The Ndebele Language Corpus: A Review of Some Factors Influencing the Content of the Corpus[*]

Samukele Hadebe, *Institutt for Nordistikk og Litteraturvitenskap, Seksjon for Leksikografi, University of Oslo, Oslo, Norway (samukeleh@yahoo.co.uk)*

Abstract: The Ndebele language corpus described here is that compiled by the ALLEX Project (now ALRI) at the University of Zimbabwe. It is intended to reflect as much as possible the Ndebele language as spoken in Zimbabwe. The Ndebele language corpus was built in order to provide much-needed material for the study of the Ndebele language with a special focus on dictionary-making and research. Like most corpora, the Ndebele language corpus may in future be used for other purposes not thought of at the time of its inception. It has been designed to meet generally acceptable standards so that it can be adaptable to various possible uses by various researchers. The article wants to outline the building process of the Ndebele language corpus with special emphasis on the challenges that faced compilers, and possible solutions. It is assumed that some of these challenges might not be peculiar to Ndebele alone but could also affect related African languages in a more or less similar situation. The main focus of the discussion will be the composition of the Ndebele language corpus, i.e. the type of texts that constitute the corpus. The corpus is composed of published texts, unpublished texts and oral material gathered from Ndebele-speaking districts of Zimbabwe. It will be argued that the use of the corpus and its reliability for research depends among other factors on its contents. It will also be shown that the contents of a corpus depend on a number of factors, some of which include sociolinguistic, political and economic considerations. These considerations have implications on both the content and quality of published and oral texts that constitute the Ndebele language corpus.

Keywords: CORPUS, ORAL MATERIALS, CODE-MIXING, CODE-SWITCHING, MOTHER-TONGUE, NDEBELE

Opsomming: Die Ndebeletaalkorpus: 'n Oorsig van sommige faktore wat die inhoud van die korpus beïnvloed. Die Ndebeletaalkorpus wat hier beskryf word, is dié saamgestel deur die ALLEX Project (tans ALRI) by die Universiteit van Zimbabwe. Dit is bedoel om soveel moontlik te weerspieël van die Ndebeletaal soos in Zimbabwe gepraat. Die Ndebeletaalkorpus is opgebou om veelbenodigde materiaal te verskaf vir die studie van die Ndebeletaal, met spesiale fokus op woordeboeksamestelling en navorsing. Soos die meeste korpora, kan die Ndebeletaalkorpus in die toekoms gebruik word vir ander doeleindes waaraan nie by tye van sy ontstaan gedink is nie. Dit is ontwerp om aan algemeen aanvaarde standaarde te voldoen sodat

[*] This article is a slightly revised version of a paper presented at the Sixth International Conference of the African Association for Lexicography, organised by the Sepedi National Lexicography Unit, University of the North, Pietersburg, 2–4 July 2001.

dit aanpasbaar kan wees vir verskillende moontlike gebruike deur verskillende navorsers. Die artikel wil die bouproses van die Ndebeletaalkorpus skets met spesiale klem op die uitdagings wat die samestellers ondervind het, en moontlike oplossings. Dit word aanvaar dat sommige van hierdie uitdagings nie eie aan Ndebele alleen mag wees nie, maar ook verwante Afrikatale in 'n min of meer soortgelyke situasie mag raak. Die hooffokus van die bespreking sal op die samestelling van die Ndebeletaalkorpus wees, d.w.s. die soort tekste wat die korpus uitmaak. Die korpus is saamgestel uit gepubliseerde tekste, ongepubliseerde tekste en mondelinge materiaal versamel in Ndebelesprekende distrikte van Zimbabwe. Daar sal geredeneer word dat die gebruik van die korpus en sy betroubaarheid vir navorsing op onder andere sy inhoud berus. Daar sal ook getoon word dat die inhoud van die korpus op 'n aantal faktore berus, sommige waarvan sosiolinguistiese, politieke en ekonomiese oorwegings insluit. Hierdie oorwegings het implikasies vir beide die inhoud en gehalte van gepubliseerde en mondelinge tekste wat die Ndebeletaalkorpus uitmaak.

Sleutelwoorde: KORPUS, MONDELINGE MATERIAAL, KODEVERMENGING, KODEOMSKAKELING, MOEDERTAAL, NDEBELE

Introduction

The Ndebele language corpus described here is that compiled by the ALLEX Project (now ALRI) at the University of Zimbabwe. It reflects or is intended to reflect as much as possible the Ndebele language as spoken in Zimbabwe. The Ndebele language corpus was built in order to provide much-needed material for the study of the Ndebele language with a special focus on dictionary-making and research. As would later be demonstrated in this article, the composition of texts and their conversion to machine-readable documents reflect the underlying focus of the main objective, which is lexicography. Like most corpora, the Ndebele language corpus may in future be used for other purposes not thought of at the time of its initial compilation. The main focus of this article is the content of the Ndebele language corpus, i.e. the type of material that constitutes the corpus. The corpus is composed of published texts, unpublished texts and oral material gathered from Ndebele-speaking districts of Zimbabwe. It will be argued that the use of the corpus and its reliability for research depend among other things on its contents. The contents of a corpus depend on a number of factors, some of which include sociolinguistic, political and economic considerations. These considerations have implications on both the content and quality of published and oral texts that constitute the Ndebele language corpus.

Background: The Ndebele Language

Language policy factors have a bearing on the content of the corpus of the Ndebele language. This is so because of the status of Ndebele in Zimbabwe. Ndebele, together with Shona, are the recognised national languages of Zimbabwe while English enjoys the almost exclusive monopoly as language of administration and medium of instruction in schools. Ndebele is therefore confined to informal domains while official business is done mainly in English. For this reason and other related factors, Ndebele has not developed a vocabulary for other spheres of activity. For instance, there are no Ndebele books for subjects such as history, geography, science or mathematics. In short, the Ndebele language lacks published material in and about the language. Although Ndebele is taught up to university level in Zimbabwe, this has not led to as many advantages as would be expected, that is, in terms of research on and publications in the language. One factor that has hindered this otherwise normal development is that instead of teaching Ndebele as spoken in Zimbabwe, Zulu was taught. For this reason one big question that will always evade compilers is whether to include Zulu texts in the Ndebele language corpus. Secondly, the continued use of English as language of instruction and official language of administration has denied the Ndebele language the opportunity to develop vocabulary and terminology in fields such as agriculture, commerce, law, science, etc. With this background it is only natural that creative works would dominate the corpus.

Gathering of Oral Material

Areas where Oral Material Was Collected

Ideally oral material had to be gathered in all the areas where there are mother-tongue speakers of Ndebele. This would have given the desired representative sample of spoken Ndebele from all geographical areas. However, not all areas were as well covered as researchers would have wanted. Firstly, such an endeavour was impracticable financially, considering the cost involved in such an undertaking. Secondly, mother-tongue speakers of Ndebele are not confined to Ndebele-speaking districts and towns only, but some pockets are scattered in other non-Ndebele-speaking districts. Locating all these communities would not only have been time-consuming and costly but was also felt to be unnecessary. The areas of focus were therefore the Ndebele-speaking districts, which are mainly in the provinces of Matabeleland North, Matabeleland South, and the Midlands. These three provinces constitute nearly half the size of the country geographically although accounting for probably one fifth of the country's population. This implies that human settlements are far apart and very scattered, resulting in high cost in travelling through the districts. This also became a factor in reducing potential areas for oral material collection.

The research was also to serve as a sociolinguistic survey of the language map of the country. Until that of Hachipola (1998), no prior comprehensive survey of the language situation was available. The districts which are commonly described as Ndebele-speaking areas are also populated by speakers of the so-called minority languages: Kalanga, Venda, Tonga, Nambya, Sotho, and in the Midlands districts consist of both Ndebele and Shona speakers. There were debates as to whether it was worth collecting data from areas where other languages are also spoken. There were concerns that the type of Ndebele spoken by these people who also speak other languages, was likely to be heavily influenced by these languages and therefore not appropriate for the envisaged dictionary. This argument posed another problem of how to distinguish between acceptable Ndebele and unacceptable varieties. There were fears that it would be politically wrong to exclude other people deliberately because they were speakers of other languages, as all children in these districts learn Ndebele in any case. So there were arguments that all varieties of Ndebele should be gathered as this would reflect the linguistic reality at ground level. Although it was eventually agreed that oral material should only be collected from mother-tongue speakers of Ndebele, this was impossible in practice. However, to minimise the influence of other language groups, the majority of student research assistants were deployed only in those areas where only Ndebele was the community language. For instance, the Beitbridge district was not covered because of its predominantly Venda population, and only one research assistant was deployed in Binga, which is a Tonga territory. The table below shows the rough estimate of Ndebele speakers in Beitbridge as extracted from Hachipola (1998: 32):

Areas of Language Mixture in the Beitbridge District

Area	Dominant Community	Other Communities
1. Tshipise	Venda	Shangani
2. Tshitulipasi	Venda	Shangani
3. Tshikwalakwala	Venda	Shangani
4. Dendele	Venda	Sotho
5. Maramane	Venda	Sotho
6. Shashe	Sotho	Venda
7. Malibeng	Venda	Sotho
8. Makombe	Venda	Pfumbi
9. Siyoka 2	Venda	Ndebele

As becomes clear from this table, it would have been a costly venture to collect oral material in Beitbridge owing to the paucity of mother-tongue speakers of Ndebele. Ndebele is the language taught in schools and used in the public domain in such areas as Beitbridge but the users are not first-language speakers of Ndebele and their type of Ndebele was considered not of the desired standard for a monolingual dictionary for learners.

The other determining factor in the choice of areas to be covered was the availability of student research assistants in certain areas. According to the regulations of the University of Zimbabwe on remuneration for student research assistants, there is no allowance for transport and accommodation. It therefore meant that student research assistants should come from those areas where research was to be conducted. As a result some areas could not be covered because there were no available students from these areas. Students researching in their home areas had the advantage that it was easier for them to conduct interviews among communities with which they are familiar.

Competence of Interviewers

The student research assistants were largely drawn from undergraduates who had done the course on translation and lexicography. However, some had just taken Ndebele language courses in their first year at college. None had prior experience of research in this field, but their performance was considered satisfactory by the corpus compilers and most of them fulfilled the targets that had been set. They had undergone a crash course on the basics of fieldwork that included training in the use of audio-recorders and transcribing recorded material.

All the student research assistants were fluent mother-tongue speakers of Ndebele. Of the twenty-six research assistants, eleven were female and the rest male. Their ages ranged between twenty and twenty-four years. In terms of their academic ability as well as their proficiency in Ndebele the group was competent enough to assume the task. These are some of the key issues that have a bearing on the quality and reliability of the results of the research. The fact that the student research assistants were working in their home districts had an added advantage in that they knew most of the people as well as their potentialities to provide certain information. Similarly it was easier to approach potential informants by people who already knew them. However, because these interviewers were almost all of the same age group, there are topics they seemingly handled very well but in some cases their youth was a limitation when considering the cultural orientation of the Ndebele society. Topics related to everyday events, which dominated the interviews, were handled well but specialised topics such as aspects of religion or sexuality were not satisfactorily treated.

The reason for this is that in Ndebele society it is considered improper to discuss certain topics with young people. In the same way sexuality cannot be discussed by opposite sexes, which meant that male interviewers could not ask certain questions to female informants and vice versa. The other limitation was the student research assistants' own lack of knowledge about certain topics so that they could not pose suitable questions to elicit more information from the informants. The researchers and compilers of the Ndebele language corpus had foreseen some of these inadequacies in the student research assistants. Therefore, the students had been given notebooks that were to be used as diaries throughout the whole fieldwork period. In these they had to give detailed descriptions of their daily contacts and work within the community as well as their own evaluation of informants they met. It is here, also, that they had to note down potential informants they could not interview or those they felt had more information but could not disclose it to them because of their age or sex.

Another way in which these limitations were noted was when the researchers went through the transcribed texts as well as through the tapes. In some instances it became clear that more could have been obtained had the interviewer been knowledgeable enough to lead the discussion fruitfully.

All these limitations had been foreseen and ways of overcoming them prepared. It was originally planned that the researchers would make follow-up interviews in those areas where it was felt follow-ups were needed. An inventory of potential informants was compiled but unfortunately so far no follow-up interviews could be conducted to fill the gaps left by the student research assistants. As there already was a lot of oral material to be processed, no immediate follow-ups were considered. There was also no money available to pursue further research. However, the potential informants are known and recorded for possible future interviews.

Method of Collecting Oral Material

Most of the oral material was collected by means of structured and unstructured interviews. Each of the student research assistants was responsible for determining whether to use a structured or an unstructured interview. Some began as structured but flowed into more or less unstructured discussions. Guidelines had been given for typical structured interviews on specific topics on which they were required to gather material, for instance, topics related to Ndebele marriage customs, child care or cattle farming. Students were given the discretion to choose between structured and unstructured interviews depending on what they thought best in prevailing circumstances. Apart from aiming at creating a word-bank for the Ndebele language, it was also envisaged that the material would be useful for oral history and cultural studies as well as for various language studies other than lexicography. Although the primary aim was an oral corpus for dictionary-making, its other possible uses were not forgotten. The student research assistants themselves had no prior knowledge of a corpus or dictionary-making based on a corpus, which could lead them to the assumption that detailed oral material is required to obtain the meanings of words. However, although their assumptions were not always entirely correct, they succeeded in collecting a rich variety of oral material.

While interviews were the most prevalent in the oral material collection, there were a few cases where recordings of dialogues or other discussions were made. For instance, there were recordings of songs, either at social functions such as weddings, or in churches and in schools. Church services were also recorded as well as classroom sessions in both primary and secondary schools. All these were done only after prior permission was sought from the authorities concerned. In some classroom recordings the teachers involved did the recordings by themselves so as to avoid the presence of a stranger, that is the student research assistant, in the class. One student research assistant managed to record a traditional court session while some recorded normal conversations in workplaces. While this type of recordings are valuable for yielding real-life situations, they have, however, some limitations, the most conspicuous being the problem of identifying the particular speaker in terms of name, age, occupation and gender. All oral interviews are marked with these details for record purposes.

As already mentioned, all student research assistants were given notebooks that they had to use like diaries to record in detail all research experiences. They would also write down the names of trees, grass, birds and plants found in their areas of research. These notebooks are therefore a rich collection of oral material, especially as far as the names of birds, animals, trees and the like are concerned. Family praise names were sometimes similarly obtained. Audio-recording informants could not easily have yielded this kind of valuable information. Although the bulk of the oral material the student research assistants collected was through audio-recordings, the notes they made in their notebooks have proved very useful.

Written Texts

Renouf (Sinclair 1987: 2) makes the following observation:

> When constructing a text corpus, one seeks to make a selection of data which is in some sense representative, providing an authoritative body of linguistic evidence which can support generalisations and against which hypotheses can be tested.

As it describes an ideal situation, this observation holds true for any language corpus. However, for languages with a relatively short and recent literary history such as Ndebele, it is not always practical to have a representative selection. As Renouf (Sinclair 1987: 2) states further, a selection is possible where there is a range or variety from which a representative sample can be drawn:

> The first step towards achieving this aim is to define the whole of which the corpus is to be a sample.

For Ndebele, with a very small number of published books whether it be fiction or non-fiction, the whole implies all publications in the language. The long-term objective is to include all published texts in the Ndebele language corpus. The little that has been published represents a neat selection of material used for educational purposes. Apart from religious texts, most publications in Ndebele, both fiction and non-fiction, are in fact targeted at schools. A number of factors account for this. One reason is that the cost of producing and publishing books in Zimbabwe is relatively high and in order to offset these costs there is also a need for a ready market for the books. In a country where there is not yet a reading culture, only schools offer that ready market, and publishers would publish only those works that could be used in schools. However, the Ndebele language corpus in its current state does not reflect the long-term ideal nor is it likely to do so in the near future. It is a sample of what has so far been published in Ndebele and this sample cannot be described as representative until qualified.

It should be explained why certain texts were excluded from the Ndebele language corpus. The early written works in the Ndebele language may be categorised as falling between 1852 and 1950. The first date marks the first publications in Ndebele by the London Missionary Society, while the latter date marks significant departures from the early Ndebele orthography. Publications spanning this period, few as they may be, are very important in the history of Ndebele but had to be excluded. These are in the old Ndebele orthography, which few people can read today and unless these are rewritten in the current orthography (which is very unlikely) they cannot be included in the corpus. Some of the symbols used would even pose problems for the scanner to detect. Therefore, all texts in the old orthography, which include scripture texts and Ndebele language newspapers and leaflets, have been deliberately excluded owing to the orthography used in them.

The earliest publication of fictional work in Ndebele dates from 1957. It should be noted that the Ndebele language corpus is largely composed of novels. However, a number of novels originally planned to be included in the text corpus were later excluded, some temporarily. Most of the books published in the sixties and seventies were on cheap quality paper. It is difficult and time-consuming to scan works on this kind of paper, most of these works also having been printed in a small font size. If such texts had to be scanned, it means that the time of proofreading them is almost the same as that of typing them. However, the compilers of the Ndebele language corpus had a time frame and target to meet and apart from corpus building, they were also compiling a dictionary, which had to be completed within a given time and target date. Under these time constraints the compilers preferred to scan and proofread those texts that only took the minimum time.

The majority of books included in the corpus were therefore published within the last twenty years, that is, between 1979 and 1999. As already mentioned, the bulk of these are creative works, especially narratives. No poetry collections or anthologies have been included and there are no immediate plans to do so. Poetic language is not popular in general corpus work and for lexicographic purposes it would be less useful. One drama text has been included and so far also one textbook. More textbooks will be included as the corpus keeps growing. Scanning, proofreading and tagging textbooks are more demanding than doing the same with novels, for instance. For this reason it seems the compilers have postponed the inclusion of textbooks, which they will have to do eventually, if they keep to their original plan.

As the Ndebele language corpus consists mostly of novels, the selection criteria for this category must be discussed. Firstly, there were efforts to bring about a balance between male and female writers. There are more published male than female Ndebele writers, so an effort to include a representative sample of novels by women was made. Another selection criterion was the popularity of the works. Writers who are considered popular had their works included. Writers usually become popular when their works are either prescribed in schools or broadcast in the media. Two such leading Ndebele writers are a male, Ndabezinhle Sigogo, and a female, Barbara Makhalisa. All their works other than drama and poetry have been included. Some works were included on the basis of the richness of their language. Corpus compilers who are competent literary critics made these judgements. Novels were further chosen according to the themes they handle, for instance, attempts were made to have a representative sample of war novels, love and marriage themes, witchcraft themes, and historical novels. Some themes dominate, partly because of the colonial policy prescribing certain themes for writers.

Composition of the Corpus

The corpus consists of both oral and written texts, all transcribed and converted into machine-readable texts. The oral material can be subdivided into oral interviews, oral recordings (of classroom lessons, church sermons, court sessions, etc.) and radio and television recordings. The written texts include publications and manuscripts. Within the category of manuscripts are unpublished dissertations and some selected documents and manuscripts. The published texts are divided into novels, drama and textbooks. There are other materials that have been included such as newspaper articles and advertisements.

Texts in the Ndebele Corpus

Type of Texts	**Quantity in %**
Written Material Publications Novels Drama Textbooks Manuscripts Unpublished dissertations Unpublished documents	80
Oral Material Oral interviews Oral recordings Radio/Television recordings	18
Other Materials	2

(The percentages given are estimates; as the corpus keeps growing its composition is not static.)

The composition of the Ndebele language corpus reflects the history of publishing in Zimbabwe, especially that of the indigenous languages. The case of Ndebele is further complicated by the reliance on Zulu literature for the teaching of Ndebele. As the majority of publications are biased towards school textbooks and novels, these texts dominate the Ndebele corpus. Some efforts were made to include types of texts other than school textbooks and novels. One way of offsetting this imbalance was to include what has been categorised as manuscripts. These are mostly unpublished dissertations and other documents and reports. The dissertations were collected from Ndebele departments in the various teacher-training colleges. As they are research papers, they contain some form of formal academic language. For instance, some dissertations are on Ndebele grammar while others are on teaching methods. Dissertations on literary criticism of Ndebele were also sampled. These papers have a potential of yielding language that is not ordinarily found in novels. However, the major limitation of these manuscripts is that as unpublished works they remain private and personal, and the language they contain is not standardised.

Similarly, as far as the oral material is concerned, it was felt that more was needed than the data collected through interviews throughout the Ndebele-speaking districts. The oral interviews were complemented by recordings of programmes from radio and television stations. The advantage of these is that compilers would have listened to or seen the programmes and so could choose whether to include them or not. Such material could also be systematically chosen to find the desired types of material. The disadvantage, however, was that compilers could not obtain any previous recordings, as the stations of the Zimbabwean Broadcasting Corporation destroy all tapes about two weeks after having broadcast them. Therefore, material was to be limited to that broadcast during the collection of the corpus material. One other disadvantage of radio and television material is that it lacks adequate biographical details of informants in terms of age, sex, occupation and educational background. Such information is essential for various research purposes and all oral interviews therefore have such details marked.

What has been labelled as other materials include various types of language, those that can be found in advertisements, posters or letters. One other notable kind of material in this category is the unfinished Ndebele dictionary that was supposed to be published by Longmans Zimbabwe.

Conversion of Texts

The coming of computers into language study has helped to address the question of corpus accessibility to other researchers. However, before the corpus can be shared by many researchers, it must be made machine-readable. The ALLEX Project corpora (both Ndebele and Shona) use the Standard Generalised Mark-up Language (SGML). This is in line with the international choice and preference of this method. "Because of its power, flexibility and independence of particular software systems," says Kennedy (1998: 82), "the Standard Generalised Mark-up Language (SGML) has become increasingly accepted outside the publishing industry as the standard way of encoding texts."

In addition to the SGML, the text encoding initiative guidelines were also followed. Although these are not standardised, they are flexible and adaptable to the compilers' needs. According to Kennedy (1998: 83) "the TEI Guidelines were designed to apply to any texts regardless of the language, the date of production or the genre". The use of these internationally used mark-up techniques makes the Ndebele corpus accessible to most international users and it can be rated as user-friendly and up-to-date.

As mentioned previously, the compilers tagged the texts to suit their immediate lexicographic needs. As "the corpus compiler has flexibility as to how much detail is marked-up for any particular corpus" (Kennedy 1998: 84), there is room for additional tags depending on the needs of the researcher. For instance, most of the Ndebele oral corpus has tags giving the biographical details of informants such as age, sex, education and occupation. There are also details on the header about the district where the material was gathered.

Implications of the Content of the Corpus

Collecting oral material was not easy for the compilers of the Ndebele language corpus. Apart from the large financial resources that were expended on the activity, there was also the problem of who should be interviewed. The majority of those who are counted as Ndebele speakers today are in fact mother-tongue speakers of the so-called minority languages such as Venda, Kalanga, Nambya, Sotho and Tonga. An ideal Ndebele language corpus should be the language of mother-tongue speakers of Ndebele; however, in reality, the Ndebele language is spoken by people whose first language is not Ndebele. The question of choosing from which districts to collect oral material was therefore problematic as it brought a number of sociolinguistic and political factors into play. Even if one were strictly to isolate mother-tongue speakers of Ndebele (if that is ever possible) there is still the problem of English that seems to be a characteristic of most Ndebele speakers, especially the middle-aged and most urban dwellers. There is a lot of code-mixing and code-switching and outright use of English words. This becomes problematic in transcribing the tapes as it gives orthographic problems. Decisions had to be made in certain cases whether to write some words in their English spelling or give them a Ndebele version. Notwithstanding the above-mentioned problems, the ALLEX Project of the University of Zimbabwe began to compile the Ndebele language corpus basing it on the language as currently spoken by mother-tongue speakers.

Conclusion

It has been shown that the content of a corpus depends on a number of factors that include sociolinguistic, political and economic considerations. It has also been shown how these factors have a bearing on both the content and quality of published and oral texts that constitute the Ndebele language corpus. As this corpus grows year by year, the present limitations would be addressed gradually. For instance, as the present state of the corpus is predominantly creative work, a deliberate effort would be made to include more textbooks so as to create a balance. The changes in the status of Ndebele as a language in Zimbabwe might significantly influence writing and publishing in the Ndebele language and thus influence the content of the corpus.

References

Hachipola, J.S. 1998. *A Survey of the Minority Languages of Zimbabwe*. Harare: UZP.

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

Sinclair, J.M. (Ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London: Collins.