The Shona Corpus and the Problem of Tagging[*]

Emmanuel Chabata, *African Languages Research Institute, Department of African Languages and Literature, University of Zimbabwe, Harare, Zimbabwe*

Abstract: In this paper the writer examines problems the African Languages Lexical (ALLEX) Project (at present the African Languages Research Institute (ALRI)) encountered while tagging the Shona corpus. The problems to be highlighted include general problems which apply to more than one language as well as problems peculiar to Shona. The paper was inspired by the challenges the writer encountered when he took part in building the Shona corpus. An analysis of the problems that most corpus builders face shows that more problems are likely to be encountered when dealing with spoken corpora than with written corpora. The paper demonstrates that tagging is an important component of corpus building as it makes it easier for a researcher to extract relevant data. To utilise the benefits of a tagged corpus, the tagging should be thorough and accurate. Well-informed decisions form an integral part of the tagging process since the utility of a tagged corpus depends largely on the input of the tagging process. This paper shows the need to take the tagging process seriously.

Keywords: ALLEX PROJECT, COMPUTER, CORPUS, ENCODING, FOREIGN WORD, LEMMATIZATION, LEXICOGRAPHY, MONITOR CORPUS, PART OF SPEECH, SCANNING, SHONA, SLANG, TAGGING, TRANSCRIPTION, WORD

Opsomming: Die Shonakorpus en die probleem van etikettering. In hierdie artikel ondersoek die outeur probleme wat die African Languages Lexical (ALLEX) Project (tans die African Languages Research Institute (ALRI)) teëgekom het terwyl die Shonakorpus geëtiketteer is. Die probleme wat bespreek word, sluit algemene probleme in wat van toepassing is op meer as een taal, sowel as spesifieke probleme wat eie aan Shona is. Die artikel het sy ontstaan in die uitdagings wat die outeur teëgekom het terwyl hy deel gehad het aan die opbou van die Shonakorpus. 'n Ontleding van die probleme waarvoor die meeste korpusbouers te staan kom, toon dat daar waarskynlik meer probleme teëgekom word wanneer daar met gesproke korpora as met geskrewe korpora gewerk word. Die artikel toon dat etikettering 'n belangrike komponent van korpusbou is, aangesien dit dit vir die navorser makliker maak om relevante data te onttrek. Om die voordele van korpusetikettering te realiseer, moet die etikettering deeglik en akkuraat wees. Ingeligte besluite vorm 'n integrale deel van die etiketteringsproses aangesien die bruikbaarheid van 'n geëtiketteerde korpus hoofsaaklik afhang van die inset tydens die etiketteringsproses. Hierdie artikel toon die noodsaaklikheid om die etiketteringsproses ernstig op te neem.

[*] This paper was presented at the Fourth International Conference of the African Association for Lexicography, held at the University of Pretoria, Pretoria, 5-6 July 1999.

Keywords: ALLEXPROJEK, REKENAAR, KORPUS, ENKODERING, VREEMDE WOORD, LEMMATISERING, LEKSIKOGRAFIE, MONITORKORPUS, WOORDSOORT, SKANDERING, SHONA, SLENG, ETIKETTERING, TRANSKRIPSIE, WOORD

1. Introduction

Recent linguistic researches have shown corpora to be important as a source of data for linguistic analyses. As Kennedy (1998: 88) for example observes, a linguistic corpus, in whatever form, is important as a basis for more accurate and reliable descriptions of how languages are structured and used. Thus, they are a source of evidence for linguistic descriptions and other researches that have to do with exploring language use. Renouf (1987: 1) has defined a corpus as "a collection of texts, of the written or spoken word, which is stored and processed on computer for the purposes of linguistic research". As implied in this definition, the term corpora is synonymous to machine-readable and computer-processed data. This does not, however, play down traditional corpora where texts were put on slips of paper and where relevant information could only be accessed manually.

Corpora may exist in two forms, namely, nontagged/unannotated or tagged/annotated. The difference between these two forms is that an unannotated corpus is just a plain text in its raw form, whilst a tagged corpus is enhanced with different kinds of information attached to each text or to items in a text. A close look at these two forms would probably show that it is more advantageous to work with an annotated corpus than with a plain one. McEnery (1996: 24) observes that annotating considerably increases the utility of a corpus. This is because information that may be implicit in plain text is made explicit through concrete annotation. In this case annotation makes it quicker and easier to retrieve and analyse relevant data from the corpus.

The term "tagging" has received different interpretations from different scholars. Svartvik (1982: 92) defines it as "the assignment of a lexical-grammatical description to linguistic units in the transcription from the audio tapes". Whilst this definition captures the idea that tagging is a way of marking those linguistic features that describe linguistic units, it tends to restrict the tagging process to spoken corpora, that is, to corpora built from oral material recorded on audio tapes and then transcribed. However, material in a corpus can also come from other different sources which do not necessarily involve transcription. This material can also be marked with codes that indicate various linguistic features. As a result we may need a definition that also incorporates corpus material collected from books, magazines, newspapers as well as material in electronic form. We can thus work with a more general and inclusive definition by McEnery (1996: 36), who defines tagging as "the attachment of specific codes to words in order to indicate particular features". The choice of linguistic features to be marked, depends on the kinds of studies for which the researcher wants to use the corpus. An example of the kinds of linguistic information that can be attached to items in a corpus, are shown in the sentence **Amai vanoda sadza** (Mother likes sadza) which can be tagged as follows:

<subject>Amai</subject> <subject concord><tense>va</tense></subject concord><aspect>no</aspect><verb stem>da</verb stem>
<object>sadza</object>.

The tagging in this short sentence is aimed at showing the different roles that each item plays in the sentence. **Amai** plays the role of the subject, **v**- is the subject concord, -**a**- is the present tense marker, -**no**- marks aspect, -**da** is the verb stem and **sadza** is the object. It is, however, important to note that these are not the only kinds of information with which each of these items can be marked. One can also add information about word category, noun class and any other features peculiar to each of the items. Using the codes presented in this example, we could, for instance, extract all subjects, verb stems or objects that are marked in the corpus. This is done by instructing the computer to extract all data with a specific code.

The process of tagging can be compared to the process of putting clothes into a wardrobe where one does not just dump all of them into one pile. Instead, one would put shirts in their own compartment, trousers in another and socks in yet another. This is not a purposeless arrangement. Instead, this is done so that one can have easy access to a particular type of clothing for which one is looking. As already noted, in corpus building, tagging involves the marking of a plain text. Specific features of linguistic units, for example, word category and grammatical function, would be indicated through the use of specific codes. As with clothes in a wardrobe, the codes place linguistic units with similar features into specific compartments which the researcher has created. Using particular codes, a researcher can easily retrieve only the information which is relevant to his/her study.

Tagged corpora have been found to be more and more useful in language-related studies such as lexicography, dialectology, semantics, syntax, psycholinguistics, sociolinguistics and code-switching and code-mixing. These disciplines either focus on how language is structured or how it is used, and conclusive evidence for any controversial issues in these areas can only come from instances where language is found in use. A well-collected corpus should be a true reflection of the everyday use of language, and it should be a useful source of linguistic evidence. This is especially so in cases where the corpus or corpora have been built from texts collected from ordinary speech. Ordinary speech is taken here to refer to conversations that take place while people are interacting in the manner they do in their daily life.

It is, however, important to note that a tagged corpus is most useful when the tags are accurate. Producing an accurately tagged corpus is not an easy task; it has a number of challenges. This paper will highlight some of the constraints by looking at the tagging of the Shona corpus. However, before looking at problems, we look at the Shona corpus, that is, how and why it was built. This would make it easier to understand and appreciate some of the challenges, especially those that have already been encountered by the ALLEX team.

2. The Shona corpus

The Shona corpus, with a current size of about two and a half million words, is a product of the African Languages Lexical (ALLEX) Project (at present the African Languages Research Institute (ALRI)). The ALLEX Project is housed in the Department of African Languages and Literature at the University of Zimbabwe. The Project was launched in 1992 and its major objective is to publish reference works that enhance the development of the indigenous languages of Zimbabwe. Some of the Project's envisioned publications include both monolingual and bilingual dictionaries. The Project also hopes to publish glossaries, for example, glossaries of musical, linguistic and literary terms, as well as specialized glossaries for ZimSign (Zimbabwean sign language with glossaries in Shona, Ndebele and English), science and technology (Chimhundu 1994: 21). All these activities need evidence of language use. As a result, the ALLEX Project made the building of corpora one of its priorities. In this case the corpora provide instances in which particular word forms are used for purposes of headword selection, defining and creation of usage examples.

To date, the ALLEX Project has published *Duramazwi ReChiShona*, the first Shona monolingual dictionary, and is currently working on two dictionaries, the Advanced Shona Dictionary (ASD) and the General Ndebele Dictionary (GND). Whilst the production of *Duramazwi ReChiShona* was corpus-aided, that is, its compilation was assisted with material from the corpus, that of the ASD and the GND is corpus-based. This means that headwords, senses and other relevant linguistic information required in the compilation of these dictionaries come from corpora in the two languages, Shona and Ndebele.

The research done by the ALLEX Project is based not only on theoretical linguistics, but also on linguistic data which shows how speakers use language. The aim is to publish reference works that reflect and represent actual language usage, not usage that is only theoretically possible. It is because of the nature of this research by the Project that the building of corpora became necessary.

The building of the Shona corpus started in 1992 when the ALLEX Project sent out student research assistants (undergraduate Shona students) to conduct interviews in all the districts in Zimbabwe where Shona is spoken. The interviews were on various socioeconomic, cultural, religious and political issues and comprised a representative sample of Shona discourse as it is used by the total population of the language's speakers. They captured vocabulary used in all varieties of Shona by people in different age and social groups.

For the collection of data in the field, each research assistant was provided with a cassette recorder and about 20 audio tapes. In order to obtain material from a variety of text types, context-governed material was collected. Some of these contexts included public speeches, church sermons, school lessons, lectures, individual interviews on people's life experiences, narratives of historical events and descriptions of major social events. In the process of systematically collecting this oral material, details on the context of discourse, for example, date of interview, physical location, topic, gender, age, education and social status of participants, the setting and other relevant details were recorded. Extralinguistic features such as hesitations, repetitions, shouts, coughs and whispers were also recorded and marked. The collection of the oral material resulted in 750 audio tapes of spoken Shona, which constitutes about 70% of the current Shona corpus.

After recording, the research assistants transcribed the material. Since the interviews were not edited, the transcriptions represent speech as it was recorded. The decision not to edit the interviews was taken to ensure that the oral material in the corpus was as natural as possible. The transcribed material was encoded or keyed into the computer, proofread and tagged. The tags include those for the header to identify each interview as a unit different from the others, and body tags which mark each speaker's utterances and the sentences that make them up. This oral material can now be accessed in three different modes, that is, as speech on audio tapes, as transcriptions on paper and in electronic form on the computer. In the corpus each interview constitutes a corpus text.

In addition to oral material, there is also written material which constitute the remainder of the corpus. This material came from a variety of sources such as novels, poems, drama and school text books. It was scanned, proofread, tagged and then stored in electronic form. The sources of written material include newspapers published in Shona (such as *Kwayedza*), magazines, pamphlets, advertisements, evidence taken down in magistrate courts (where it is given in Shona), material from the Zimbabwe parliament (mainly translations of parliamentary sessions) and the Zimbabwe Broadcasting Corporation (where radio stations 2 and 4 mainly broadcast in Shona and Ndebele). Each publication would then constitute a corpus text. Emphasis is now on collecting material already in electronic form, which is being done through negotiations with publishing houses that publish works written in Shona, since material in such a form minimise the laborious task of scanning and proofreading. The Project's ultimate goal is to develop and maintain a monitor corpus of spoken and written uses of Shona.

The Shona corpus can be described as a general-purpose corpus. Its use is not restricted to a specific type of linguistic research. As noted earlier, the Project's aim is that the corpus should be used as a source of data for a variety of linguistic researches. The corpus can also be described as a monitor corpus. According to McEnery (1996: 22), a monitor corpus is open-ended. Texts are constantly added to it so that it gets bigger and bigger as more samples are added. A monitor corpus is important for the ALLEX Project, which specialises in dictionary making. In fact, monitor corpora, according to McEnery (1996:

22), "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".

3. The tagging process and the challenges encountered

The importance of tagging a corpus has already been alluded to. Notwithstanding its significant role in the building of corpora, tagging has its challenges. Besides the need for a thorough understanding of the structure of the language with which the researcher is working, there are problems inherent in the corpus building process itself. In this section of the paper, the writer pays attention to problems associated with tagging, including those that have already been encountered by the ALLEX team. It is important to note that when discussing the Shona corpus as it now exists, two levels of tagging are involved. The first is the tagging that deals with marking each text as a unit as well as marking individual sentences in each speaker's utterance. This is what the ALLEX team has been doing. The second level is the grammatical or morphological tagging for which the team is developing programmes at present. This section deals with problems that have already been encountered at the first level of tagging as well as those that are being encountered in the process of developing the programmes that should cater for morphological tagging.

Some tagging problems are old problems that linguists have already encountered even before the establishment of corpus linguistics as a linguistic discipline or the acceptance of a corpus as a standard tool for linguistic research. These are problems that emanate from the way languages are structured. One such problem is the definition of the word "word". The notion of what a word is, has a number of possible interpretations, some of which are beyond the scope of this paper. In fact, what may be considered a word by one researcher may not constitute a word when viewed from another angle. The definition of a word is controversial and up to date linguists have not yet agreed as to what a word is or what it is not. In corpus building, the definition of a word is especially important when one is doing part of speech tagging, that is, when marking items with codes that indicate word-class category. However, in the process of marking it is sometimes difficult to distinguish a word from a phrase. At times it is difficult to establish a dividing line between these two. To illustrate this point we can take a Shona example, **Ndakazomuona** (I eventually saw him). Graphologically, this construction can be thought of as a word. However, a deeper analysis of this example shows that it is an inflected verb phrase consisting of:

Nda- (subject concord and tense),
-**ka**- (aspect),
-**zo**- (auxiliary),
-**mu**- (object concord),

-on- (verb radical), and
-a (terminal vowel).

Although they occur together as one form, each morpheme has a grammatical function which is realised at phrasal level. Given this information, it may not be enough to look at a word just as a group of letters written together and separated from the others by spaces.

The definition of what constitutes a word also poses a problem for word division. There are cases where it is not clear where and when to separate words in a sentence. This is especially so in cases where there is no clear definition of a word. To show this problem we can take the Shona form of the sentence "I was already going". The Shona translation of this sentence can be written in two forms, both of which seem to be acceptable, that is, (a) **Ndakange ndaa kutoenda** or (b) **Ndakange ndaakutoenda**. The problem here is whether **ndaa** and **kutoenda** should be written as one word form or as two separate units. This situation is a challenge to a corpus builder whose decision on such issues are reflected through the codes that he/she attaches to each unit.

Linked to the issue of word division are problems that emanate from the process of reducing spoken Shona to the written form. These problems are shared by the transcriber and encoder before they are passed on to the tagger, and they arise from the difficulties involved in the process of listening to and transcribing a spoken text in a way that is representative of the spoken language. In fact, it is difficult to have hundred percent accuracy when transcribing a spoken text. The whole problem according to McEnery (1996: 35) emanates from the fact that "in speech there is no explicit punctuation and any attempt at breaking down spoken language into sentences and phrases is an act of interpretation on the part of the corpus builder". Along with the unsettled problem of what a word is, the reduction of speech to writing posed a big challenge for those who worked with data collected through oral interviews. This is especially so given the fact that oral material in a corpus, when tagged, should represent original speech. The challenge is to come up with decisions that ensure that the resultant corpus captures the way the Shona language is structured or used both in the spoken and written modes.

For most linguistic research, part of speech or syntactic tagging is important. As noted earlier, this tagging process involves the assignment of codes to each word in a text as a way of labelling the word-class category to which it belongs in a particular context. Word-class tagging is especially important for researchers working with data restricted to a specific word-category, for example, noun, verb, adjective. If the tagging is done accurately, it should be possible for the researcher to capture only those words in which he/she is interested. However, in Shona, like in many other languages, the assignment of word-class tags is not always an easy task. There are times when the tagger comes across words which do not fit into any of the conventional word categories of the language one is working with. In Shona, for example, there are a substantial number of words that pose classificatory problems. In Chimhundu (1996)'s *Duramazwi ReChiShona*, most of these words that could not be fitted into any of the conventional word categories of Shona ended up being put into the **kanu** (interjective) category. Examples of such words include the following:

ahiwe (word used when giving someone a warning against his/her behaviour),
bodo (word used when denying something),
chagwa (word used for claiming something that has fallen),
mazvita (word used for thanking), and
diko (word used to show that one agrees to something).

Despite the fact that these words were classified under the same word-category, they are different in a number of respects. For example, whilst **ahiwe** can carry an exclamation mark as a sign of strong interjection, **bodo** cannot. The problem with such a classification is that words are tagged with similar codes irrespective of the fact that they have different linguistic features if looked at closely. The adoption of such an approach when tagging would not help much in giving each of these words its distinctive features through the codes.

Related to the problem of word class assignment is the problem of lemmatization. Lemmatization involves the reduction of words in a corpus to their respective lexemes, that is, to the forms that one would look up if one were looking for the words in a dictionary. Besides ensuring that related words, for example, homographs are not classified under the same lemma, lemmatization also shows how derived word forms evolve from their respective base forms. A lemmatized text is therefore important to lexicographers, especially during the process of headword selection. If successfully done, it resolves the problem that most lexicographers face: that of determining what should be and should not be a headword. However, lemmatization is not an easy task. In Shona, for example, it is sometimes difficult to trace extended verb stems to their base forms. This is because sometimes when a verb is extended, it can become lexicalised, that is, it acquires new meanings that may not be related to the meaning of the base form, thus qualifying to stand on its own as a separate lexeme. To illustrate this point, we can take an example of the verb stem **-gadza** (1. make someone sit 2. put something on a fire 3. put someone in a position of responsibility) which is derived from **-gara** (sit) by the causative verbal extension **-dz-**. As we can see from the translations of the two forms, **-gadza** has acquired, besides its first sense, other new meanings which are not easily traceable to **-gara**. Also, in the everyday use of the verb **-gadza** no connection is made with **-gara**. The problem arises when one tries to link **-gadza** to **-gara** since, besides the literal meaning that **-gadza** has by virtue of adding the causative **-dz-** to **-gara**, it acquires more senses that have nothing to do with sitting as it is expressed in **-gara**.

Another difficult task with lemmatizing a text is the fact that most of the work has to be done manually. Although automatic lemmatization is possible, it is sometimes difficult to instruct a machine to make accurate decisions about tags, especially given the complexity of natural language (Biber 1998: 262). As noted by Kennedy (1998: 208), reliable automatic lemmatization also depends on reliable grammatical tagging. However, as we have already noted above, coming up with reliable grammatical tagging is a huge task.

The problem of reducing words in a corpus to their base forms is related to that of identifying the semantic input each morpheme makes to a construction. This is a tagging process that involves marking of semantic relationships between items in the text, for example, the relationship between agents and patients of particular actions. The tagging can also be used to show grammatical functions such as subjects and objects. This kind of tagging would be useful for those who would want to study the grammar of a particular language. Though important, the exercise has its constraints. Sometimes, a single morpheme is used to perform more than one grammatical function at the same time. In Shona, for example, the subject marker can also function as the tense marker in the same construction. To illustrate this we can take an example such as **Aenda mangwanani** (He went (today) in the morning). In this sentence, the **a**- in **aenda** is a marker for the subject as well as for recent past tense: **a**- cannot, therefore, be distinct in terms of its grammatical function. To try and show both functions may be very difficult.

Corpus builders also face challenges that are linked to language contact and language development. It is a known phenomenon that when two languages come into contact, there is a tendency to borrow vocabulary from one another. Words are adopted and adapted from one language to another. Since adoption and adaptation is a process, it is difficult to determine the status of some words at a particular time in their history. A word would enter a language as foreign, and with time it may settle by changing its linguistic characteristics in order to fit into the new linguistic environment. Once the word has settled, it becomes part and parcel of the new language and is no longer perceived as foreign. However, if it keeps its form and pronunciation in the new environment, it remains foreign and should be marked so in a corpus. To indicate that words are foreign is important for linguistic studies such as code-mixing and code-switching. Marking them as foreign, in this case, would make it possible for the researcher to easily access those sentences or contexts where these words are found without going through the whole text. However, we have just noted that the process to fit into the new language is slow. The slowness of this change poses the problem of determining whether, at a certain time, a word has remained foreign or whether it is still in the process of change or whether it has already settled.

In the Shona corpus, the problem is encountered when dealing with words that are borrowed from either English or Ndebele and are attached to Shona word forms. This can be seen in the following constructions: **ku**receiver (to receive), **ku**mix**a** (to mix) **ku**accept**a** (to accept) and **Ndaka**appl**a**y**a** (I applied). The problem is to represent the English forms as foreign in their respective contexts. One option of solving the problem would be to remorphologise and rephonologise such forms. However, the problem with such an option is to ensure that corpus users would only recognise easily the original form of the words in spoken language. For example, **Ndaka**apppl**a**y**a** may become **Ndakaapuraya** in which case, it no longer "looks like" English. Although this may be a way of trying to deal with the problem, it may not capture the manner in which most Shona people use or recognise this construction. Furthermore, remorphologising and rephonologising would also mean that the spoken text is edited. This in turn contradicts the principle that spoken corpora should mirror speech. The other option would be to tag the English parts as foreign. In this case, a construction like **ku**mix**a**, would be tagged as follows: **ku**<foreign>mix</foreign>**a**. Straightforward cases like this one are not a problem. The problem comes when one is trying to tag 'apply' as a foreign word in the construction, **Ndaka**appl**a**y**a** where a Shona vowel -**a**- has been introduced between consonants **l** and **y**. It is difficult to pick out and mark the English form without misrepresenting the way it is pronounced by most Shona people.

Related to the issue of foreign words is the problem of handling slang words. Some slang words come into a language for a short time and immediately fall out of use again, whilst others stabilise and end up being conventionalised in the language. This means that what may be regarded as slang at one point in the history of a language, may not be slang at another time. This poses the problem of determining whether a word is still slang or has settled in the language.

4. Conclusion

In this paper, we have seen some of the challenges that corpus builders face when tagging. Although examples were drawn from Shona alone, the problems discussed here can apply to most languages, particularly those in the Bantu family. What this paper has shown is that some problems come from the nature of the language under consideration, the language's interaction with other languages and also from the way the corpus material is collected.

This paper has shown that any aspect in a corpus can be tagged, depending on what one wants to study. This paper has given few examples of linguistic information that can be tagged, but has rather highlighted some problems that may be encountered in the process of tagging. Tagging is an important but big task in the building of corpora and should therefore be taken seriously.

References

Biber, D. et al. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Chimhundu, H. (Ed.). 1994. *The ALLEX Project: Second Progress Report.* Harare: University of Zimbabwe.

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

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

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

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

Svartvik J. et al. 1982. Tagging the London-Lund corpus of Spoken English. Johansson, S. (Ed.). *Computer Corpora in English Language Research.* Bergen: Norwegian Computing Centre for the Humanities.