Orthographic and Morphological Problems in Headword Identification, Selection and Presentation in ALLEX

Herbert Chimhundu, *University of Zimbabwe, Harare, Zimbabwe*

Abstract: This article discusses aspects of an on-going lexical computing project at the University of Zimbabwe, known as ALLEX, the acronym for African Languages Lexical Project. ALLEX is a collaborative, multi-faceted, long-term, computer-aided lexicography project that is intended to produce a series of dictionaries, glossaries and other language reference works in the indigenous languages of Zimbabwe, starting with the first ever monolingual Shona dictionary, which is already at an advanced stage of preparation, and which will also be the first corpus-based dictionary in Zimbabwe.

The article confines itself to problems relating to word formation processes in Shona, specifically with reference to the lexicographers' need to ensure consistency in (i) identifying word forms as headwords in the running texts of the corpus, (ii) selecting from among these headwords to decide which ones to enter in the dictionary, and (iii) presenting the entries in the standard orthography.

First, an outline is given of the project's background, objectives, priorities, guidelines and work in progress. The article then focuses on specific problems encountered, and discusses these, and some of the solutions, in the light of advances in computer lexicography, with particular reference to concordancing. It will become evident that the problems encountered by the ALLEX Team have to do with the unlimited capacity of the Shona language's basic and derivational word formation processes, which it shares with the other languages in the Bantu family. Therefore, it will be suggested that these problems, and the solutions that are being worked out, have much wider implications that go beyond Shona and Zimbabwe.

Keywords: AFRICAN LANGUAGES, HEADWORD IDENTIFICATION, HEADWORD PRESENTATION, HEADWORD SELECTION, LEXICAL PROJECT, LEXICOGRAPHY, MORPHOLOGICAL PROBLEMS, ORTHOGRAPHIC PROBLEMS

Opsomming: Ortografiese en morfologiese probleme by lemma-identifikasie, -seleksie, en -aanbieding in ALLEX. Hierdie artikel bespreek aspekte van 'n voortgesette leksikale rekenariseringsprojek by die Universiteit van Zimbabwe, bekend as ALLEX, die akroniem vir African Languages Lexical Project. ALLEX is 'n gesamentlike, veelfasettige, langtermyn-, rekenaargesteunde leksikografieprojek wat hom dit ten doel stel om 'n reeks woordeboeke, woordelyste en ander taalnaslaanwerke in die inheemse tale van Zimbabwe te produseer.

Die projek begin met die eerste eentalige Shonawoordeboek, wat reeds in 'n gevorderde stadium van bewerking is en ook die die eerste korpusgebaseerde woordeboek in Zimbabwe sal wees.

Die artikel beperk hom tot probleme in verband met woordvormingsprosesse in Shona, spesifiek met verwysing na die leksikograaf se behoefte om eenvormigheid te handhaaf by (i) die identifikasie van woordvorme as lemmas in die lopende teks van die korpus, (ii) die seleksie uit hierdie lemmas om te besluit watter om in die woordeboek in te sluit, en (iii) die aanbieding van hierdie inskrywings in die standaardortografie.

Ten eerste word 'n skets gegee van die projek se agtergrond, doelstellings, prioriteite, riglyne en die werk wat aan die gang is. Dan fokus die artikel op spesifieke probleme wat teëgekom is, en bespreek hulle en sommige van die oplossings in die lig van rekenaarleksikografie, met besondere verwysing na konkordansieskepping. Dit sal duidelik word dat die probleme wat deur die ALLEX-span teengekom is, te doen het met die Shonataal se onbeperkte vermoë tot basiese woordvorming en woordvorming deur afleiding, 'n eienskap wat dit met die ander tale in die Bantoefamilie deel. Daarom word daar aangedui dat hierdie probleme en die oplossings wat uitgewerk word, baie wyer implikasies het wat verder reik as Shona en Zimbabwe.

Sleutelwoorde: AFRIKATALE, LEKSIKALE PROJEK, LEKSIKOGRAFIE, LEMMASELEKSIE, LEMMA-AANBIEDING, LEMMA-IDENTIFIKASIE, MORFOLOGIESE PROBLEME, ORTOGRAFIESE PROBLEME

1. Introduction

This article discusses only a few aspects of an on-going lexical computing project at the University of Zimbabwe, coordinated by the present writer and known as ALLEX, the acronym for African Languages Lexical Project. ALLEX is a collaborative, multifaceted, long-term computer-aided lexicography project that is intended to produce a series of dictionaries, glossaries and other reference works in the indigenous languages of Zimbabwe. The first volume planned is a monolingual Shona dictionary, which is already at an advanced stage of preparation. The article discusses some of the problems encountered during work in progress, but it does not go into theoretical, computational, defining and editorial issues, all of which, it is hoped, will be handled at appropriate stages by different members of the ALLEX Team.

The current article confines itself to problems relating to the word formation processes of Shona and its adopted system of spelling and word division, insofar as these are relevant to the design of the dictionary, particularly at the macrostructural level, and specifically with reference to the lexicographers' need to ensure consistency in how they identify word forms as headwords in the running texts of the corpus, how they select from among all these headwords to decide which ones to enter in their dictionary, and then how to present these entries in the standard orthography.

What follows below is partly based on the writer's previous experience in compiling the *Addendum* for Hannan's *Standard Shona Dictionary* (1981), but the greater part is based on more recent experiences in ALLEX. From the project's technical reports (Chimhundu 1992c, 1993a and 1993b), related documents and correspondence, an outline is drawn of the project's background, objectives, priorities, guidelines and work in progress. The article then focusses on specific problems relating to headword identification, selection and presentation that are relevant to the methodology that is being used to produce the first corpus-based dictionary in Zimbabwe, which will also be the first monolingual dictionary in Shona, the language spoken by at least three-quarters of the population. It will be suggested that the problems encountered by the ALLEX Team have to do with the unlimited capacity of the Shona language's basic and derivational word formation processes, which it shares with the other languages in the Bantu family. Therefore, the problems encountered and the solutions being worked out in ALLEX at present have much wider implications that go beyond Shona and Zimbabwe.

These problems range from how the conjunctive spelling further compounds the problem of identifying word forms generally; what criteria to use to determine qualification for selection of monomorphemic and multimorphemic word forms as lexical entries; how to create concordances or, more specifically, how to select key-words for concordancing of running texts; how to use the concordances for identification and selection of lexical items by the agreed criteria; and then, how to present these in the dictionary without violating the rules of the orthography in Standard Shona and the lexicographical conventions already established in existing dictionaries.

Defining as such will not be discussed, as this is a different problem altogether which, it is hoped, the writer will come back to in a follow-up article.

Obviously, because of the very close similarity in the morphosyntactic structure of the languages of the Bantu family, and despite the fact that some of them use conjunctive while others use disjunctive spelling, the problems that are being encountered and the solutions that are being worked out in ALLEX should generate much wider interest. It should also be of much greater significance and application in the region, especially since, to the best knowledge of the writer, ALLEX is a pioneering collaborative project in computer-aided lexicography in the indigenous languages of the region, and perhaps of Africa as a whole, that is unique in terms of both design and methodology.

2. The ALLEX Project

When the ALLEX Project was launched in September 1992 there was neither a language policy nor a lexicographical policy in Zimbabwe, and bilingual dictionaries for second-language users predominated, as they still do, in the whole region:

> The indigenous African languages display a lack of comprehensive monolingual lexicographical description (Gouws 1990: 55).

The timing of the rather belated Zimbabwean effort has been fortuitous in the sense that the launch of ALLEX was preceded by advances in computer technology. These advances concerned specifically automated text processing, that are now providing African language lexicographers with at least a theoretical chance to catch up, that is, if we ignore the futuristic automated dictionary that is already being talked about in the technologically advanced countries as a potential rival for the conventional form of the dictionary that we all know of as a book.

Internationally, the same advances in computing have pushed the lexicon to the centre of linguistic analysis:

> There is now an ever-increasing interest in lexical matters among linguists and others working in the field of cognitive science (Ralph 1988: 219).

Being aware of these developments, the ALLEX Team designed their project, from its inception, to take advantage of both the advantages of using computers in corpus-based linguistic research and the revived international interest in lexical matters that have resulted from these technical advances.

Consequently, ALLEX eventually came to be launched as an experimental, north-south cooperative venture involving a team of researchers at the University of Zimbabwe, assisted by an internordic group of lexicographers and computer scientists from the University of Oslo in Norway and the University of Gothenburg in Sweden, as well as a consultant lexicographer from the University of Botswana. The bulk of the funding and equipment was provided by the two Scandinavian institutions and a serious training component was built into the three-year period that was budgeted for the production of the first dictionary that was envisaged, a 500 page monolingual Shona dictionary containing 25-30 000 entries targeted for secondary school and general users.

This pioneering project has already yielded a very substantial corpus of texts that are potentially of multiple use and the work that is already in progress, specifically the monolingual Shona dictionary, itself, is viewed both as an experiment in technical and collaborative methodology, as well as a model on which other dictionaries will be based. The next major publication that is planned in the series is a similar monolingual dictionary in Ndebele. Two training and planning workshops have already been held involving about 60 University of Zimbabwe staff and students working on the Shona dictionary alone. Some research visits have also been exchanged between the participating universities. More workshops and research visits are planned.

At the time of writing, the data collection and encoding stages have already been completed for the Shona dictionary and a nucleus is already in place for a mini-archive, most of which is already in machine-readable form.

Tape-recorded materials, mainly from interviews during extensive field trips across the country, were transcribed, typed and tagged in the computer during the encoding stage of corpus building. Selected books, articles in magazines, newspapers and other publications of various kinds were also encoded. Some texts were also encoded mechanically by scanning. Concordancing is now in progress, as well as headword selection, defining and preparation of a draft manuscript, from which both the macrostructure (of the dictionary as a whole) and the microstructure (of the organisation of individual entries within the dictionary) have already been fixed and described in a comprehensive style manual. A Shona metalanguage is being developed and, along with it, a comprehensive list of abbreviations. Where gaps in the terminology and list of abbreviations still remain, English ones are being used in the draft manuscript, to be replaced later by global editing.

From the foregoing, it will be clear that ALLEX is trying to achieve several things as quickly as possible and in tandem, notably:

(a) to formulate comprehensive lexicographical policies and plans;
(b) to actually carry these through;
(c) to nurture the framing of consistent policies in the related fields of education and language planning;
(d) to accelerate the standardisation and development of African languages through codification and documentation in order to strengthen them and to enhance their status so that they can reclaim the recognition they deserve in the motherland;
(e) to provide a variety of language reference works for a variety of mother-tongue users;
(f) to make synchronic dictionaries and specialised dictionaries that are not only varied in their uses and applications, but are also as user-friendly as possible.

With reference to (f), for instance, it was decided from the outset to use computers, not only to reduce the drudgery of sorting, arranging and analysing data, but also to run concordances from which at least some of the senses and examples can be drawn from natural language. For the same reason, the defining style that ALLEX has settled for is a judicious mixture of conventional and COBUILD formats (Sinclair 1987), as is evident from the following examples representing four different types of draft entries:

muti DK z 3>4 mi-. 1 *Muti* imhando yezvinhu zvinomera zvoga asi kana kudyarwa asi iwo uchireba kupfuura zvimwe zvose. 2 *Muti* mushonga unoshandiswa pakurapa, kana kukuvadza.

(LH n 3>4 mi-. 1 A tree is a plant that grows from the ground but is taller than other such plants. 2 *Muti* is medicine for treatment or harmful magic.)

-taura D 1 itik; *Kutaura* kubudisa mazwi nomuromo. 2 it; Kana vanhu vachitaura nyaya vari kukurukura. FAN *-bwereketa, -reketa.*

(L 1 i; To *speak is* to utter words through the mouth. 2 t; When people are talking about things they are discussing them. SYN *-bwereketa, -reketa.*)

zii K- ny. Kana munhu akakati *zii* anenge akanyarara kana kuti asingadi kutaura. FAN *mwii, mwiro, kwaka.*

(H- ideo. When one is like this one will be quiet or unwilling to talk. SYN *mwii, mwiro, kwaka.*)

-tsva K pr. Chinhu chitsva ndechisati chamboshandiswa, kuonekwa kana kuitika. *Vana vanowanzotengerwa hembe itsva pakisimisi. Kundengendeka kwenyika hachisi chinhu chitsva kunyika dzakaita seJapan. Paakaenda kunze kwenyika, gungwa chakava chinhu chitsva kwaari.* FAN *nyuwani.* PIK *-tsaru, -sharu.*

(H adj. A *new* thing is one that has not been used before or seen or happened. *Children usually have new clothes bought for them at Christmas. Earthquakes are not new things in countries like Japan. When he went abroad, the sea was a new experience for him.*)

asi DK bat. Izwi rinoshandiswa kuratidza rimwe divi renyaya kana mamiriro akaita zvinhu. *Kusevenza ari kuda zvake, asi ari kurwara.*

(LH conj. A word used to show the other side of the story. *He wants to work, but he is sick.*)

It will be evident from these examples that some set formulas of the COBUILD-type are being used in ALLEX definitions. The advantage of employing formulas in defining in an inflecting language like Shona is that verbal and other forms of entries that are neither phonological nor graphological words are actually presented in some inflected form within the context of the definition itself. This is not only more user-friendly, but it also reduces the need for separate illustrative examples. Where such examples are still deemed to be necessary, they actually serve to give further elaboration of contexts of usage.

It is generally accepted that COBUILD methodology has "moved the science of lexicography into a new phase" (Clear 1987: 61). However, this new phase, and its revolutionary technological advances, have their own problems. Some of these problems, which have already been encountered by ALLEX in the two areas of orthography and morphology, will be discussed in the remainder of this article.

3. Orthographic Problems

Shona settled for a conjunctive system of word division in 1931 when Doke's recommendations on the unification of its dialects were accepted (Doke 1931). This choice was indeed logical for an inflecting language although some critics of what was called Union Shona in the 1930s, such as Father Baker, later attacked what they called its excessive conjunctivism (African Languages and Literature, s.a.). It was recognised, even at that time, that the phonological word in Shona was marked by penultimate stress or length. Therefore, the word in the spoken language can, in fact, be defined as a group of syllables characterised or marked off by greater duration on the last but one syllable in the group.

One might expect that, in the writing system that represents the spoken language, the rules of word division are designed such that these groups of syllables are marked off by spaces and then called words. However, the written word in Standard Shona has been defined on the basis of meaning. The guiding principle, which was set out by Fortune (1972), is that a meaningful unit in the language is to be written as a separate word *if it cannot be divided* into smaller meaningful units. The picture is then complicated by a series of rules that qualify this guiding principle by enumerating various exceptions which must be made, notably in the case of compounds or complex nominal constructions, deficient verbs with compound predicates, conjunctives and interjectives, reduplicated verbs and substantives, as well as forms that have to be hyphenated.

The result of these contradictions and of the complexity in the statement of rules is that, although speakers have no problem in identifying the words that they use as units, they generally have considerable difficulty in applying the rules of word division in the written language, mainly because they are expected to identify lexical words on the basis of semantic content while, at the same time, recognising and respecting morphological forms and grammatical functions. For the lexicographer, this conflict is further compounded by the language's affixational word formation processes which will yield many entries that are either sublexical or multilexical (Gouws 1990: 62).

Conscious of the normative influence of dictionaries, the ALLEX Team often encountered problems that sometimes forced them to make quite arbitrary decisions about how to use the letters, digraphs and trigraphs of the Shona alphabet as it was prescribed in 1967, not only to spell entries consistently correctly, but also to select from the many variant forms across the language's dialects and registers, as well as new forms that have been incorporated or coined as a result of language contact, mainly through borrowing.

Consequently, certain decisions were included in the dictionary's style manual that depart from the conventions already set in the current orthography and in the existing bilingual dictionaries. It is generally acknowledged that some of the current rules and conventions are resented or resisted because they create problems for speaker-writers. In such cases, the ALLEX Team felt that some changes were necessary in order to arrive at solutions that would be more in line with the pronunciation and / or linguistic feelings of the speakers. Still, however, it is not the ALLEX Team's intention to depart from the system of spelling and word division that has been adopted for Shona, but only to make things easier for the users of that system. Frequent revisions of the rules do not help either the users or the standardisation process. For their part, lexicographers must be aware that the dictionaries they compile will be authoritative sources which the users, particularly teachers and students, regard as prescriptive instruments that are available for them to check the correctness or norms of orthography, pronunciation, morphology, the usage and the status of lexical items, as well as indications of lexical variants and language interference.

Variants in particular pose a problem for ALLEX, especially since the planned volume has to be quite selective because of its size and target readership. Furthermore, for reasons that are explained elsewhere (Chimhundu 1983, 1992a and 1992b), the new Shona dictionary will not attach dialect labels to its headwords. While it is accepted that each of the 35 Bantu languages listed by Michael Mann as being among the 85 African languages spoken by more than one million people, represents a chain of interintelligible lects, ALLEX's experience with Shona leads the writer to dispute Mann's further suggestion that there is so much variation at both local and individual levels that each of these languages can only be served by a single dictionary with considerable difficulty (Mann 1990a: 1-7).

Much depends on what Zgusta (1971: 15-20) describes as the lexicographers' preliminary work in studying the language situation in order then to make informed decisions on even how to go about data collection, and then how to select, construct and arrange their entries. Where previous Shona dictionaries have indicated dialect for variant forms as distinct from synonyms, it is evident that some of them are morphological, in the sense that different allomorphs have been selected for either the base form or the inflectional component, while others are phonological, in the sense that different tone patterns are used or, occasionally, forms pronounced with and without breathy voice occur in free variation. A few examples may be given here in these categories as follows:

1. SYNONYMS

-viga ~ -kotsa (hide)
-wana ~ -roora (marry)
gudo ~ diro (baboon)
tezvara ~ mukarahwa ~ bambo (wife's father / brother)
mukunda ~ mwanasikana (daughter)
mage ~ mashoronga (curds of milk)

2. ALLOMORPHS

nzeve ~ zheve (ear)
tsuro ~ tsuro (hare)
hwowa ~ bwowa (mushroom)
-dzimara ~ -dzamara (until)
nyange ~ nyangwe (eve though)
nyambo ~ nyn'ambo (joke, humour)

3. INFLECTIONS

handidi ~ handidiba (I don't want)
semurume ~ somurume (like a man)
ngatiende ~ hatiende (let us go)
ngekuti ~ nekuti ~ nokuti (because)
(nyama) ingonaka ~ inonaka ((meat) is nice)

4. TONE PATTERNS

tezvara DKD ~ KKD (wife's father / brother)
minyu DK ~ DD (ideo. of dislocating or spraining)
sekuru KKK ~ DKD (grandfather, uncle)
-chonya D ~ K (wink)

5. BREATHY VOICE

-nhonga ~ -nonga (pick up)
rinhi? ~ rini? (when?)
vhazu ~ vazu (ideo. of being startled)
mujonhi ~ mujoni (white police officer)

A variety of historical and sociolinguistic factors have now made the language situation so fluid that any assumptions about who uses which of these forms, and where, are bound to be misleading.

Sometimes, the variant forms actually turn out to be contracted forms that are used interchangeably with the fuller forms, as in the case of:

-zurura ~ -zura (open)
gambimbisirwa ~ gambiswa ~ gambi (in actual fact)
hondohwe ~ hondo (ram)
dovamutova ~ dova (dew).

Occasionally, different derivational processes create shorter and longer forms of a noun from the same verb root, as in:

-kwidiba (close) > hwidobo ~ hwidibiro ~ hwidibidzo (lid).

It would be pointless to try to consistently indicate dialect usage in such cases, just as it would be pointless to indicate dialects against different forms of borrowed words, as in the case of the nouns:

sibhedyera ~ chibhedyera (<Nguni *isibhedlela*: hospital)
kero ~ keri (<English *c/o*: address)
jekiseni ~ *jekisoni* ~ *jakisoni* (<English *injection*).

It is a well known fact that language thrives in variation. One way or another, the lexicographer's consistent application of criteria for selection and presentation of entries will influence the users' perception of what may be deemed to be the norm or standard usage.

4. Morphological Problems

The inflecting capacity in the morphosyntactic structure of Bantu languages creates a major problem on how to determine what should be identified as a headword in a dictionary and what should not. As Charles Bwenge (1990: 5) has observed:

> The central problem is particularly the method of arranging the nominal and verbal items of the language, emanating from the complex morphological structure common to Bantu languages, of a morphological classification system categorising nouns by means of prefixes and a verbal derivation system forming new verbs by means of derivational affixes.

The arbitrariness and inconsistency that Bwenge has described in the treatment of relatedness between base and derivative forms in Swahili dictionaries have also been observed in Shona dictionaries, but to a lesser degree, particularly in Hannan's *Standard Shona Dictionary* (2nd edit., 1974), in which meticulous care was taken to fully utilise the linguistic findings of the time that it was compiled, as these were described within the framework of the constituent structure analysis by Fortune (1980, 1984).

This analysis allocates all the base forms to three categories in which all constructions take their place in three hierarchies — verbal, substantival (nominal and qualificative) and ideophonic. The monomorphemic verbal and substantival stems at the bottom of the first two hierarchies take on a whole range of affixes and inflections in prefixal, infixal and suffixal positions to produce word forms. Some of these are quite complex, as, for example, reduplicated verbs, multiple-extended verbs and complex nominal constructions. Further, various derivational processes may be used to form words in one hierarchy, that are built on base forms from another hierarchy. There is unlimited potential for the derivation of nouns from verbs and ideophones, for example, and for the derivation of verbs from ideophones and nouns.

The present writer found the problem of complex morphological structure acute about fifteen years ago when he was compiling the *Addendum* to Hannan (2nd. edit., Reprint, 1981). Consider, for instance, the following examples that are given with morpheme boundaries indicated and base forms capitalised:

mu-GARA-dza-ka-SUNG-w-A
(Lit.: one who STAYs with them (handcuffs) TIEd up always,
i.e. policeman)

chi-VAKA-SHURE
(Lit.: that which BUILDs from BEHIND)

chi-URAYE-URAYE (reduplicated ideophone < verb stem: -uraya)
(Lit.: manner of KILLING and KILLING, i.e. indiscriminate killing))

It is very easy to coin polymorphemic nominalisations, including gnomes or reduced sentences, because the productive patterns available are numerous. Therefore, selective criteria need to be set for inclusion in the dictionary. No compiler can be expected to list all the complex nominal constructions that are permitted by the grammar and are also acceptable to the speakers.

Similar problems arise with verbal extensions which are numerous and which are often combined, even in reduplicated forms, as is shown in the following sets of examples, all derived from the verb root *-bat-* (hold, catch):

(i) -bat-W-a (passive)
-bat-IR-a (applied)
-bat-IK-a (neuter)
-bat-AN-a (reciprocal)
-bat-IS-a (causative)
-bat-IS-a (intensive)
-bat-ISIS-a (double intensive)
-bat-IRIR-a (perfective);

(ii) -batirwa, -batiswa-, -batisiswa, -batisiswa, -batiririrwa;

(iii) -batwabatwa, -batirabatira, -batirwabatirwa, -batikabatika,
-batanabatana, -batisabatisa, -batiswabatiswa, -batirirabatirira,
-batiririrwabatiririrwa.

The verb root can also be used to form simple nouns such as *mubati* (one who holds / occupies) and *mubato* (handle), complex nominals such as *mabatakii* (one who is in charge / holds a key position), or ideophones such as *bate, batei, batebate, bateibatei.*

After the problem of selection has been resolved, the lexicographer will still have to deal with arrangement. The above examples should make self-evident, related problems of conjunctive spelling and what has been described as "the Western European habit of strict alphabetical arrangement" (P.R. Bennett quoted in Bwenge 1990: 7).

The writer's experience from ALLEX and the *Addendum* to Hannan leads him to support Bwenge's view that, for the Bantu languages, affixation morphology must be treated as a property of the lexicon because the affixation rules are so basic to the whole system. Not only are they syntactic in that they govern the concordial system, but they also yield the most productive processes in word formation. Therefore, the dictionaries must make it easier for learner-users to recognise and to appreciate how this happens, so that they, in turn, can produce well-formed lexical items. The lexicographer must not merely assume that the user will somehow be able to distinguish and to see the relatedness between derived and non-derived forms. This is why there is a general tendency in modern dictionaries to include explicit grammatical information, especially syntactic data:

> In the dictionary of today, the reader is offered more sophisticated information, more or less transparent, owing to the codification of syntactic data. (Gellerstam 1988: 103)

5. Precedents and Decisions

For the monolingual Shona dictionary, ALLEX has already worked out a comprehensive style manual that takes all these problems into account , the precedents already set in existing dictionaries and new conventions that were found to be necessary, even where there would have to be some deviation from past practice. Such a style manual is a must for any new lexicographic enterprise, and, with reference to the problem areas under discussion, the style manual must beforehand consider the following:

(a) which word forms will be selected as main entries;
(b) which other entries will be cross referenced to the main entries;
(c) which affixes, if any, will be listed as entries and in what manner;
(d) how to present the rules that will help the learner-user to produce well-formed lexical items; and
(e) how to handle problems of alphabetization that might arise from any of the above.

Concordances which are being run on a special programme will be used to provide varied contexts from the corpus that will help the editors to make appropriate decisions regarding selection of headwords and senses. ALLEX recognised from the outset that

> concordances are the most popular product of literary and linguistic data processing today (Kipfer 1984: 166)

but, unlike COBUILD, it was never the intention of the ALLEX Team that each observation about semantics, syntax and lexis should be adequately exemplified from text drawn from the corpus (Clear 1987: 42).

For the Shona dictionary, concordances will only be used selectively rather than routinely to augment a pre-selected headword list in a core-manuscript and to refine definitions. The selection process must therefore be done very carefully. It is especially important to use concordances to check for irregular words and usages, even of familiar words. For example, the present writer only realised that the sense "about to" should be added to the others previously listed for the verb *-da*, generally defined as "like, want", after studying the concordance run by project consultant Daniel Ridings on a specially designed programme. With this programme, items are worked backwards from right to left and different forms of words can be called for. Various endings must therefore be noted first before searches are initiated. In the case of the root *-end-* (go), for example, searches were made for, among other items, *-enda, -endawo, -endapo, -endai, -ende,* and *-endei,* for which concordances were run from the encoded corpus. Thus, the printed concordances already at hand show that any sequence of letters can be picked out of any graphological form regardless of that sequence's location in that form and regardless of word class. For the noun *amai* (mother), for example, *amai, vanaamai, dzaamai, kwaamai,* etc. would also be picked up.

Word forms selected for concordancing on specified criteria, such as frequency of occurrence in the encoded corpus, are arranged systematically and by consistent criteria. Verbs, for example, are worked backwards and arranged by their endings because they are inflected with prefixes, while the nouns may be arranged by their singular forms only because Bantu languages have alternating singular-plural noun class prefixes (Mann 1990b: 44-51). A list of the additional headwords selected from the concordance can be created mechanically from such a frequency count and items in the encoded corpus file can be cross-checked against the core-manuscript datafile. The concordance can thus be used for selection of senses to be used in defining both headwords already existing in the datafile and new headwords selected as additional entries from the corpus.

The concordances are also useful in identifying truncated forms and neologisms, as well as special uses and phonological environments of variant forms of such formatives as causative extensions *-is-* / *-es-*, *-idz-* / *-edz-*, and *-its-* / *-ets-*. The general idea is to search for concordance evidence for headwords, especially variants that are predictable in their form, and derivative forms such as extended verbs and deverbative nouns with meanings that are not predictable. The concordance, being an exhaustive index of the immediate contexts in which a particular word form occurs in the corpus, is ideally suited for this purpose as it will show all the possible forms of the word.

6. Conclusion

It is not possible, within the limited scope of this article, to outline all the identification, selection and presentation problems that have been noted so far as ALLEX progresses. The above should suffice to indicate, not only their nature, but also the implications for any similar lexicographic enterprise in Bantu.

Another general observation that can now be made is that advances in computer technology only reduce the drudgery and time taken, but they do not necessarily reduce the stages of dictionary-making that have been described by Zgusta and others. Neither does technology help to solve the problems that arise from lexicography's multidisciplinary nature, the need to follow or to establish a tradition and conventions, the lexicographer's normative responsibility, the need for a system or theoretical framework, and the need to remain conscious of the fact that one is doing scientific work for general practical use.

There are already a lot of things that can be done by computer to ease the lexicographers' tasks, while others cannot be mechanised because they require human intervention. By identifying and separating these tasks in their project design, compilers can produce dictionaries of high scientific quality much more quickly than ever before. Tasks such as homograph separation, defining and editing must remain part of the creative process that cannot be done mechanically as they are beyond the limits of artificial intelligence.

However, in addition to sorting and arranging or alphabetization, machines are already able to do a number of things. Concordance programmes can be manipulated to mechanically produce various contexts that provide useful information about shades of meaning and word-usage by lining up words by their prefixes and suffixes, as well as in free and infixed positions; by ranking the entire vocabulary of a file by word frequency; and by comparing vocabularies of different files. Lemmatisation by computer is also possible by using a pre-determined set of rules to classify words under their correct dictionary headwords. Using advanced forms of computer sorting, it is already possible to assign the correct lemma of each word; to have each word accompanied by its lemma; and to use what Kipfer has called a look-up dictionary for comparing and matching (1984: 167).

It seems as if African lexicographers are not yet eager to take advantage of these technological advances. The ALLEX experiment already shows two important things: that imported technology can be adapted to local, language-specific tasks; and that the lexicographers do not need to know all the computing technicalities involved themselves. After all, dictionary-making has almost always been team-work anyway. It would also be of great help to compare notes at a regional level, especially in view of the common problems anticipated in Bantu languages, some of which have already been indicated in this article.

References

African Languages and Literature. s.a. Efforts Towards the Unification of Shona. University of Zimbabwe. s.a.

Bwenge, Charles. 1990. Lexicographical Treatment of Affixation Morphology: A Case Study of Four Swahili Dictionaries. Hartmann, R.R.K. 1990: 5-17.

Chimhundu, Herbert. 1983. *Adoption and Adaptation in Shona.* Unpublished D.Phil thesis. Harare: University of Zimbabwe.

Chimhundu, Herbert. 1992a. Standard Shona: Myth and Reality. Crawhall, N.T. (Ed.). 1992: 77-88.

Chimhundu, Herbert. 1992b. Early Missionaries and the Ethnolinguistic Factor during the 'Invention of Tribalism' in Zimbabwe. *Journal of African History* 33: 87-109.

Chimhundu, Herbert. (Ed.). 1992c. *Report on the African Languages Lexical Project (ALLEX) Planning and Training Workshop.* Harare: University of Zimbabwe.

Chimhundu, Herbert. (Ed.). 1993a. *The ALLEX Project: First Progress Report.* Harare: University of Zimbabwe.

Chimhundu, Herbert. (Ed.). 1993b. *Report on the Second ALLEX Project Planning and Training Workshop.* Harare: University of Zimbabwe.

Clear, Jeremy. 1987. Computing. Sinclair, J.M. (Ed.). 1987: 41-61.

Crawhall, N.T. (Ed.). 1992. *Democratically Speaking: International Perspectives on Language Planning.* Salt River: National Language Project.

Doke, Clement M. 1931. *Report on the Unification of the Shona Dialects.* Hertford: Steven Austin and Son.

Fortune, G. 1972. *A Guide to Shona Spelling.* Salisbury: Longman.

Fortune, G. 1980-84. *Shona Grammatical Constructions.* 2 Vols. Harare: Mercury Press.

Gellerstam, Martin. 1988. Verb Syntax in a Dictionary for Second Language Learning. Gellerstam, Martin et al. 1988: 103-123.

Gellerstam, Martin et al. 1988. *Studies in Computer-Aided Lexicology.* Gothenburg: University of Gothenburg.

Gouws, Rufus. 1990. Information Categories in Dictionaries with Special Reference to South Africa. Hartmann, R.R.K. (Ed.). 1990: 52-65.

Hannan, M. 1974-81. *Standard Shona Dictionary, 2nd edit. — Reprint with Addendum.* Harare: College Press.

Hartmann, R.R.K. (Ed.). 1990. *Lexicography in Africa.* Exeter: Exeter University Press.

Kipfer, Barbara Ann. 1984. *Workbook on Lexicography.* Exeter: University of Exeter.

Mann, Michael. 1990a. A Linguistic Map of Africa. Hartmann, R.R.K. (Ed.). 1990: 1-7.

Mann, Michael. 1990b. The Impact of Computer Technology with Special Reference to Eastern Africa. Hartmann, R.R.K. (Ed.). 1990: 44-51.

Ralph, Bo. 1988. Basic Semantic Verb Structures. Gellerstam, Martin et al. 1988: 219-27.

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

University of Zimbabwe. s.a. *The Principal Dialects of Shona and the Development of the Standard Language.* Unpublished course notes. Harare: University of Zimbabwe.

Zgusta, Ladislav. 1971. *Manual of Lexicography.* The Hague: Mouton.