New Advances in Corpus-based Lexicography*

Arvi Hurskainen, *Institute for Asian and African Studies, University of Helsinki, Helsinki, Finland (arvi.hurskainen@helsinki.fi)*

Abstract: This article presents various approaches used in corpus-based computational lexicography. A claim is made that in order for computational lexicography to be efficient, precise and comprehensive, it should utilize the method where the corpus text is first analysed, and the results of this analysis is then processed further to meet the needs of a dictionary. This method has several advantages, including high precision and recall, as well as the possibility to automate the process much further than with more traditional computational methods. The frequency list obtained by using the lemma (the equivalent of the headword) as basis helps in selecting the words to be included in the dictionary. The approach is demonstrated through various phases by applying SALAMA (the Swahili Language Manager) to the process. Manual work will be needed in the phase when examples of use are selected from the corpus, and possibly modified. However, the list of examples of use, arranged alphabetically according to the corresponding headword, can also be produced automatically. Thus the alphabetical list of headwords with examples of use is the material on which the lexicographer works manually. The article deals with problems encountered in compiling traditional printed dictionaries, and it excludes electronic dictionaries and thesauri.

Keywords: LEXICOGRAPHY, DICTIONARY, LANGUAGE TECHNOLOGY, COMPUTATIONAL LINGUISTICS, AUTOMATIC COMPILATION, DICTIONARY TESTING, INFORMATION RETRIEVAL, MORPHOLOGICAL ANALYSIS, SEMANTIC ANALYSIS, DISAMBIGUATION, HEURISTICS

Opsomming: Nuwe ontwikkelinge in korpusgebaseerde leksikografie. Hierdie artikel beskryf verskillende benaderings wat in korpusgebaseerde rekenaarleksikografie gebruik word. Daar word aangevoer dat vir rekenaarleksikografie om doelmatig, noukeurig en omvattend te wees, dit die metode behoort te gebruik waarby die korpusteks eers ontleed word, en die resultaat van hierdie ontleding dan verder verwerk word om te voldoen aan die behoeftes van 'n woordeboek. Hierdie metode het verskillende voordele, insluitende 'n hoë mate van noukeurigheid en herwinning, sowel as die moontlikheid om die proses baie verder as met meer tradisionele rekenaarmetodes te outomatiseer. Die frekwensielys verkry deur die lemma (die ekwivalent van die trefwoord) as basis te gebruik, help met die keuse van woorde vir insluiting in die woordeboek. Die benadering word geïllustreer deur verskillende fases van die aanwending van SALAMA (die Swahili Language Manager) in die proses. Werk met die hand sal nodig wees gedurende die stadium wanneer gebruiksvoorbeelde uit die korpus gekies en moontlik aangepas word. Die lys ge-

* An earlier version of this article was presented as a keynote address at the Seventh International Conference of the African Association for Lexicography, organised by the Dictionary Unit of South African English, Rhodes University, Grahamstown, 8–10 July 2002.

bruiksvoorbeelde, alfabeties gerangskik volgens die ooreenstemmende trefwoord, kan egter ook outomaties voortgebring word. Die artikel behandel probleme wat teëgekom word by die samestelling van 'n tradisionele gedrukte woordeboek, en dit sluit elektroniese woordeboeke en tesourusse uit.

Sleutelwoorde: LEKSIKOGRAFIE, WOORDEBOEK, TAALTEGNOLOGIE, REKENAARLINGUISTIEK, OUTOMATIESE SAMESTELLING, WOORDEBOEKTOETSING, INLIGTINGSHERWINNING, MORFOLOGIESE ONTLEDING, SEMANTIESE ONTLEDING, ONDUBBELSINNIGMAKING, HEURISTIEK

1. Introduction

The use of computers in lexicographical work has gone through various phases, where enthusiasm on the one hand and disappointment on the other have alternated. The calculating power and speed of computers were thought to revolutionise the compilation of dictionaries, and high expectations were held for automating the process. It was thought that text corpora could be transformed into dictionaries with minimal human intervention.[1]

In this kind of thinking, two major mistakes were made. It was thought that strings in text would, with minimal modifications, become lexemes and possible dictionary entries. The other mistake was that there was no linguistic insight built into the system.[2] At best this approach resulted in various kinds of concordances where the occurrence of a word or a group of words could be retrieved from text with a needed amount of context, and sorted in selected ways. Much of the usefulness of computers in lexicography was seen just in these terms (Jones and Sondrup 1989; Panyr and Zimmermann 1989). The automatic concordancing was, of course, a huge improvement compared with manual compilation, but there was nothing linguistically intelligent in it. These retrieving programs, often called KWIC (Key Word In Context), continue to be standard tools in dictionary work, but they are suitable only for selected tasks.

Because a good dictionary is much more than a list of words, linguistic sophistication is required from computer-based lexicography. In order for the computer-based lexicographical work to be really meaningful, the computer system used for the work has to acquire and make explicit the linguistic information attached to each of the potential lexemes in the dictionary. These requirements include, *inter alia*

— the category of each word (part of speech),

— sufficient information for guiding in the use of a word, such as inflection, concordance, tone pattern, argument structure, etc.[3],

— semantic information, including glosses in bilingual dictionaries,[4]

— etymological information,[5] and

— the commonness of a word (frequency category).

Only fairly recently computational lexicography has come to the level where both realism and know-how make it possible to achieve significant advances (Teubert 2001). Much of the current work is still concentrating on the problems encountered in the lexicography of English and other Western languages. African languages raise different kinds of problems, including complex morphology, tonology, disjoining writing systems, etc., and these have to be faced and solved.

A major problem in the computational analysis of language is ambiguity. The extent of ambiguity varies among languages, but in every language it is a problem and needs to be solved. Ambiguity occurs on the morphological level, as well as on the syntactic and semantic levels. A word in isolation may have more than one morphological interpretation. It may have more than one syntactic function, and more than one semantic role, especially several textual meanings.

The computer system designed for lexicographical work should be able to address each of these problems and solve them. This calls for a full computational description of a language, a description that in great detail makes use of linguistic rules and is lexically comprehensive. In other words, the system should be able to analyse unrestricted text of a particular language.

In order to make the subsequent discussion more comprehensible, a description will be given of SALAMA (the Swahili Language Manager), a computer system designed for Swahili, a major Bantu language. Work on the computer description of this language started in 1985, and by now has reached a phase where almost all the problems have at least been addressed, and most of them solved.[6] The system will be briefly described phase by phase, and then by means of examples it will be shown how the system can be applied for dictionary compilation.

2. Choice of headwords

Data in language dictionaries are usually arranged under headwords ordered alphabetically. Good dictionaries also have sub-entries for listing such lexical words that are either derivatives of headwords or are in some other way closely related to the headword. Lexicographers consider the choice of headwords fairly difficult.[7] Because the final product of dictionary work has to be limited in size, a choice of headwords has to be carried out. Here we will discuss the choice of entries for a general language dictionary, although methods for semi-automatic compilation of domain-specific dictionaries have also been developed.[8]

We may think that a large enough and balanced corpus of general language text is a base for such a dictionary, and by retrieving the lemmas of words in the corpus we will get a reliable list of dictionary entries. The task is

not so simple, however. We need large amounts of various types of text for the corpus, and we also have to think about its representativeness. A problem with text-based lexicography is that words used mainly in spoken contexts will not be represented in text, and such words need to be considered separately. One method is to use transcriptions of spoken corpora as source for spoken language, but sufficiently large and representative spoken corpora are rarely available.

A systematic and comprehensive analysis of written language starts from the identification and analysis of individual words. More specifically, what we find in text is actually word-forms and not such words we find as dictionary entries. Such word-forms will be analysed morphologically, and each interpretation will be made explicit. Thus the interpretation of many word-forms becomes ambiguous, i.e. the word-form has more than one legitimate interpretation.

The concept of 'word' itself is also not as clear as it seems. In lexicography, we are more interested in grammatical words than orthographic words. Grammatical words fairly closely correspond to concepts, and it is the concepts and their definitions we need to deal with in lexicography. A concept may be represented in text by more than one string of characters. The treatment of such multi-word concepts may already be problematic in counting word frequencies of English (Kilgarriff 1997), but it can be detrimental in languages with a disjoining writing system (Hurskainen and Halme 2001).

Multi-word concepts can be treated as single concepts in automatic processing, especially if their constituent parts do not inflect and if they are adjacent to each other. This can be done by temporarily joining such word clusters together, and in the final version the words can be returned to their original shape. Grammatical words allowing other words between the constituent parts cannot be treated in this simple way, but there are means for treating them too (Tapanainen and Järvinen 1998).

One requirement for a useful system is that it has to be comprehensive. In other words, it should not leave words in text without interpretation, however rare or strange they are. There are two major reasons for this. There should be a 'master dictionary' that contains all the grammatical information of the language, as well as all lexical information. When compiling a smaller dictionary for a specific purpose, it is easier to filter out unnecessary analysed material than to cope with unrecognised (and unanalysed) words. Another reason for comprehensiveness is that in order for a disambiguating program to fulfil the task reliably there should not be unanalysed words in text.

If the text corpus is large and balanced enough, the core vocabulary of the dictionary can be selected on the basis of the lemma list arranged in frequency order. For example, we may think of choosing the 10 000 most frequent lemmas for a dictionary. Except for special purpose dictionaries, it is a good policy to include words in order of frequency in the dictionary. The point where the frequency list will be cut depends on the intended size of the dictionary. This

method ensures that at least all common words will be included.

This statement sounds trivial, but it is not trivial at all. In the comprehensive computer evaluation of five Swahili dictionaries (Hurskainen 1994, 2002), it was found that the two most authoritative dictionaries[9] had serious omissions in core vocabulary, although they had a fairly large percentage of words not found in any texts at all. The tests were made with three different corpora, totalling 4 227 362 words. The results show that the monolingual dictionary *Kamusi ya Kiswahili Sanifu* (KKS) was able to recognize between 89.7 and 91.8% of the words of the three corpora, and *Kamusi ya Kiswahili–Kiingereza* (KKK) recognized 90.7 to 92.9% of the words. At the same time, both dictionaries listed a number of such words not found in the corpus. Only half the nouns (precisely 50%) of classes 1/2, 3/4, 5/6, 7/8, and 9/10 listed in KKS were found in the corpus. The corresponding percentage in KKK was 55, i.e. it had less 'excessive' words. With verbs the situation was better: 78% for KKS and 85% for KKK.

If we compare these results with *Swahili–Suomi–Swahili-sanakirja* (Abdulla et al. 2002), which was also tested, we find interesting differences. This dictionary was produced by using a corpus as base for selecting headwords. Its success rate in recognising the words of the corpus ranges between 91 and 94%. In other words, it covers the vocabulary of the corpora slightly better than KKS and KKK. On the other hand, the percentage of 'excessive' nouns of the classes mentioned above was only 24%, and with verbs it was practically zero. In other words, only such verbs also used in the corpora were listed in the dictionary.

These statistics reveal the possibilities of modern language technology to show in detail weaknesses of existing dictionaries, as well as the improvements technology can bring to dictionary compilation.

This lengthy discussion on the problems of selecting headwords for a dictionary reveals that it is a major issue. The use of a frequency list of corpus lemmas is a safe method of avoiding at least major omissions.

The frequency list is, however, not the final entry list of the dictionary. The corpus is rarely so large and balanced that it alone provides all words needed, even for a fairly modest dictionary. Many words used in everyday life are often missing in the corpus, because such matters are not dealt with in texts. Names of flora and fauna are also insufficiently found in texts.

3. Format of the corpus

It was pointed out above that for the corpus to be maximally useful in dictionary compilation, the linguistic information of the text must be made explicit. Even the first task, i.e. the production of the lemma list, does not succeed in languages with left-branching (prefixing) inflection without a morphological analysis program capable of returning the correct lemma of each word-form. For automatic inclusion of relevant linguistic information needed in a dictionary, the linguistic analyser is an absolute necessity.

Therefore, it is not a question of whether the corpus should be tagged or not, but how and in what phase the tagging is to be performed. Principally there are two methods of tagging, both of them automatic. In one method, which is more traditional, the raw text is tagged with a computer program, and the tagged version of the corpus is then used by the lexicographer as source text. Queries are made to the tagged version, and tags can be used as search keys.

In another method, which basically performs the same operations as the one described above, the lexicographer works with raw text and uses the whole array of programs and utilities in compiling the dictionary. In this method, the user has the raw material (text) and a comprehensive set of tools (programs, utilities, filters, scripts, etc.), which can be used in a number of ways, depending on the type of task.

The latter method is better than the former for several reasons. The user is free to select or prepare their own texts without resorting to tagged corpora prepared by someone else, often for purposes not ideal for the current task. The user also avoids handling of excessively large files. On average, the analysed Swahili text is 16 times larger than the original text, and even after disambiguation it is still 11 times larger than the original. Any editor has difficulties in handling files of this magnitude.

The size problem can be conveniently solved so that the analysis and disambiguation are carried out 'in flight', which means that the user does not even see the results of these phases, because further processing can be carried out in pipe. In lexicography we do not need to see all occurrences of a word in the corpus. We rather want to know in what senses the word occurs in the corpus, and how many times it occurs in each sense. By condensing the format of the information, we do not lose any lexically important information, but the space required for presenting this is cut to a minimum. The larger the corpus, the bigger is the advantage. This method of lexicography requires a working environment, where piping of processes is possible, such as Linux and Unix.

4. Searching headwords from the corpus

How can the occurrences of a lexical word be found in the corpus? There are currently at least three methods for doing this. Each of these and their suitability for African languages will be briefly discussed below.

4.1 Direct string search — traditional approach

In languages with right-branching inflection and derivation, direct string search is not a major problem, because the potential headwords and their inflected and derived forms are adjacent to each other in alphabetical listing. In languages with predominantly left-branching inflection, the problem is more

serious, as is demonstrated in (1). Our task is to extract all occurrences of the verb *soma* (to read). As can be seen, the search string cannot be the whole verb stem, but only the root *som*, because the verb may also be ending in e or i, and various types of derivative suffixes can be added. Similarly, a large set of (strings of) prefixes has to be taken into account.

(1) Example of string search

```
[486] donner$ cat maj1999 | kw-alg 'som'
       ambaye ana***som***a darasa la pili √
    simu akiwa Mu***som***a, Dk. Mazara
  Masatu, Rajab M***som***a, Elia
    ilikuwa ni ku***som***a ile barua. √
    Kwa sababu wa***som***aji ndio wateja
             Wa***som***ali hao wanadaiwa
   kituo cha Tran***som***a Mabibo na Bw.
  sekondari ya Ka***som***a wilaya ya Musoma
    ndiye aliyewa***som***ea mashitaka √
     maelezo yali***som***eka kuwa anakufa √
     ikiwemo kuwa***som***esha. √
        Azizi ali***som***ewa mashitaka hayo √
   viongozi na wa***som***i ambao wamejaa tele
    huko pia ni m***som***i kwa kuwa ana
     na baada ya ***som***o, baadhi ya
   aendelee na ma***som***o. Bw. Hiza
    uamuzi huo ku***som***wa. √
   ili ushahidi u***som***we hadharani √
```

With the keyword *som* we are likely to get all the real cases, but also a lot of wrong words.[10] If we try to modify the search string so that wrong hits will be reduced, we run the risk of excluding real cases.

4.2 String search with regular expressions

The search is much more accurate if we use regular expressions in formulating the search key. If language analysis tools are not available in dictionary compilation, this is a valuable alternative. It is far more efficient than direct string search, but it is not even nearly as accurate and efficient as the compilation by employing language analysis tools.

Instead of using *som* as search key we have to approach the problem by also trying to describe other elements of the verb that are distinctive enough for separating them from other word categories. As the verb final vowels may be a, e, i and u, this is not a promising approach, because many word categories have similar endings.

A more promising approach is the description of verb prefixes, because there is usually a longer string of characters typical to verbs only. The problem

is that there are at least tens of thousands of such grammatical character combinations. Regular expressions, however, make the formulation of such queries possible, even practical. In (2), such a query has been used, and as the result shows, all findings now are verbs.

(2) Example of search by using regular expressions

```
[487] donner$ cat maj1999 | \
kw-alg '(ha)?(ni|u|a|tu|m|wa|i|li|ya|ki|vi|zi|ku|pa|mu)\
(na|li|ta|me|si)(ye|o|yo|lo|cho|vyo|zo|ko|po|mo)?\
(ni|u|m|mw|i|li|ya|ki|vi|zi|ku|pa|mu)?som'
        ambaye ana***som***a darasa la pili √
     ilikuwa ni ku***som***a ile barua. √
     ndiye aliyewa***som***ea mashitaka √
   na maelezo yali***som***eka kuwa anakufa √
      ikiwemo kuwa***som***esha. √
         Azizi ali***som***ewa mashitaka hayo √
     uamuzi huo ku***som***wa. √
    ili ushahidi u***som***we hadharani √
```

Even this search string is not accurate, because it leaves out the so-called general present tense, subjunctive, present tense negative, infinitive, and several more rare tense/aspect forms. It is difficult, and dangerous, to include such possibilities in the same search key, because the danger of getting unwanted strings will multiply.

Let us modify our previous task, so that instead of searching the verb *soma*, we look for all occurrences of each verb in the corpus. We cannot use the verb stem as part of the search key now, because there are thousands of verbs, and we do not know in advance what they are. We may try to simulate the verb stem by defining its minimum length. With some verb forms of monosyllabic verbs it is as short as two characters. Unfortunately this is also the length of the stem in many independent relative constructions, and in some it is even three characters. Thus it seems impossible to get an unmixed list of verbs only. Examples of found strings are shown in (3). Verb roots are in bold face.

(3) An attempt to retrieve verbs by using regular expressions[11]

```
[489]$ cat maj1999 | \
kwic -s '(ha)?(ni|u|a|tu|m|wa|i|li|ya|ki|vi|zi|ku|pa|mu)\
(na|li|ta|me|si)(ye|o|yo|lo|cho|vyo|zo|ko|po|mo)?\
(ni|u|m|mw|i|li|ya|ki|vi|zi|ku|pa|mu)?[a–z][a–z]+'
     wawakilishi wa CUF alichoki**it**a kuwa √
    mwanachama wa chama alichoku**wa**mo wakati √
        NAFCO, anadaiwa alii**tum**ia hali hiyo √
      Malera aliongeza, ali**it**wa mtuhumiwa namba √
      jana kuwa Mohamed ali**kir**i kosa hilo lakini √
      Bibi Subira kwani aliku**f**a kutokana na √
```

```
      mfupi baba yake alikuja na kuanza √
           Rais Mkapa alilia**mb**ia jopo hilo √
  kufungwa kwa duka, ali**lip**a faini hiyo. √
    nusu, Bw. Kahale alim**p**a fomu za kukata √
         wa mjadala, alivi**p**a changamoto vyombo √
       kwa shuti kali lililom**shind**a kipa Masuke √
 Hata hivyo, ilivuta usikivu wa washabiki.
   kwa shule hizo ni usimamizi mbovu wa
     tayari kupoteza utaifa wao na kama
 Magharibi ambayo si utamaduni wa wananchi
polisi waliambiwa na wanakijiji kuwa baadhi
    jijini jana kuwa wanamichezo hao walifariki
    kutokea Kenya na wana**milik**i silaha kali √
 Ilala Boma baada ya wanamuziki hao kudai
   na timu ya Vijana wanaume itashiriki
            Mmoja wa wasimamizi mlangoni
```

The search found 5,770 verb candidates, and as expected, there were independent relatives and also nouns that fulfilled the search criteria. Some of these are shown in (3). The precision was, however, very good: more than 98%. The recall was much worse. The analysis with SALAMA showed there were in addition 2 659 such words that were unambiguously verbs. Thus the recall was as low as 68%. This could be improved considerably by using search strings, which were excluded above and which could not be included in the same search.

The identification of a verb lemma is even more difficult than the identification of a verb. We could think of writing a program that would mark the beginning of a verb lemma for each verb in text. This code could then be used in retrieving the lines. In this way we would get a concordance list where the beginning of each verb lemma is marked. It would then be fairly easy to isolate the correct lemma, although a fairly large amount of manual work would be necessary.

4.3 Advanced approach — analyse text first

Although the use of regular expressions facilitates complicated search strings, it is still far from the precision, recall, and ease of the use of an approach where the text is first analysed linguistically. In this method, the following features are made explicit:

— The lemma or base form of the word can be defined so that it is identical with the headword of the dictionary. As a consequence, we get a list of words to be included in the dictionary.

— Part-of-speech information is given by the analysis program.

— The program produces a detailed list of morphological features of the

word-form found in text.

— Semantic features can be added. For example, the information on animality or humanness, may be necessary for defining the correct concordance pattern. Verbs may also be given information on their argument structure (SV, SVO, SVOO, etc.).

— If the dictionary is intended to be bilingual, semantic glosses in another language can be automatically produced for each dictionary entry.[12]

— Syntactic features (subject, object, various roles of verbs, dependent constituents in noun phrases, etc.) can be added. In dictionary compilation, such features are usually omitted.

— Information on the etymology of words can be added.

— Variant, or non-standard, orthography can be reported.

5. The problem of ambiguity

Word-forms often have more than one interpretation. A word-form may belong to more than one word class. English is a good example of this kind of ambiguity. In Bantu languages, ambiguity is often caused by the fact that the same morpheme is a marker of more than one noun class. Although word-forms may be ambiguous on the word level, in context they normally have only one interpretation. A general rule is that the more comprehensive the analyser is, the more ambiguity the result has.

There are two major approaches for solving ambiguity. One method relies on probabilities. If a word-form has two interpretations and one of these is common and the other rare, then the common one is chosen. The result is often correct, but one is never certain whether it is correct or not, because the choice was made on the basis of probability. In another method, ambiguity is resolved with context-sensitive 'linguistic' rules. For the vast majority of cases, context-sensitive rules fulfil the task.

Heuristic rules are used only for cases where there is no basis for constructing a linguistic rule. On the basis of morphological features, such rules try to guess the correct interpretation of the word. For example, if a word begins with m- and ends with -aji, the word is very likely a deverbative noun of noun class 1. It is self-evident that ambiguity can be resolved only in context, i.e. as part of real text.

(4) An example of ambiguity in Swahili

```
"<ofisi>"
        "ofisi" N 5a/6-SG ENG 'office'
        "ofisi" N 9/10-0-SG ENG 'office'
        "ofisi" N 9/10-0-PL ENG 'office'
```

```
"<ya>"
        "a" GEN-CON 3/4-PL
        "a" GEN-CON 9/10-SG
        "a" GEN-CON 5/6-PL
        "a" 5/6-PL-SP
"<kampuni>"
        "kampuni" N 5a/6-SG ENG 'company'
        "kampuni" N 9/10-0-SG ENG 'company'
        "kampuni" N 9/10-0-PL ENG 'company'
"<yake>"
        "ake" PRON POSS 3/4-PL SG3 'his/her/its'
        "ake" PRON POSS 9/10-SG SG3 'his/her/its'
        "ake" PRON POSS 5/6-PL SG3 'his/her/its'
"<iko>"
        "iko" 3/4-PL-SP LOC-17 'be (in place)'
        "iko" 9/10-SG-SP LOC-17 'be (in place)'
"<ghorofa>"
        "ghorofa" N 5a/6-SG AR 'storey, floor'
        "ghorofa" N 9/10-0-SG AR 'storey, floor'
        "ghorofa" N 9/10-0-PL AR 'storey, floor'
"<ya>"
        "a" GEN-CON 3/4-PL
        "a" GEN-CON 9/10-SG
        "a" GEN-CON 5/6-PL
        "a" 5/6-PL-SP
"<tano>"
        "tano" NUM 9/10-PL NUM-INFL CARD 'five'
        "tano" NUM NUM-INFL ORD 'fifth'
```

By using a Constraint Grammar parser (CG2) ambiguity is resolved with the help of context-sensitive rules. The process of resolving ambiguity is also called 'disambiguation'. The result is shown below.

(5) Ambiguity resolved

```
"<ofisi>"
        "ofisi" N 9/10-0-SG ENG 'office'
"<ya>"
        "a" GEN-CON 9/10-SG
"<kampuni>"
        "kampuni" N 9/10-0-SG AR 'company'
"<yake>"
        "ake" PRON POSS 9/10-SG SG3
"<iko>"
        "iko" 9/10-SG-SP LOC-17 'be (in place)'
"<ghorofa>"
        "ghorofa" N 9/10-0-SG AR 'storey, floor'
"<ya>"
        "a" GEN-CON 9/10-SG
```

```
"<tano>"
        "tano" NUM NUM-INFL ORD 'fifth'
```

6. Removing excessive tags

Experience has shown that the more detailed the analysis of words, the better possibilities it offers for linguistically motivated disambiguation. Therefore, all features should be made explicit in morphological and semantic analysis, because they may be needed in writing disambiguation rules. An example of complexity is provided in (6), where a few word-forms of the verb *andika* (to write) have been analysed. Note that morpheme boundaries (+) have been manually added, and ambiguity has been removed by rules, so that each form has only one interpretation.

(6) All tags retained

```
"<wa+me+mw+andik+i+a>"
"andika" V 1/2-PL3-SP VFIN PERF:me 1/2-SG3-OBJ SV SVO SVOO 'write' APPL
"<wa+me+ji+andik+ish+a>"
"andika" V 1/2-PL3-SP VFIN PERF:me REFL-SG-OBJ SV SVO SVOO 'write' CAUS
"<ni+li+andik+ish+w+a>"
"andika" V 1/2-SG1-SP VFIN PAST SV SVO SVOO 'write' CAUS PASS
"<a+li+andik+ish+w+a>"
"andika" V 1/2-SG3-SP VFIN PAST SV SVO SVOO 'write' CAUS PASS
"<a+li+ye+andik+a>"
"andika" V 1/2-SG3-SP VFIN PAST 1/2-SG-REL SV SVO SVOO 'write'
"<a+li+ye+zi+andik+a>"
"andika" V 1/2-SG3-SP VFIN PAST 1/2-SG-REL 9/10-PL-OBJ SV SVO SVOO 'write'
"<a+li+i+andik+i+a>"
"andika" V 1/2-SG3-SP VFIN PAST 9/10-SG-OBJ SV SVO SVOO 'write' APPL
"<a+li+yo+i+andik+a>"
"andika" V 1/2-SG3-SP VFIN PAST 9/10-SG-REL 9/10-SG-OBJ SV SVO SVOO 'write'
"<a+li+li+andik+i+a>"
"andika" V 1/2-SG3-SP VFIN PAST 5/6-SG-OBJ SV SVO SVOO 'write' APPL
"<a+mekwisha+mw+andik+i+a>"
"andika" V 1/2-SG3-SP VFIN PERF:mekwisha 1/2-SG3-OBJ NON-STD SV SVO SVOO
    'write' APPL
```

The description in (6) has much such information we do not need in a dictionary. Therefore we remove part of the tags and leave those that are useful. After having removed excessive tags, we get a more readable output as in (7).

(7) Part of tags removed

```
"<wamemwandikia>"       "andika" V SVOO 'write' APPL
"<wamejiandikisha>"     "andika" V SVOO 'write' CAUS
"<niliandikishwa>"      "andika" V SVOO 'write' CAUS
```

```
"<aliandikishwa>"          "andika" V SVOO 'write' CAUS
"<aliyeandika>"            "andika" V SVOO 'write'
"<aliyeziandika>"          "andika" V SVOO 'write'
"<aliiandikia>"            "andika" V SVOO 'write' APPL
"<aliyoiandika>"           "andika" V SVOO 'write'
"<alililiandikia>"         "andika" V SVOO 'write' APPL
"<amekwishamwandikia>"     "andika" V SVOO 'write' APPL
```

In (7), the analysis program was used in the mode that returned the basic verb lemma but retained the information on verbal extensions. For finding out verb frequencies in the corpus, this mode is useful, because it returns the base form of the verb regardless of its actual form in text. In dictionaries, we often need listing at least part of the extended forms, especially if their meanings are not directly derivable from linguistic rules. For such purposes, a format shown in (8) is better, because it returns extended forms as lemmas. These extended forms are often alphabetically listed as sub-entries after the headword.

(8) Verbal extensions in verbs retained.

```
"<wamemwandikia>"          "andikia" V SVOO 'write' APPL
"<wamejiandikisha>"        "andikisha" V SVOO 'write' CAUS
"<niliandikishwa>"         "andikisha" V SVOO 'write' CAUS
"<aliandikishwa>"          "andikisha" V SVOO 'write' CAUS
"<aliyeandika>"            "andika" V SVOO 'write'
"<aliyeziandika>"          "andika" V SVOO 'write'
"<aliiandikia>"            "andikia" V SVOO 'write' APPL
"<aliyoiandika>"           "andika" V SVOO 'write'
"<alililiandikia>"         "andikia" V SVOO 'write' APPL
"<amekwishamwandikia>"     "andikia" V SVOO 'write' APPL
```

7. Post-processing of the analysed corpus

When each word in the corpus is analysed and the ambiguity resolved, the result can be manipulated in a number of ways. In dictionary work, we in fact need several kinds of modifications to the result.

For the selection of dictionary entries, we need a frequency list according to the lemma. In order for the list to be correct, we need to remove the actual word-form and all such tags that describe inflection, as well as the codes of verbal extensions. By doing this, we may collapse the list in (7) above and get a single line as shown in (9).

(9) A format needed for counting frequencies of headwords

```
10 andika V SVOO 'write'
```

If verbal extensions are also counted as separate lexical entries as in (8) above, we get a list as shown in (10). Note, however, that if the list is sorted in fre-

quency order, the extended forms will not be adjacent to each other.

(10) Counting verbal extensions

```
3 andika V SVOO 'write '
4 andikia V SVOO 'write ' APPL
3 andikisha V SVOO 'write ' CAUS
```

When we have a list of words in lemma form we want to be included from the corpus in the dictionary, we sort the list according to the lemmas. The result is the skeleton of the dictionary, and the headwords are arranged alphabetically. The top part of such a frequency list is shown in (11). We note that it is not merely a list of lemmas, because different functions of the same word cause them to be counted separately. For instance, the word *na* has four different functions, and due to the function of the disambiguation program, we have four different frequencies for this word.

(11) Top part of the frequency list

```
145306  na CC 'and'
 62611  kwa PREP 'at, to, for'
 55907  katika PREP 'in, at'
 49686  ni DEF-V:ni 'be'
 31873  na AG-PART 'by'
 30087  na PREP 'with'
 21416  kama ADV 'like, such as (ar)'
 20649  wa V 'be'
 19084  na NA-POSS 'of'
 10814  baada_ya PREP 'after'
  9788  pia ADV 'also, likewise, too'
  9417  hata ADV 'definitely not, not even'
  8629  kwenye PREP 'in, at, about'
  8612  sasa ADV 'now (ar)'
  8089  tu ADV 'only, just'
  7955  sana AD-ADJ 'much, very, a lot (ar)'
  6498  pamoja_na PREP 'together with'
  6340  zaidi ADV 'more, beyond (ar)'
  6213  jana ADV 'yesterday'
  5748  hadi PREP 'till, until (ar)'
  5269  juu_ya PREP 'above, concerning'
  5240  si ADV NEG 'not'
  5096  kutokana_na PREP 'deriving from'
  5059  kila ADJ A-UNINFL 'all'
  5030  tena ADV 'again'
  4475  mbalimbali ADV 'different, various'
  4047  leo ADV 'today'
  3951  kati_ya PREP 'between'
  3818  bila PREP 'without (ar)'
```

The dictionary itself is ordered according to the headword, and for this reason we have to rearrange the data. We also want to retain information on the frequency of the words. Selected entries from the alphabetically arranged data, extracted from a small section of the news corpus, are shown in (12).

(12) Selected dictionary entries produced by SALAMA

43	awali	ADV 'first, originally (ar)'
32	awali	N 9/10 '1 first. 2 origin, cause. 3 above (ar)'
7	awamu	N 9/10 'phase'
4	azimio	N 5a/6 'declaration'
6	azma	N 9/10 'intention; desire, purpose'
13	baa	N 9/10 'bar, pub. (eng)'
561	baada ya	PREP 'after'
117	baadaye	ADV 'thereafter, afterwards, then, later (on). (ar)'
103	baba	N 9/10 'HUM father, (zamani) sire.'
28	badala ya	PREP 'in stead of'
20	badala yake	PREP 'in stead of him/her/it'
14	badiliko	N 5a/6 AR 'change'
92	bado	ADV 'not yet, still (ar)'
2	bagua	V SVO '1 separate. 2 discriminate against, segregate'
14	baina ya	PREP 'between'
14	baini	V SVO 'realize, recognize (ar)'
11	baiskeli	N 9/10 'bicycle, (hist) velocipede (eng)'
6	baki	N 5a/6 '1 remainder, residue; balance. 2 (chakula) left-overs (ar)'
36	baki	V SV '1 remain. 2 stay/be left behind (ar)'
1	bakiza	V SV SVO 'leave behind; leave (not taking everything)'
6	banda	N 5a/6 'shed, barrack, barn, hut; hovel'
5	bandia	N 9/10 '1 doll, dummy. 2 imitation (ar)'
6	banja	V SVO '1 crack; break, split (nuts, firewood etc). 2 strike. 3 (ms) bark up the wrong tree'
49	bara	N 9/10 'continent (ar)'
86	barabara	N 9/10 'highway, road, street, turnpike, way, avenue'
7	baraka	N 9/10 '1 blessing, benediction, boon, favour. 2 prosperity, progress, abundance (ar)'
	...	
1	plastiki	N 5a/6 'plastic (eng)'
3	plastiki	N 9/10 'plastic (eng)'
20	pombe	N 9/10 'local brew, beer'
21	ponda	V SV 'pound, crush, mash; smash, crash'
17	posho	N 9/10 '1 allowance. 2 food, ration'
13	potea	V SV '1 be lost. 2 be wrong, err'
13	potoa	V SVO '1 twist, make crooked/curved/slanting. 2 ruin, pervert, spoil'
2	potofu	ADJ A-INFL '1 stray; misleading. 2 spoiled'
19	profesa	N 9/6 'AN HUM professor. (eng)'
5	pumziko	N 5a/6 'pause; half-time, interval, break, recess'

5	punde	ADV 'soon, in a short while, shortly. (ms) ~ si ~ suddenly'
9	puuza	V SVO 'disregard, ignore, snub'
69	pya	ADJ A-INFL '1 new, recent, modern. 2 novel, strange'
16	rafiki	N 9/6 'AN HUM friend; comrade. (ar)'
30	raia	N 9/10 'AN HUM citizen. 2 civilian (ar)'
113	rais	N 9/6 'AN HUM president (ar)'
38	rasmi	ADJ A-UNINFL 'official, formal (ar)'

We see that some nouns are used in two different noun classes, and the frequencies of each usage are shown. Inflecting adjectives and non-inflecting adjectives have separate codes, which is necessary information for the dictionary user. Verb types are classified and marked with transitive (SVO) and intransitive (SV) tags. Etymological information, if applicable, is given at the end of the gloss.

In (13), we finally have a form where frequency information has been transformed into classes, the most frequent ones being marked with three dark dots, and the least frequent ones with no dots at all. Some further formatting has also been incorporated, all without manual intervention.

(13) Dictionary entries with frequency classes

awali	*adv* 'first, originally (*ar*)' ••
awali	*n* 9/10 '1 first. 2 origin, cause. 3 above (*ar*)' ••
awamu	*n* 9/10 'phase'
azimio	*n* 5a/6 'declaration'
azma	*n* 9/10 'intention; desire, purpose'
baa	*n* 9/10 'bar, pub. (*eng*)' •
baada ya	*prep* 'after' •••
baadaye	*adv* 'thereafter, afterwards, then, later (on). (*ar*)' •••
baba	*n* 9/10 'HUM father, (zamani) sire.' •
badala ya	*prep* 'in stead of' ••
badala yake	*prep* 'in stead of him/her/it' •
badiliko	*n* 5a/6 AR 'change' •
bado	*adv* 'not yet, still (*ar*)' •••
bagua	*v* SVO '1 separate. 2 discriminate against, segregate'
baina ya	*prep* 'between' •
baini	*v* SVO 'realize, recognize (*ar*)' •
baiskeli	*n* 9/10 'bicycle, (hist) velocipede (*eng*)' •
baki	*n* 5a/6 '1 remainder, residue; balance. 2 (chakula) left-overs (*ar*)'
baki	*v* SV '1 remain. 2 stay/be left behind (*ar*)' ••
bakiza	*v* SV SVO 'leave behind; leave (not taking everything)'
banda	*n* 5a/6 'shed, barrack, barn, hut; hovel'
bandia	*n* 9/10 '1 doll, dummy. 2 imitation (*ar*)'
banja	*v* SVO '1 crack; break, split (nuts, firewood etc). 2 strike. 3 (ms) bark up the wrong tree'
bara	*n* 9/10 'continent (*ar*)' ••
barabara	*n* 9/10 'highway, road, street, turnpike, way, avenue' •••

baraka *n* 9/10 '1 blessing, benediction, boon, favour. 2 prosperity, progress, abundance (*ar*)'

. . .

plastiki *n* 5a/6 'plastic (*eng*)'
plastiki *n* 9/10 'plastic (*eng*)'
pombe *n* 9/10 '1local brew, beer' •
ponda *v* SV 'pound, crush, mash; smash, crash' •
posho *n* 9/10 '1 allowance. 2 food, ration' •
potea *v* SV '1 be lost. 2 be wrong, err' •
potoa *v* SVO '1 twist, make crooked/curved/slanting. 2 ruin, pervert, spoil' •
potofu *adj* A-INFL '1 stray; misleading. 2 spoiled'
profesa *n* 9/6 'AN HUM professor. (*eng*)' •
pumziko *n* 5a/6 'pause; half-time, interval, break, recess'
punde *adv* 'soon, in a short while, shortly. (*ms*) ~ si ~ suddenly'
puuza *v* SVO 'disregard, ignore, snub'
pya *adj* A-INFL '1 new, recent, modern. 2 novel, strange' •••
rafiki *n* 9/6 'AN HUM friend; comrade. (*ar*)' •
raia *n* 9/10 'AN HUM citizen. 2 civilian (*ar*)' ••
rais *n* 9/6 'AN HUM president (*ar*)' •••
rasmi *adj* A-UNINFL 'official, formal (*ar*)' ••

If we want to furnish the dictionary with examples of use, as we normally do, we need to retrieve such examples from the corpus. In order to automate the process, we need a third kind of list where the lemmas (i.e. headwords) are attached to the actual word-forms in the corpus. Basically the production of such a list is simple, because it is the default format of the analysis result of SALAMA. The problem is that if we do a selection of lemmas according to frequency, it is not easy to delete the correct lemmas from the original list, because the frequency order there is completely different compared with the lemma list. The solution is to retrieve all such lines from the main list where the lemmas of our selection list occur. As a result, we have a list of only those words we intend to include in the dictionary, and the list also has accurate information on the actual word-forms we can use as key for retrieving examples of use in the corpus.

The search for examples of use can be performed in two ways. One possibility is interactive where the dictionary compiler checks from the corpus the use of each lemma by employing one of several search programs or a more user-friendly interface. The other possibility is to retrieve the needed examples with a program. The resulting file will have all those words in the context, for which we want examples of use. By sorting such lines according to the lemma, we get a list of examples of use in the same order as in the dictionary. It is then fairly simple for the dictionary compiler to select and modify suitable examples of use to be included in the final dictionary. In (14), we have an extract from an alphabetically ordered list of the use of words in context. This list was produced by a program which used the word-form (not lemma) as search key.

(14) Words in context

dai: *barua hiyo imesainiwa na watu 10 <waliodai> kuwawakilisha wenzao.
dai: *habari <zilidai> kuwa hatua hiyo inatokana na kile kilichoelezwa kuwa ni mtindo wa *bw.
dai: *hamad kuwataka wanachama wafanye subira kila wanapotaka kufanya jambo fulani la <kudai> haki.
dai: *hata_hivyo, <alidai> kuwa wafuasi wengine wa chama hicho waliendelea kushikiliwa na polisi na kwamba hadi jana mchana walikuwa hawajaachiwa.
dai: *hata_hivyo, majina ya wafuasi wengine <waliodaiwa> kushikiliwa na polisi hayakuweza kupatikana mara_moja.
dai: *ngawaiya <alidai> kuwa baada_ya yeye kufuatilia suala hilo polisi, alielezwa kuwa gari hilo lilikamatwa kwa_kuwa dereva wake hakuwa na leseni.
dai: *profesa *lipumba alisema chama hicho kitafanya maandamano hayo <kudai> mambo matatu.
dai: <*alidai> kuwa kwa sasa wafuasi hao wamefunguliwa mashitaka ya uzururaji.
dai: <*walidai> kuwa uamuzi wa kuteua nyumba zinazostahili kubomolewa ndani_ya bonde hilo umefanywa bila tathmini ya kitaalamu.
fariki: *gabriel *ngwilulupi alisema jana nyumbani kwa marehemu *ukonga *staki *shari, kwamba marehemu <alifariki> juzi usiku katika hospitali ya *taifa *muhimbili kwa ugonjwa wa kiharusi.
fariki: *hezron *mhela <kufariki> muda mfupi kabla_ya uchaguzi.
fariki: *mtumishi wa umma na mwanasiasa wa siku nyingi nchini *mzee *brown *ngwilulupi (76) <amefariki> dunia.
fuatilia: *omari pia wamewaagiza wakaguzi wa kahawa wa bodi hio pia <kufuatilia> kwa karibu suala hilo na kutoa taarifa kwake mwisho wa mwezi.
fuatilia: *wiki moja kabla_ya siku kuu ya *krismasi, mwaka jana, walionekana baadhi ya viongozi wa serikali za vijiji katika wilaya ya *rombo, *moshi na *hai <wakifuatilia> ushuru huo kwenye makampuni hayo bila mafanikio.
hatua: *alisema uamuzi wa serikali wa kununua umeme kutoa nchini *zambia ni <hatua> thabiti kwani inaonekana ni utekelezaji wa dira ya taifa ya mpango wa kuinua uchumi wa *taifa.
hatua: *omari alisema kuwa, ifikapo mwishoni mwa mwezi huu, kama makampuni hayo yatashindwa kulipa ushuru <hatua> za kisheria zitachukuliwa dhidi_yao kwa_mujibu_wa kanuni na sheria za ununuzi wa kahawa kutoka_kwa wakulima chini_ya mfumo wa soko huru.
hatua: *taarifa hiyo ilisema <hatua> hiyo inatokana na ukweli kwamba ujenzi wa makazi ya watu katika eneo hilo hauruhusiwi na ni kinyume cha sheria.
hatua: *wamelalamika kuwa ujenzi wa nyumba zao ulitokana na hali ngumu ya kubana matumizi kutokana_na kipato kidogo wanachokipata lakini *mkurugenzi huyo amefikia <hatua> ya kutoa agizo lenye athari kubwa kwao na familia zao.
ingia: "*yatakuwa maandamano ya amani, lakini kwa kadri tunavyowajua polisi wetu <watatuingilia> kwa lengo la kuvuruga amani ... wakija na magari yao msiwakimbie na muwe imara kukabiliana nao", alisema *profesa *lipumba alipokuwa akiwahutubia wanachama wa chama hicho katika ukumbi wa *diamond *jubilee, *dar_es_*salaam jana.
ingia: *aidha, baada_ya kustaafu shughuli za utumishi, *mzee *ngwilulupi <aliingia> kwenye siasa, ambapo alikuwa miongoni_mwa watu waliopigania mfumo wa

vyama vingi nchini na kufanikiwa.

ingia: *hata_hivyo, wakazi hao wamemuomba *rais *benjamin *mkapa <aingilie> katika hatua hiyo kwa madai kuwa ni ya uonevu.

ingia: *mwenyekiti wa *chama cha *wananchi (*cuf) *profesa *ibrahim *lipumba, amewahimiza wafuasi wa chama hicho kujitokeza kwa wingi kwenye maandamano yaliyopangwa kufanyika nchi nzima *jumamosi ijayo na kwamba wawe imara kukabiliana na polisi pindi <watakapoingilia> maandamano hayo.

8. Conclusion

After a fairly long period of research and testing, computational lexicography has reached a stage where computers and corpora can be put into effective use. For many years, computers have been used for producing word lists with frequencies from a corpus, as well as for retrieving concordances of word use. This article has shown that the use of regular expressions can significantly increase the precision and recall of search. However, the inclusion of the full linguistic analysis in dictionary work brings the work to a level where precision and recall meet high standards. SALAMA, the working environment developed for Swahili, facilitates the testing of various phases in dictionary compilation based on extensive use of the computer. This article demonstrates that computer-based lexicography does not only greatly benefit from the described approach; it is in fact a necessity in working with highly inflectional left-branching languages.

The system brings the automation of dictionary compilation to the point where the benefits of further automation become questionable. It accurately describes what can safely be described, and leaves ambiguous cases for human checking. Its great advantages are morphological accuracy and coverage, great speed, and ease of use.

The system can be developed still further, especially in the area of semantic disambiguation, so that correct senses of words in each context can also automatically be defined. Research is currently concentrating on the problems in this area.

Endnotes

1. There were also more realistic opinions that reflected the contemporary state-of-the-art in this field (Calzolari 1989; Wegera and Berg 1989).
2. By linguistic insight we here mean a kind of simulation of linguistic regularities, which a computer system utilizes and translates as 'linguistic rules'.
3. There has been discussion on the need of sufficient and systematic grammatical information in dictionaries (Salerno 1999). The approach discussed in this article effectively facilitates the inclusion of this feature.
4. The need of semantic information in dictionaries has increasingly been emphasized, whether

in terms of frame semantics (Fontenelle 2000, 2000a) or in terms of some other semantic theory. Statistical methods have also been used for identifying such word clusters that seem to occur together. On the basis of such clusters it is possible to carry out cluster analysis (Watters 2002).

5. In SALAMA, the Swahili Language Manager, etymological information on words of non-Bantu origin has been included by means of specific tags (Hurskainen 1999).
6. SALAMA is based on two-level morphology, and it is implemented by using finate state automata (Koskenniemi 1983; Hurskainen 1992, 1999). The disambiguation is based on the Constraint Grammar formalism (Karlsson 1995; Tapanainen 1996; Hurskainen 1996).
7. In fact, according to a survey, the choice of headwords was considered the most difficult among the 13 tasks asked from the team working on the third edition of the *Longman Dictionary of Contemporary English* (Kilgarriff 1998).
8. Based on SALAMA, the Swahili Language Manager, Sewangi (2000) has developed a system that retrieves term candidates from domain-specific text. This method facilitates the extensive use of domain-specific texts, such as educational books, handbooks, and other written materials of the domain, for compiling domain-specific dictionaries.
9. These two dictionaries are *Kamusi ya Kiswahili Sanifu* (1981) and *Kamusi ya Kiswahili–Kiingereza* (2001), both produced by the Institute for Kiswahili Research, University of Dar es Salaam.
10. The strings we wanted to find are shown with √.
11. Alternative strings are separated with a vertical bar and all alternatives are enclosed in parentheses. The question mark (?) stands for optionality, and the plus sign (+) means that the preceding unit may occur one or more times. The set a–z within square brackets means any character. The backslash (\) in the end of the line signifies that for the computer the same line continues.
12. The accuracy of the semantic glosses depends on how they were acquired in the analysis system. The most obvious way not requiring too much manual work is to use an electronic version of a good normal dictionary and include relevant parts of its entries in the dictionary of the analysis system. This was done in SALAMA, and the glosses produced are largely the same as those in the original dictionary, for good and bad. We should not, however, be content with these glosses, because they are just approximations of the various meanings of the lexemes and they should be checked and amended on the basis of the information available in the corpus. In addition to helping in the selection of headwords, the corpus is useful in identifying various meanings of the lexemes.

References

Calzolari, N. 1989. Computer-Aided Lexicography: Dictionaries and Word Data Bases. Steger, H. and H.E. Wiegand (Eds.). *Handbuch zur Sprach- und Kommunikationswissenschaft, Band 4*: 510-519. Berlin: Walter de Gruyter.

Fontenelle, T. 2000. A Bilingual Lexical Database for Frame Semantics. *International Journal of Lexicography* 13(4): 232-248.

Fontenelle, T. 2000a. Introduction: Dictionaries, Thesauri and Lexical-semantic Relations. *International Journal of Lexicography* 13(4): 229-231.

Hurskainen, A. 1992. A Two-Level Computer Formalism for the Analysis of Bantu Morphology: An Application to Swahili. *Nordic Journal of African Studies* 1(1): 87-122.

Hurskainen, A. 1994. Kamusi ya Kiswahili Sanifu in Test: A Computer System for Analyzing Dictionaries and for Retrieving Lexical Data. *Afrikanistische Arbeitspapiere* 37 (Swahili Forum I): 169-179.

Hurskainen, A. 1996. Disambiguation of Morphological Analysis in Bantu Languages. *COLING-96, Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, August 5–9, 1996*: 568-573. Copenhagen: Center for Language Technology.

Hurskainen, A. 1999. SALAMA: Swahili Language Manager. *Nordic Journal of African Studies* 8(2): 139-157.

Hurskainen, A. 2002. Tathmini ya Kamusi Tano za Kiswahili. *Nordic Journal of African Studies* 11(2): 283-301.

Hurskainen, A. and R. Halme. 2001. Mapping between Disjoining and Conjoining Writing Systems in Bantu Languages: Implementation on Kwanyama. *Nordic Journal of African Studies* 10(3): 399-414.

Jones, R.L. and S.P. Sondrup. 1989. Computer-Aided Lexicography: Indexes and Concordances. Steger, H. and H.E. Wiegand (Eds.). *Handbuch zur Sprach- und Kommunikationswissenschaft, Band 4*: 490-509. Berlin: Walter de Gruyter.

Karlsson, F. 1995. Designing a Parser for Unrestricted Text. Karlsson, F. et al. (Eds.). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*: 1-40. Berlin: Mouton de Gruyter.

Kilgarriff, A. 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10(2): 135-155.

Kilgarriff, A. 1998. The Hard Parts of Lexicography. *International Journal of Lexicography* 11(1): 51-54.

Koskenniemi, K. 1983. *Two-level Morphology: A General Computational Model for Word-form Recognition and Production*. Publications No. 11. Helsinki: Department of General Linguistics, University of Helsinki.

Panyr, J. and H.H. Zimmermann. 1989. Information Retrieval: Überblick über active Systeme und Entwicklungstendenzen. Steger, H. and H.E. Wiegand (Eds.). *Handbuch zur Sprach- und Kommunikationswissenschaft, Band 4*: 696-708. Berlin: Walter de Gruyter.

Salerno, L. 1999. Grammatical Information in the Bilingual Dictionary: A Study of Five Italian–French Dictionaries. *International Journal of Lexicography* 12(3): 209-222.

Sewangi, S. 2000. Tapping the Neglected Resource in Kiswahili Terminology: Automatic Compilation of the Domain-specific Terms from Corpus. *Nordic Journal of African Studies* 9(2): 60-84.

Tapanainen, P. 1996. *The Constraint Grammar Parser CG-2*. Publications No. 27. Helsinki: Department of General Linguistics, University of Helsinki.

Tapanainen, P. and T. Järvinen. 1998. Dependency Concordances. *International Journal of Lexicography* 11(3): 187-203.

Teubert, W. 2001. Corpus Linguistics and Lexicography. *International Journal of Corpus Linguistics* 6 (Special Issue): 125-153.

Watters, P.A. 2002. Discriminating English Word Senses Using Cluster Analysis. *Journal of Quantitative Linguistics* 9(1): 77-86.

Wegera, K.-P. and E. Berg. 1989. Computergestützte Grammatikographie: Eine Fallstudie. Steger,

H. and H.E. Wiegand (Eds.). *Handbuch zur Sprach- und Kommunikationswissenschaft, Band 4*: 519-527. Berlin: Walter de Gruyter.

Dictionaries

Abdulla, A., R. Halme, L. Harjula and M. Pesari-Pajunen (Eds.). 2002. *Swahili–Suomi–Swahili-sanakirja.* Helsinki: Suomalaisen Kirjallisuuden Seura.

***Kamusi ya Kiswahili–Kiingereza* (Swahili–English Dictionary).** 2001. Dar es Salaam: Taasisi ya Uchunguzi wa Kiswahili.

Kamusi ya Kiswahili Sanifu. 1981. Dar es Salaam: Oxford University Press.