Arvi Hurskainen
Institute for Asian and African Studies
University of Helsinki

Computational testing of five Swahili dictionaries

Abstract

This paper introduces a computational method for testing dictionaries. It discusses the implementation of this method on testing five current dictionaries of Swahili and provides a number of test results. The tested dictionaries are Kamusi ya Kiswahili Sanifu (TUKI), Kamusi ya Maana na Matumizi (OUP), Modern Swahili - Modern English Dictionary (MS-tryck), Kamusi ya Kiswahili - Kiingereza (TUKI), and Swahili - Suomi - Swahili -sanakirja (SKS). Each of the dictionaries was tested by using a dictionary-specific version of SWATWOL, a two-level parser of Swahili. The recall of each dictionary was tested by using three test corpora. Also, the proportion of unused words in each dictionary was tested. Furthermore, the performance of each dictionary in some word classes was tested. The results of tests are summarized in tables and graphs.

Keywords: lexicography, evaluation, computational testing, finite-state methods

1. Problem

Dictionaries commonly have two major defects. The more serious of these is that the dictionary does not contain all those words that the user would need. Often the necessary information on the use of the word is either missing or defectively given. The less serious problem is that the dictionary has words that never appear in the kind of text for which the dictionary is intended. In brief, the dictionary does not match the target language.

Most dictionary users have faced both of these problems, but there is very little that they can do to improve the situation. Even for the dictionary compiler, manual checking of the accuracy of the dictionary is difficult and so time-consuming that it is rarely done. As a result we have nice-looking dictionaries, which, however, harbour serious defects.

The advent of computational language analysis programs has fundamentally changed the situation, because now it is possible to construct precise methods for testing dictionaries in several respects. It is also important that vast masses of text of various genres can be used in tests.

2. Method

The basic module in computational dictionary testing is the morphological parser that is able to identify all correct word-forms in text. The parser is made to mirror the dictionary to be tested, for better or for worse. In practice it means that a comprehensive 'good' computational parser is 'worsened' in those points where the dictionary to be tested is bad. Only those key words are allowed in the parser that are in the dictionary to be tested. If some morphological information is missing in the dictionary to be tested, it should also be missing in the parser. With this method we can simulate the process that the human dictionary user goes through when trying to find words and information on their use.

3. Dictionaries to be tested[1]

Five general-purpose Swahili dictionaries were tested. Four of those were Swahili-English bilingual dictionaries, and one was a monolingual dictionary.

Kamusi ya Kiswahili Sanifu, Taasisi ya Uchunguzi wa Kiswahili, Dar-es-Salaam, 1981. It claims to have 20,000 headwords and 50,000 words, but in reality it has only 14,288 dictionary entries. This dictionary, to be referred to below as KKS, was compiled by the Institute of Kiswahili Research, University of Dar-es-Salaam, to be a then modern dictionary of Standard Swahili.

Kamusi ya Maana na Matumizi, Oxford University Press, Nairobi, 1992. It claims to have 9,000 headwords and 12,000 example sentences. In reality it has 8,057 headwords. The primary purpose of this dictionary, prepared by Salim K. Bakhressa, is to give examples of use for selected Swahili words. It is not intended to be comprehensive, and, as the test results below will show, its recall is not good in any test.

Swahili-English Dictionary, MS-tryk, 1991. It claims to have 13,300 headwords and 3,000 examples of use. The real number of headwords is 10,461. This dictionary was compiled by Gerald Feeley, a Swahili teacher in the Danish Volunteers' Training Centre, Tanzania.

Kamusi ya Kiswahili - Kiingereza, Taasisi ya Uchunguzi wa Kiswahili, Dar-es-Salaam, 2001. It claims to have more than 30,000 headwords, but in reality the number is only 14,533. As in the case of KKS, this dictionary was also compiled at the Institute of Kiswahili Research, and it is intended to be a standard dictionary of current Swahili.

Suomi - Swahili - Suomi -sanakirja, Suomalaisen Kirjallisuuden Seura, 2002. It claims to have about 10,000 headwords, but in reality it has 11,500. In compiling this dictionary, a frequency list based on a corpus of about one million words was used for selecting the main part of headwords.

4. What was tested?

In tests, we wanted to get illumination on the following aspects of the dictionaries:

Terminological coverage: The major aim of the test was to see how well each dictionary covered the terminology found in three text corpora (to be described in more detail below).

Accuracy of linguistic information: We also wanted to see how well the dictionary helped in finding the correct forms of words. Such features include plural forms, noun classes, concordance, and inflection of adjectives and numerals.

Surplus: We also wanted to test to what extent the dictionaries contained words that are not found in texts.

Coverage in various word categories: The proportion of various categories of words, such as verbs and nouns of different noun classes, was also a subject of the study.

Comparison of dictionaries: We wanted to determine what differences there were in performance between dictionaries.

Influence of text type: Finally, we wanted to investigate what kind of effect the type of corpus had on the performance of each dictionary.

[1] Our aim in this computational testing is not to rank the dictionaries in any way. We wish to point out distinctive features of dictionaries in general and investigate how well these features are represented in each of the dictionaries.

5. Test corpora

For carrying out the research, three text corpora were constructed. Corpus 1 contains recent news texts from the years 1998-2001, and its size is 2,484,852 words. Corpus 2 consists of fiction, drama and scientific texts plus some older news texts from the years 1988-1994, and it has 1,190,489 words. Corpus 3 is a smaller collection of recent news texts (2002) and fiction with 552,021 words.

6. How was testing implemented?

First, SWATWOL[2], the Swahili morphological two-level parser, was adapted to 'mirror' each dictionary. As a result we had five more or less defective parsers. In constructing these parsers, the following principles were followed:

(1) Only those words were included that were found as entries in dictionaries.

(2) Only that information was encoded in the parser that was available in the dictionary, e.g. missing morphological information in the dictionary was mirrored in the parser.

(3) SWATWOL, a comprehensive Swahili parser, was used for extracting the lemma form of each word-form in the corpus. Because SWATWOL is comprehensive, practically every Swahili word was linguistically analyzed and a lemma form was found.

(4) The calculation and comparison was made on the basis of lemma forms of words rather than on the basis of actual word-forms in the text.

7. Limitations of the test

The test had some limitations. For example, definitions of terms and glosses in the target language were not tested. Also, subentries of headwords were not taken into account if they consisted of more than one word. Multi-word concepts in general were excluded and homonyms were not differentiated.

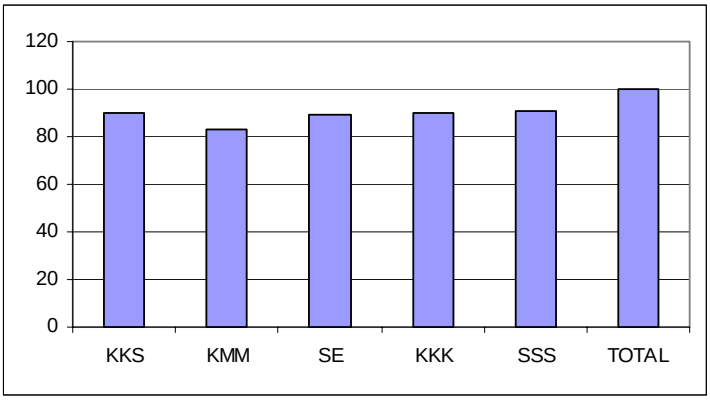
8. Identified word-forms

In Table 1 we see how well each of the five dictionaries covered the word-forms found in Corpus 1.

Table1. Identified word-forms in Corpus 1 (%).

KKS	90.08
KMM	82.73
SE	89.52
KKK	90.04
SSS	90.70
TOTAL	100.00

[2] Description of two-level morphology in general can be found in Koskenniemi (1983) and its application to Swahili in Hurskainen (1992). This testing method was initially used for testing KKS (Hurskainen 1994a), and it has since improved considerably and become more accurate and detailed (Hurskainen 2002). However, it is not possible to disambiguate (Karlsson 1995, Tapanainen 1996) the text, because the 'worsened' parsers are not good enough for successful disambiguation.



The total number of unique word-forms in Corpus 1 was 111,342. The percentages were calculated on the basis of unique occurrences of word-forms. Table 1 shows that the recall of the dictionaries was between 82.73% and 90.08%. In fact all the four more or less similar dictionaries have a surprisingly similar recall. The differences in performance do not reflect the differences in dictionary size, so that larger dictionaries would have a better coverage than the smaller ones. The results show that smaller dictionaries, such as e.g. SE, have managed to include those words that are actually used in media. On the other hand, none of the dictionaries is sufficiently good, because a general-purpose dictionary should contain more than 90% of words used in media.

The recall of dictionaries was calculated also on the basis of word-form tokens, so that all occurrences of each word-form were taken into account. The results were almost identical with those in Table 1.

9. Recall of lemma-forms in Corpus 1

The base forms of words are even more appropriate than word-forms in evaluating the performance of dictionaries. This is particularly true of languages where a word may appear in several different forms. It is the base form that we look for in the dictionary, and not one of its inflected forms. SWATWOL provides us with this capability, and below we shall deal with base forms only.

It is also important to know how frequent the missing words are. It is a different thing to miss a commonly used word and a very rare word in the dictionary. Therefore, we have done also some frequency calculations of missing words in each of the tested dictionaries.

Table 2. Words missing in Corpus 1, based on unique lemma-forms.

	All	2 or more	5 or more	10 or more
KKS	2,223	1,505	911	611
KMM	4,792	3,164	1,952	1,324
SE	3,029	1,828	968	596
KKK	1,999	1,278	704	467
SSS	1,940	1,047	463	264

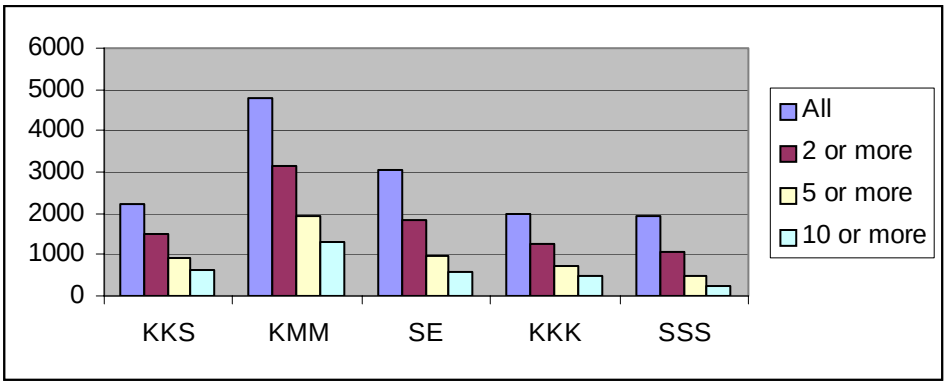


Table 2 shows how fully various dictionaries have listed words with different frequencies. The lemma-forms found in Corpus 1 were divided into four groups according to the frequency. The leftmost column shows the total number of missing lemmas, which vary between 1,940 and 4,792.[3] We see that when all missing words are taken into account KKK and SSS are the best dictionaries. The number of missing lemmas is 1,999 and 1,940 respectively. KKS is more defective (2,222), but not as bad as KMM (4,792) and SE (3,029).

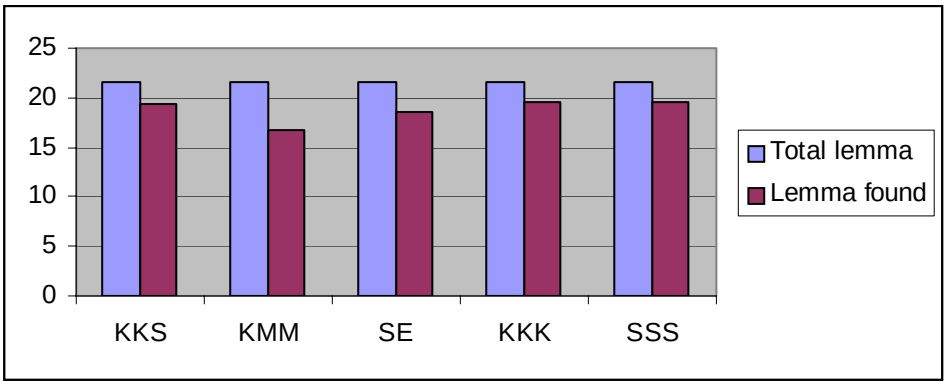
When we look at the statistics in various frequency classes, we get a somewhat different picture. KMM and SE performed badly also in the group of lemmas that occurred at least twice in Corpus 1. The situation was similar in the group of five or more occurrences. In the group of highest frequency, the one with at least ten occurrences, KKS, SE, and KKK were close to each other. SSS was the best dictionary in each frequency group, but especially in the group of the highest frequency. This can be explained by the fact that entries for the dictionary were selected on the basis of a frequency list based on current Swahili texts.

Table 3 shows the performance of each dictionary as a percentage from all lemmas in Corpus 1. No frequency information is included in the table. KKK and SSS are the best dictionaries, and KKS comes very close behind. The worst is KMM, which is not actually directly comparable with the others, because its purpose is to give examples of use for selected words (Hurskainen 1999).

Table 3. The performance of each dictionary in finding lemmas in Corpus1.

	Total lemma	Lemma found	Recall efficiency
KKS	21,509	19,286	89.7
KMM	21,509	16,717	77.7
SE	21,509	18,480	85.9
KKK	21,509	19,510	90.7
SSS	21,509	19,569	91.0

[3] The total number of lemmas found in Corpus 1 was 21,509. This fairly high number vastly exceeding the number of entries in any of the dictionaries, is explained by the fact that, because there was no possibility of using a disambiguator (Hurskainen 1996, Hurskainen 2004) due to the 'badness' of the morphological parsers, many lemmas have alternative interpretations and all of them are in the final count. This does not, however, cause a bias in comparison, because it affects each dictionary in a similar way.

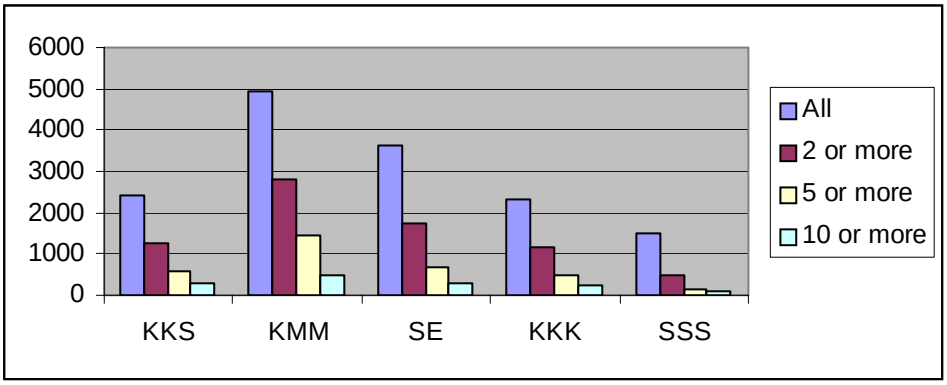


10. Performance with Corpus 2

Corpus 2 has two kinds of texts. About 80% of it consists of fiction books and texts from literature research. The rest of the corpus is older newspaper texts from 1988-1995.

Table 4. Missing lemmas in dictionaries compared with Corpus 2.

	All	More than 1	More than 4	More than 9
KKS	2,426	1,256	567	309
KMM	4,924	2,798	1,442	482
SE	3,647	1,758	694	296
KKK	2,311	1,146	503	258
SSS	1,512	490	146	88

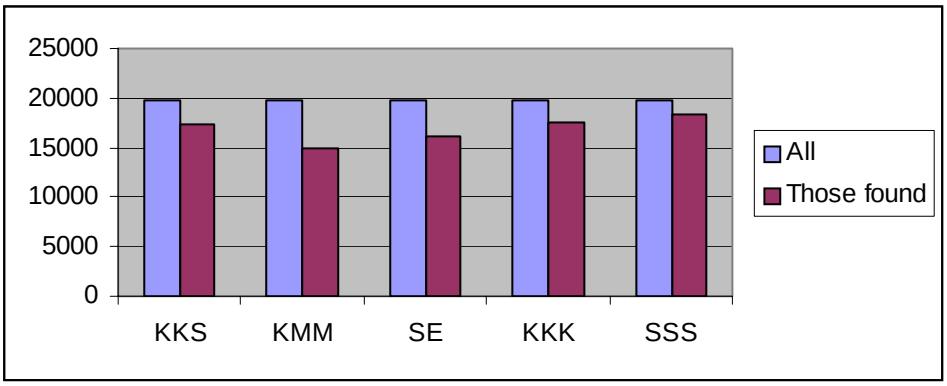


In Table 4 we see that KKS and KKK were almost identical in all frequency groups. As expected, KMM was the worst, followed by SE. SSS performed exceedingly well, especially in the groups of more frequent words. With Corpus 2 SSS performed clearly better than with Corpus 1, although such differences were not found in the performance of other dictionaries. The good performance of SSS with Corpus 2 can be explained by the fact that a large part of Corpus 2 was also part of the corpus which was used for preparing the frequency list for compiling SSS. It should be noted that the performance was good also in the groups of less frequent words, although not as good as in the groups of more frequent words. The result is precisely as expected, because in compiling SSS, the 10,000 most common lemmas were chosen.

In Table 5 the performance of each dictionary is compared with 'the ideal dictionary', i.e. the fictional dictionary that contains all words in Corpus 2.

Table 5. The performance of each dictionary in identifying lemmas in Corpus2.

	All	Those found	%
KKS	19,799	17,379	87.8
KMM	19,799	14,875	75.1
SE	19,799	16,152	81.6
KKK	19,799	17,488	88.3
SSS	19,799	18,287	92.4

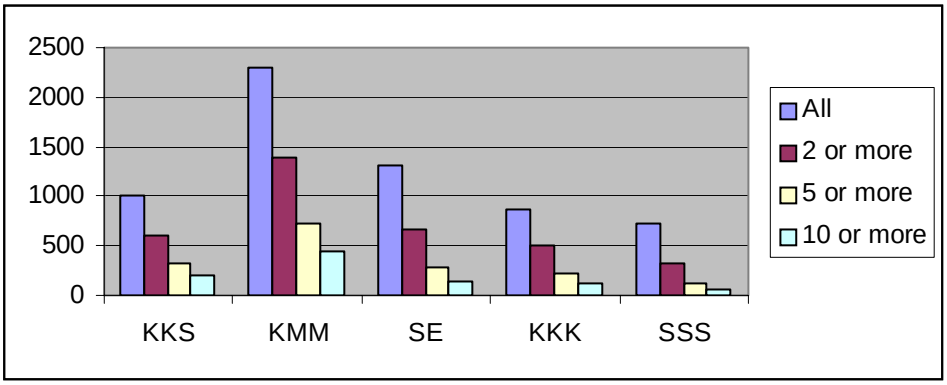


11. Results with Corpus 3

Corpus 3 contains news texts from the year 2002, and it is smaller than the other two corpora. This corpus was included into the research for testing whether the rapidly developing language has an effect on the performance. However, because of its considerably smaller size it is not directly comparable with the other two corpora. This should be remembered when interpreting the results.

Table 6. Missing lemmas in dictionaries compared with Korpus 3.

	All	More than 1	More than 4	More than 9
KKS	1,001	605	325	193
KMM	2,304	1,387	731	449
SE	1,309	665	287	136
KKK	871	494	229	125
SSS	731	326	128	64



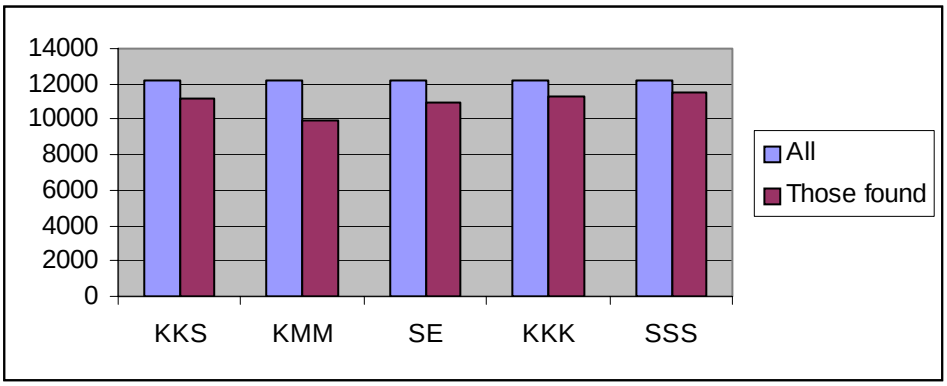
Again in this test SSS was the best dictionary in all four frequency groups. Its good performance is discernible particularly in the groups of high frequency. Corpus 3 represents modern news text, and SSS seems to cover it well. The second was KKK, but clearly behind

SSS. A small surprise is that SE was better than KKS in frequent words, although the latter is considered a kind of standard dictionary of Swahili. One explanation may be that KKS is about 20 years older than SE, and for this reason it does not have all modern terms. On the other hand, KKS is better than SE in the groups of less frequent words.

The number of missing words here is smaller than in the two earlier tests, because the small corpus has fewer words, and consequently also fewer words were missing from the dictionaries. Also, the news texts have a fairly limited vocabulary, and many themes recur in different news media and articles. The percentages of performance in Table 7 illuminate the performance of each dictionary.

Table 7. The performance of each dictionary in identifying lemmas in Corpus3.

	All	Those found	%
KKS	12,209	11,208	91.8
KMM	12,209	9,905	81.1
SE	12,209	10,900	89.3
KKK	12,209	11,338	92.9
SSS	12,209	11,478	94.0



In Corpus 3, four of the dictionaries have a performance rate of roughly 90% or better. Only KMM is clearly behind with 81.1%. As a whole the performance of all dictionaries is the best in this corpus. The corpus contains common news texts without vocabulary of specific domains or otherwise rare words.

12. Unused words in dictionaries

Above we have discussed the question of how well the dictionaries covered the vocabulary used in three test corpora. That research did not reveal anything about otiose words in the dictionaries, i.e. words that did not occur in corpus texts at all. Here we shall look into this question by investigating to what extent the dictionaries had such 'extra' words in different word categories.

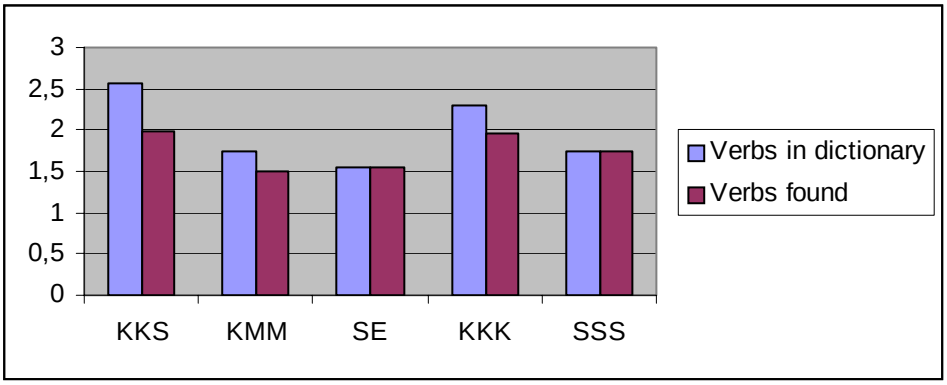
12.1. Unused verbs

In this test, SWATWOL was used in the mode where it returns the base form from all verb forms, including extended verbs. Only lexicalized extended verbs with a clearly distinct meaning were treated as separate verbs. In this way we tried to imitate the dictionary entries as closely as possible.

Table 8 shows how many verbs each dictionary identified in Corpus 1 and how many it listed in the book. The difference between these two shows the number of unused verbs.

Table 8. Verbs listed in dictionaries and found in Corpus 1.

	Verbs in dictionary	Verbs found	Efficiency index
KKS	2,560	1,991	77.77
KMM	1,742	1,500	96.11
SE	1,557	1,540	98.91
KKK	2,305	1,948	84.51
SSS	1,751	1,748	99.83



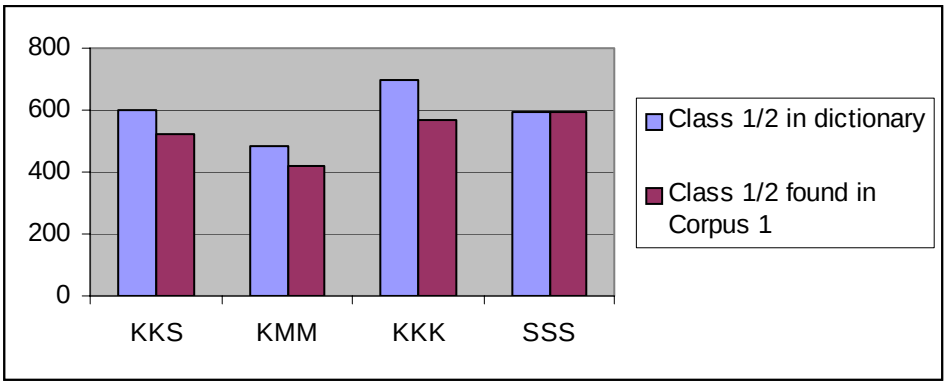
We see that KKS had the best coverage, closely followed by KKK. On the other hand, KKS also listed more than other dictionaries verbs that were not used in Corpus 1. In KKK a fairly large number of unused verbs were found, although not to such a large extent as in KKS. SSS scored the third in coverage, and what is interesting is that almost all the verbs listed in it were also used in the corpus. SE had the smallest number of verbs, but almost all verbs listed in it were also used in Corpus 1. KMM had almost the same number of verbs as SSS, but they did not match well with the requirements of the corpus.

12.2. Unused nouns

The use of nouns was tested according to noun class categories. Because SE did not show clearly the noun class affiliation of each noun, it had to be excluded from this test. Class 11 was left out, because it was not possible to formulate a reliable test program because of the poor marking of this class in some of the dictionaries. Therefore, Class 11 is excluded also in the summary comparison below.

Table 9. Class 1/2 in dictionary and found in Corpus 1.

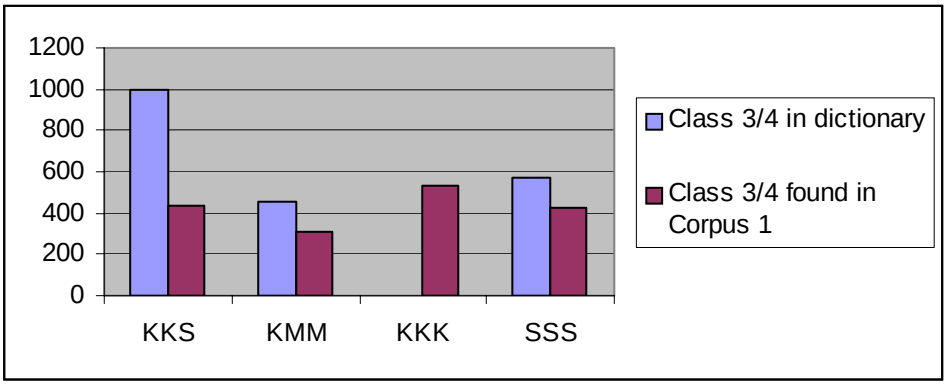
	Class 1/2 in dictionary	Class 1/2 found in Corpus 1	Efficiency index
KKS	597	520	87.10
KMM	483	417	86.34
KKK	696	568	81.61
SSS	595	593	99.66



We see that SSS has the best recall and practically all the words of this class are also found in Corpus 1. Because this is the class of human beings, and because news texts normally handle activities of people, it is no wonder that a fairly high percentage of nouns of this class listed in the dictionary is also used in texts. The poorest match is in KKK, although it is considered a standard dictionary of modern Swahili. KKS lists fewer nouns of this class and also its match is fairly poor. KMM has the smallest number of nouns, and yet many of them are not used in texts.

Table 10. Class 3/4 in dictionary and found in Corpus 1.

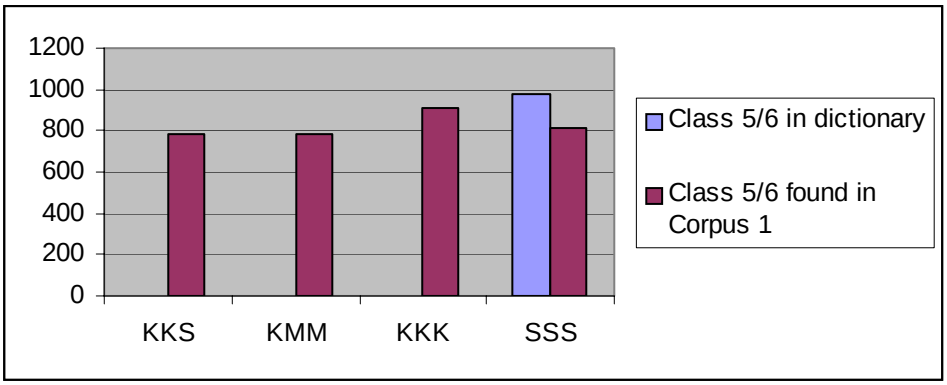
	Class 3/4 in dictionary	Class 3/4 found in Corpus 1	Efficiency index
KKS	997	439	44.03
KMM	455	306	67.25
KKK	1,085	529	48.76
SSS	572	427	74.65



In Noun Class 3/4, KKK has the best recall, but it has also a large number of nouns that never occurred in Corpus 1. KKS is the second in recall, but it also has many unused nouns of this class. The match on SSS is comparatively the best, because it has a reasonably good recall but just a moderate number of unused nouns. Here again we see that KMM does not score well.

Table 11. Class 5/6 in dictionary and found in Corpus 1.

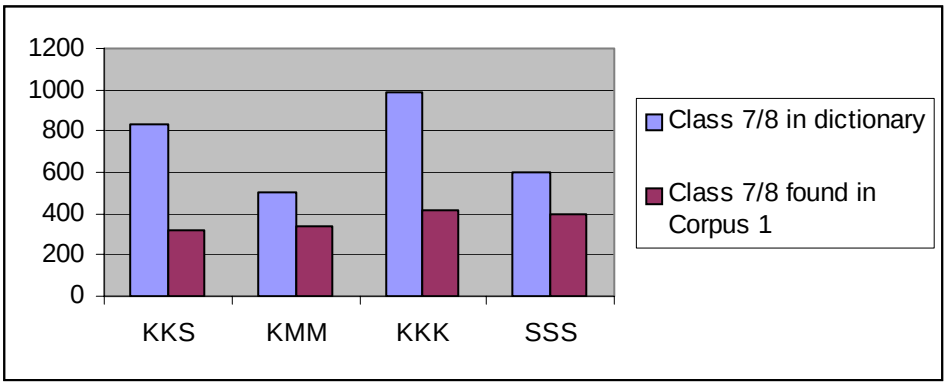
	Class 5/6 in dictionary	Class 5/6 found in Corpus 1	Efficiency index
KKS	1,584	788	49.75
KMM	1,020	786	77.06
KKK	1,319	914	69.29
SSS	981	815	83.08



It is a bit surprising that in Class 5/6 KKS, although it has listed the largest number of nouns, has almost the poorest match. KKK has the best recall, but it also has a fairly large number of unused nouns. Here again, SSS has the best match. Its recall is the second best, but it has a small number of unused nouns of this class.

Table 12. Class 7/8 in dictionary and found in Corpus 1.

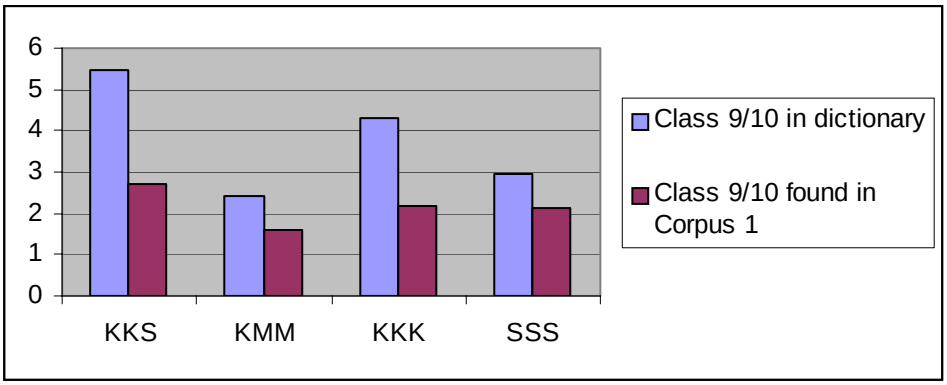
	Class 7/8 in dictionary	Class 7/8 found in Corpus 1	Efficiency index
KKS	836	324	38.76
KMM	501	336	67.07
KKK	984	420	42.68
SSS	600	401	66.83



In Class 7/8 we see a big mismatch between what is listed in dictionaries and what is found in the corpus. KKK has a little better recall than KKS, but less than half of the nouns of this class are found in the corpus. Also, KKS has a lower recall than any other dictionary, including KMM, which usually does not do well in this research.

Table 13. Class 9/10 in dictionary and found in Corpus 1.

	Class 9/10 in dictionary	Class 9/10 found in Corpus 1	Efficiency index
KKS	5,448	2,713	49.80
KMM	2,441	1,576	64.56
KKK	4,295	2,192	51.04
SSS	2,940	2,106	71.63



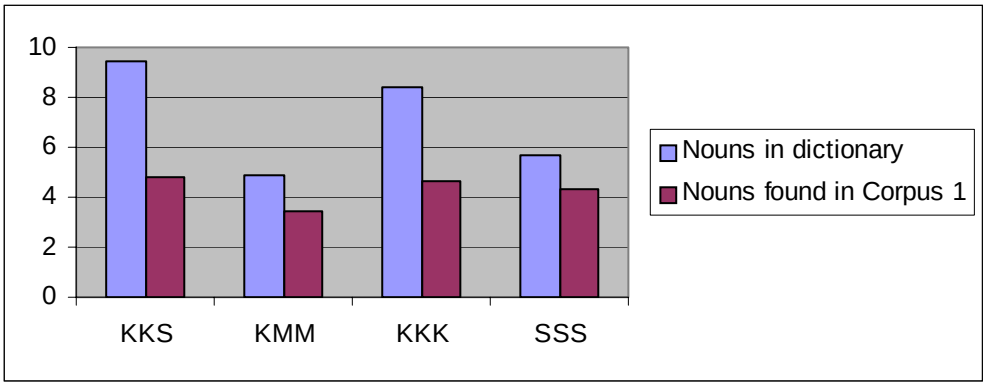
In Class 9/10 about half the words listed in KKS and KKK were not found in Corpus 1. Especially KKS has a large number of such words. This is significant, because Class 9/10 comprises about 39% of all Swahili nouns (Hurskainen 1994b). KMM had a fairly good match but a poor recall. SSS is the best also in this group, because its match and recall are fairly good.

12.3. Summary comparison of Noun Classes 1/2, 3/4, 5/6, 7/8, and 9/10

We see in Table 14 how the various dictionaries performed with nouns in general. Only the classes discussed above are included, and Class 11 has been excluded due to difficulties in constructing a reliable test of this class.

Table 14. Nouns in dictionary and found in Corpus 1.

	Nouns in dictionary	Nouns found in Corpus 1	Efficiency index
KKS	9,462	4,784	50.56
KMM	4,900	3,421	69.82
KKK	8,379	4,623	55.17
SSS	5,688	4,342	76.34



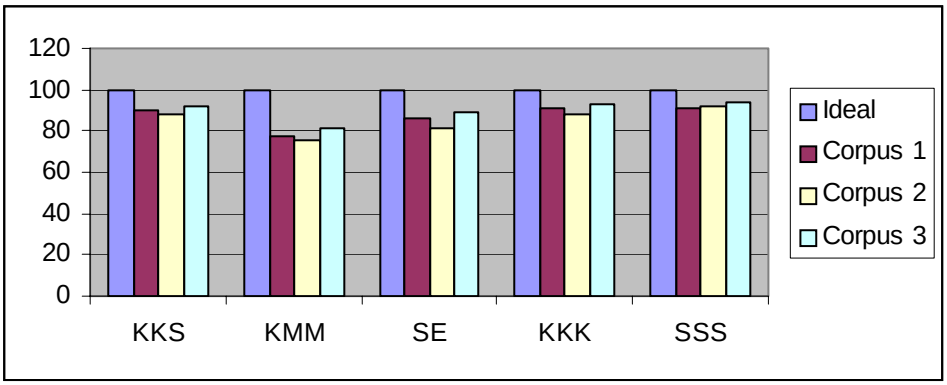
When all noun classes are taken into account, KKS had the best recall, followed by KKK, SSS, and KMM, in this order. The overwhelmingly biggest surpluses of nouns were in KKS and KKK. SSS had the best match, so that the recall was fairly good and the number of unused nouns was quite small.

13. Performance of dictionaries with all three corpora

Table 15 shows how each of the dictionaries performed with each of the three test corpora. Only recall is shown, and the results are shown as a percentage of the total.

Table 15. Summary of the performance of dictionaries with test corpora.

	Ideal	Corpus 1	Corpus 2	Corpus 3
KKS	100	89.7	87.8	91.8
KMM	100	77.7	75.1	81.1
SE	100	85.9	81.6	89.3
KKK	100	90.7	88.3	92.9
SSS	100	91.0	92.4	94.0



In Table 15 all word categories are included. Calculations were made on the basis of lemmas and comparison was made with the 'real' facts. These facts, i.e. the real number of lemmas in each corpus, were obtained with SWATWOL for each corpus. This figure was then used for calculating the recall of each dictionary from each corpus.

Table 15 shows that, except for SSS, recall in Corpus 2 was lower than with other two corpora. The high recall of SSS with Corpus 2 is largely due to the fact that part of its texts were used in compiling SSS. All dictionaries performed better with Corpus 3 than with the other two corpora. The difference with Corpus 1, however, is fairly small. Corpus 1 and 3 contained news texts, and although they are from different periods, the ratio of the recall rates of the two corpora for the various dictionaries was relatively constant.

14. Summary

The tests made with three corpora show that all five dictionaries were deficient in recall, although significant differences were found. Because of its special nature, KMM was the most deficient. The size of a dictionary did not guarantee high recall. KKS and KKK had the largest number of headwords, but they were not the best in recall. Among the four dictionaries, when KMM is excluded, SSS with the smallest number of headwords had the best recall. And as could be expected, the largest dictionaries had the largest number of unused headwords.Tests also indicate clearly that if a dictionary is compiled by using an appropriate frequency list of possible dictionary entries, it is likely to be more efficient than dictionaries compiled using traditional methods.

The method discussed above is suitable for testing dictionaries in languages where part-of-speech ambiguity on the word level is not substantial. Also, dictionaries with mainly single-word entries are generally more suitable than those with many multi-word entries.

However, if the parser is constructed to handle multi-word entries also, as is increasingly the case, this limitation is not significant.

References

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.

1994a *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.

1994b *Quantitative analysis of Swahili noun classes*. Working Papers in Linguistics 21, Department of Linguistics. University of Trondheim. Pp. 1-16.

1996 Disambiguation of morphological analysis in Bantu languages. In *COLING-96, Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, August 5-9, 1996. Pp. 568-573.

1999 *Salim K. Bakhressa, Kamusi ya Maana na Matumizi*. Nairobi: Oxford University Press. Book review. Journal of African Languages and Linguistics 20. Book review.

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

2004 Optimizing disambiguation in Swahili. *COLING-04, Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, August 23-27, 2004. Pp. 254-260.

Karlsson, F. 1995.
Designing a parser for unrestricted text. In Karlsson et al (eds.), *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin. Pp. 1-40.

Koskenniemi. K. 1983
Two-level morphology: A general computational model for word-form recognition and production. Publications No. 11. Department of General Linguistics, University of Helsinki.

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

Dictionaries tested

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

Bakhressa, Salim K. 1992.
Kamusi ya Maana na Matumizi. Nairobi: Oxford University Press.

Feeley, Gerald, 1994.
Modern Swahili - Modern English Dictionary (Revised and enlarged second edition). Denmark: MS-tryk.

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.