ndau speakers would prefer. Some examples of these are given under (6).

(6)	**Original Ndau**	**Adapted form**
	-hlaba	*-shava* (pierce through)
	-hlupa	*-shupa* (trouble someone)
	muhlobo	*mutyovo/mutyowo* (way/method/type)

Once in the original form, they are marked as foreign, but are unmarked when adapted. In the case of *-hlaba* which has only one meaning, the adopted form becomes *-shava* which has three meanings in Shona. This again causes complications and more inconsistencies. All the stages of corpus building must represent what the creator of text really meant and intended. As a result, the tendency is to consider anything not conforming to the standard as foreign. If not marked foreign, there is forced adaptation.

When these words were brought to the team panel for consideration, it was agreed that as long as the text was Ndau they would not be marked as foreign. If they appeared in texts of other dialects they had to be given a foreign tag. However, this is a temporary solution implemented for the processing of the corpus leading to evident inconsistencies. It only suits the corpus processors but the solution does not address the major issue which causes these discrepancies. What is evident here is that the problems concern language planning and language policy. Corpus building is a grant project that should involve all language stakeholders including the government who has to act on matters of language planning and language policy. As for the corpus that has been produced, the problem of inconsistencies should be explained. The language situation should be outlined in order to inform the user about the existence of the inconsistencies and the reasons for it.

Another form of language contact also exists in Zimbabwe. Zimbabwe was a British colony in which English was declared the official language. English was raised above all the indigenous languages of the country. With the attainment of independence, Shona and Ndebele were also accorded official status alongside English. As a result, some English lexical items have found their way into the indigenous languages. In cases where these lexical items have been partially adapted into Shona, they were problematic in the processing.

Partially adapted adoptives from English

English, which enjoyed the monopoly of being the only official language for a long time before independence had much influence on the local languages. As a result there are many adoptives from English in Shona. The majority of people in Zimbabwe are bilingual, speaking their local indigenous languages and having English as a second language. It is this knowledge of the two languages which results in speakers adapting one part of English words and leaving the other part unaltered.

The handling of these partially adapted adoptives from English was problematic, because of this partial borrowing of lexical elements. Once words have been adopted, they become acceptable lexical items of the borrowing language. However, these are instances where some elements of the words have remained partly in their original form. Examples of these are given under (7), the unaltered elements being shown in bold italics.

(7) *kusi**lai**dha* (to slide)
*hazvisi ra**ight**ka* (it is not right, is that not so?)
*kusasi**pect**a* (to suspect)

Because the majority of the sounds in the words are also found in Shona, some processors of the corpus felt the words could be considered as already accepted into the language. As a result they were left unmarked in some of the Shona corpus texts. Although the majority of the syllables have been remorphologised and rephonologised, some elements remain unchanged. So some corpus processors would mark only the unchanged part of the word as foreign as shown in the examples under (8).

(8) *kusi*<foreign>***lai***</foreign>*dha* (to slide)
hazvisi ra<foreign>***ight***</foreign>*ka* (it is not right, is that not so?)
kusasi<foreign>***pect***</foreign>*a* (to suspect)

What was marked as foreign are just sections of words that have maintained their identity in spelling from the source language.

While the other parts of the words conform to the writing system of Shona, those marked as foreign are not full words, neither are they morphemes which have a clear meaning for the user of the corpus. The processors of the corpus have tried to indicate that there are notable foreign elements in the word structures but what has been marked is not useful for meaningful linguistic research.

In the absence of a clear policy, the way texts may be handled by different processors will vary, leading to inconsistencies. As has been demonstrated above, different aspects of the language may be perceived differently and uniformity in the processing of the corpus may be difficult to achieve. The guide formats just mention that foreign words should be marked, but as the above examples under (8) demonstrate, the criteria for identifying foreign words are not laid down. They do not address these new challenges.

Consequently there are inconsistencies in the processing of various words particularly loanwords. The other alternative which was suggested in the discussion of the problem was to mark the whole structure as foreign, as shown in the examples under (9).

(9) <foreign>*kusilaidha*</foreign> (to slide)
hazvisi <foreign>*raightka*</foreign> (it is not right, is that not so?)
<foreign>*kusasipecta*</foreign> (to suspect)

This would override the fact that certain parts of these lexical items have already been adopted in and adapted to Shona.

However, when looked at more closely, the problem really stems from the fact that those handling the corpus are bilingual. The way a word may be presented in scripted form from the orally produced text is subjective. This is the reason why these words have various presentations in the corpus files. There has to be full communication with all involved in the processing of the corpus. In areas where there are divergent approaches, discussion is necessary and the consensus reached should be recorded. Such records should serve as a guide to the processors of the corpus. This is important for consistency. These records should be used as manuals guiding users of the corpus in the way some words were handled during the processing.

ALRI's standard formats for the processing of the Shona corpus have assumed that foreign words would be easily identified. No consideration has been given that some words would be changed in the process of adoption. As no prescriptions are given for the handling of partially adapted words, they are always problematic in the processing of the corpus. Without standard formats, individual preferences take precedence. This gives rise to the issue of the individual idiolect also coming into play. According to Crystal (1991: 170), an idiolect refers to the linguistic system of an individual speaker — one's personal dialect.

Varying idiolects

Individuals vary in their idiolects. This applies to both the text creator who in the case of oral texts is the interviewee and the processor of the text. This becomes evident in the punctuation of different texts. Certain aspects were handled differently because of a different conceptualisation of the punctuation of texts. One of the problematic areas is the 'probe statements' of the interviewee. Examples (10)(a), (b), (c) and (d) demonstrate the differences in the handling of the probe statements resulting from varying idiolects.

(10) (a) *Zvinenge zvichida kuti kana wasvika wodini … Wodzikama. Zvinhu zvobva zvodini … Zvofamba nenzira yazvinofanira kufamba nayo. Zvozodini … Zvopera.*

It is necessary that when you arrive you what … Remain cool. Then everything will what … Go the way it should move. Finally it will what … it comes to an end.

(b) *Zvinenge zvichida kuti kana wasvika wodini? Wodzikama. Zvinhu zvobva zvodini? Zvofamba nenzira yazvinofanira kufamba nayo. Zvozodini? Zvopera.*
It is necessary that when you arrive you what? Remain cool. Then everything will what? Go the way it should move. Finally it will what? It comes to an end.

(c) *Zvinenge zvichida kuti kana wasvika wodini, wodzikama. Zvinhu zvobva zvodini, zvofamba nenzira yazvinofanira kufamba nayo. Zvozodini, zvopera.*
It is necessary that when you arrive you what, remain cool. Then everything will what, go the way it should move. Finally it will what, come to an end.

(d) *Zvinenge zvichida kuti kana wasvika wodini! Wodzikama. Zvinhu zvobva zvodini! Zvofamba nenzira yazvinofanira kufamba nayo. Zvozodini! Zvopera.*
It is necessary that when you arrive you what! Remain cool. Then everything will what! Go the way it should move. Finally it will what! It comes to an end.

Discussions by the ALRI team on how to handle such statements revealed divergent preferences. Different processors had different ways of punctuating the statement which resulted in four versions of the same statement. In example (10)(a), the processor marked the end of the seemingly unfinished statements with ellipses. In (10)(b), the processor preferred using question marks, feeling that the speaker was asking rhetoric questions he would immediately answer himself. Commas were preferred in example (10)(c), the reason given was that after the probe there was a pause before the statement was finally completed. The preferred punctuation in example (10)(d) was the exclamation mark, the reason advanced that the interviewee was penultimately stressing a point before the actual completion of the statements. After a common approach to such texts had been debated, the use of a comma was finally agreed upon.

Whatever decision made had implications on the whole set of words in the given sentences. For example: The use of a comma would mean the following word would start with a small letter. The use of exclamation or question marks would render the examples into complete statements, having implications for the final corpus text. It is in the corpus processing that a text may be given a particular value. The processor of the text may decide whether the speaker uttered an exclamation, made a full statement or asked a question. This shows the importance of decisions on how to handle certain aspects of idiolect and register. After this particular case had been discussed, it was agreed that using a comma would be preferable. However, this background information will not be passed on to the user of the corpus text who might have an own opinion on how the statements should have been punctuated, hence the need for written records on how particular aspects were handled.

The ALRI team agreed that the best way was option (10)(c). This consensus was only reached after there had been divergent treatments of these statements resulting in four options. The original guiding formats indicated where punctuation marks should be placed before the various tags are used in the text. It did not, however, deal with challenges arising from the actual punctuation of a document by the processors. The team's agreement has become the standard way of punctuating such statements. It is nevertheless important to record this decision for the users of the corpus so that they could have a fuller understanding of how the documents were treated. All the background information on the standardisation of the formats should be recorded. Standardisation should be continuous since new challenges continue to become manifest. One of these problems was how to handle factual distortions.

Factual distortion

Few cases were found where texts were factually erroneous. These could result from either a lack of actual knowledge about a subject or an unintended mistake by the creator of the text, or a misrepresentation caused by an oversight in the line of corpus production. However, the resultant corpus product emerges with erroneous information. Some of the examples which had to be discussed by the team working on the corpus are given under (11).

(11) *Jona akamedza hove mugungwa.* (Jonah swallowed fish in the sea.)
Pamadhigirii etriangle ari 25 (of the triangle's total of 25 degrees)

There may be a quick conclusion that this is obviously wrong information that has to be corrected. It may be easy to return to the original text which in these cases will be the cassettes on which the interviews were recorded to verify whether the text really represents what was said by the interviewee. The principle in corpus processing is to be faithful to the original text of the creator. It should be kept in mind that the misinformation might not have been a deliberate but an unintended mistake. When people do research they need to find factual material from the sources they use. The team agreed that these statements had to be left unaltered. It was felt there could one day be somebody with an interest to study these slips in language. The decision, though taken with good intentions, does not fully benefit the user. A special tag could have been developed to alert the user. Why this was not very problematic to the ALRI team is that they use the corpus in dictionary making. There would not be problems in selecting headwords from such statements. For broader research such statements would need to be marked to indicate that they contain some factual errors.

Conclusion

Corpus building is a laborious task. This stems from the fact that the corpus has to be fully representative of the spoken language: of what was said and what was meant to be said. This presents challenges and new problems appear, so that the way of treating these has to be standardised. It is necessary to discuss these problems so as to develop a common approach. As the corpus continues to be processed, new tag marks have to be formulated. All these aspects need to be standardised for the researcher who fully utilises the corpus resource materials. The outstanding problems facing individuals in the corpus production line have to be discussed, agreed upon and standardised. As a result, there is the need for a manual serving as guide to both the processor and the user of the corpus. The ALRI team should consider all the aspects where there were divergent approaches. All the agreed solutions of problems should be combined and presented as front matter or indexed as corpus guide. This also gives the opportunity for users of the corpus to form their own opinion about certain aspects of the corpus.

The user of the corpus may have a better understanding of the language situation and the historical background of this situation. This will help the user to appreciate how various aspects of the texts were handled to make the corpus useful as a research tool. Even the contradictions by the processors of the corpus may be appreciated. In this way, there is room for input from the creators, the processors and the users that can help improve the corpus. Like any research resource the corpus should be analysed and criticised to create the possibility for its improvement.

Bibliography

Aarts, J and W. Meijs (Eds.). 1984. *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*. Amsterdam: Rodopi.

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

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

Crystal, D. 1991. *A Dictionary of Linguistics and Phonetics*. Third Edition. Oxford: Blackwell.

Doke, C.M. 1931. *Report on the Unification of the Shona Dialects Carried Out under the Auspices of the Government of Southern Rhodesia and the Carnegie Corporation*. Hertford: Stephen Austin.

McEnery, T. and A. Wilson. 2001. *Corpus Linguistics: An Introduction*. Second Edition. Edinburgh: Edinburgh University Press.

Sibanda, M.J. 1989. Early Foundations of African Nationalism. Banana, C.S. (Ed.). 1989. *Turmoil and Tenacity: Zimbabwe 1890–1990*: 25-49. Harare: The College Press.