AAP 37 (1994): 169-179

KAMUSI YA KISWAHILI SANIFU IN TEST: A COMPUTER SYSTEM FOR ANALYZING DICTIONARIES AND FOR RETRIEVING LEXICAL DATA

ARVI HURSKAINEN

The paper describes a computer system for testing the coherence and adequacy of dictionaries The system suits also well for retrieving lexical material in context from computerized text archives Results are presented from a series of tests made with Kamusi ya Kiswahili Sanifu (KKS), a monolingual Swahili dictionary. The test of the internal coherence of KKS shows that the text itself contains several hundreds of such words, for which there is no entry in the dictionary. Examples and frequency numbers of the most often occurring words are given. The adequacy of KKS was also tested with a corpus of nearly one million words, and it was found out that 1.32% of words in book texts were not recognized by KKS, and with newspaper texts the amount was 2.24%. The higher number in newspaper texts is partly due to numerous names occurring in news articles Some statistical results are given on frequencies of wordforms not recognized by KKS. The tests shows that although KKS covers the modern vocabulary quite well, there are several areas where the dictionary should be improved The internal coherence is far from satisfactory, and there are more than a thousand such rather common words in prose text which are not included into KKS The system described in this article is an effective tool for detecting problems and for retrieving lexical data in context for missing words

1. Introduction

Dictionary compilers have nowadays available a number of tools, which are of great help in searching for data and in arranging the material in a systematic way Various kinds of computer-based storing systems have been devised for making the management of information quick, comprehensive and reliable Less common are, however, devices for testing the internal coherence of dictionaries and their capability of covering certain types of texts.

In the following I shall describe a data searching system, which enables the automatic search of such lexical items in a running text, which are not included in a given dictionary Such a system is useful when there already exists a dictionary, but there is a need to test its adequacy. It is also helpful when there is a need to update the existing dictionary, because it enables one to create an ordered database, where words not included in the dictionary appear in their natural contexts. Such a data searching system will in fact replace the traditional databases compiled manually and written on cards.[1]

[1] Still a few years ago Tumbo-Massabo (1989) doubted the feasibility of computerized data banks in developing countries, due to the unavailability of suitable technology. There is no need to maintain such

The following description is based on tests made with Kamusi ya Kiswahili Sanifu (hence KKS), the monolingual Swahili dictionary, prepared by the Institute of Kiswahili Research, University of Dar-es-Salaam. It was first published in 1981, and although it is a useful dictionary in many respects, it has weaknesses, which should be corrected in future editions. The requirements and problems of monolingual dictionaries in general, and of KKS in particular, have been discussed elsewhere (Khamisi 1987; Chuwa 1987) I am not going to argue here on the structure of the dictionary itself, because there are more than one good way to compile a useful dictionary Rather I shall pinpoint the obvious inadequacies, which the first edition of this dictionary contains. Since KKS contains only single words as entries and does not pay attention to their etymology, I shall treat KKS strictly on the basis of this choice of the compilers. There would be an urgent need of other types of information also (Calzolari and Bindi 1990), which could be retrieved from a corpus with the tools available, but this has to be left to another context

The first requirement of a monolingual dictionary is that it should have those words as entries which are used in explaining the meanings of the words given as headwords, and that under each headword sufficient information is given concerning the forms which the headword takes. The computer system under discussion is able to detect such inadequacies, and below, after having described the structure of the system, I shall present test material of two kinds: (1) the written text of KKS will be analyzed and the discrepancies in it detected, and (2) selected modern Swahili prose texts will be filtered to show how well the dictionary covers the terminology of different kinds of texts, and how these words can be retrieved in context into a lexical databank for further use [2] The system is composed of the following set of programs, which are run consecutively:

1. Normalize the input text, e.g. a fiction book, newspaper text etc. so that upper case letters are converted to lower case, preserving the information of a capital letter where the word is always initiated with a capital letter; place a blanc around a word where preceded or followed by a diacritic or punctuation mark, etc. This can be accomplished by a rewriting program, e.g. a Beta program [3]

2. Make a list of all wordforms in a text, and delete all additional occurrences of the same wordform.

3. Filter out such wordforms, which are not included into the dictionary concerned. This phase requires the use of a morphological parser, which recognizes the wordforms of the dictionary only. See a more detailed description below.

---

doubts any more, although computers will probably never completely replace the traditional manual work in compiling dictionaries

[2] Part of the material from modern texts was retrieved by my students in 1992 when practicing the use and production of retrieving programs. I express my gratitude to them for their share in preparing this paper

[3] Beta is a programming language, which rewrites text according to user-made rules based on context restrictions and state mechanism. It was first written by Benny Brodda from the University of Stockholm. It has been further developed by him and Fred Karlsson and Kimmo Koskenniemi from the University of Helsinki, by using such programming languages as Fortran, Pascal, InterLisp, MuLisp, and C-language. See Brodda 1990

4. Find the resulting wordforms in context from the text used for filtering, and possibly sort them according to the keyword. Retrieving can be carried out e.g. with a Beta program, or any other suitable retrieving program.

For a language with a minimal amount of morphological variation, such as English, a wordform retrieving system could perhaps be built on the basis of a computerized lexicon with a straightforward list of lexical entries. For highly inflecting languages such as Swahili this is not possible. The accurate recognition system requires a powerful morphological parser, which builds up words by combining morphemes according to defined rules. In the system described here I have used SWATWOL, which has been described elsewhere in detail (Hurskainen 1992). SWATWOL is a morphophonological parser, which uses a language-independent parsing module TWOL (Koskenniemi 1983), and takes a rule file and a dictionary, as well as a text file, as input, and processes the text in a number of modes. The rule file consists of a set of morphophonological two-level rules, which define the deviant surface forms of characters in certain phonological environments. It is possible to define the environment for rule application accurately by referring to left and right contexts on lexical and surface levels. This rule facility simplifies the structure of the lexicon a great deal and speeds up processing.

For this specific purpose I have prepared such a version of the lexicon, which recognizes and accepts only those wordforms, for which there is a lexical entry in KKS, while all other strings of characters are considered unacceptable. When run in a filtering mode, SWATWOL produces a list of those discarded wordforms.

2. The internal test of KKS

The first task was to test the whole text of KKS and find out its internal inconsistencies and obvious weaknesses. After having run the program with the full text of KKS we can make the observation that there are discrepancies in all parts of speech, nouns, however, dominating. The most frequently occurring (and missing as entries in KKS) wordforms are *maneno* (279),[4] *nyingi* (108) and *mbalimbali* (108). Other commonly found forms are *nyama* (99), *nywele* (76), *ujuzi* (50), *kike* (49), *mizigo* (44), *mzigo* (23), *kuni* (24), and *mwenendo* (23). Those are wordforms occurring more than 20 times in KKS in the text, but not found as entries. The word neno is as an entry in singular, but because the plural *ma-* is not indicated, all plural forms are rendered as missing in the dictionary (or interpreted as belonging to class 10). The word *-ingine* is given as an entry in KKS but not *-ingi*.

When looking at the results in terms of word classes, nouns form the largest group.

The noun **class 3/4 (m/mi)** has such words as *mzigo* (23) and its plural *mizigo* (44), *mwenendo* (23), *mwongozo* (7), *mvurugiko* (5), *mteremko* (4), *msukumo* (4), *msisitizo* (4), *mshikio* (4), *mwendeshaji* (3), *mtukutiko* (3), *msuguano* (3), *mshughuliko* (3), *mshikilio* (3), *mpangilio* (3), *mkao* (3), and some less frequently occurring nouns as *mtutumko, mshikamano, msawazisho, msagiko, mbatilisho, mwujiza, mwonyeshaji, mwelekezo, mwangusho,*

[4] Numbers in parentheses mean the number of occurrences in the text concerned

mwanguko, mvutano, mvunjiko, mvujo, mtwazo, mtungamano, mtokezo, mtelezo, mtekenyo, mtapanyiko, mtanzuko, mtanuko, mtangamano, mtambalio, msuguo, msogeleo, msogeleano, mshororo, mshiko, mrusho, mrundiko, mropoko, mpungio, mpulizio, mpukuto, mpeto, mpambazuko, mnyanyazo, mmimininiko, mmeajamii, mkutaniko, mkurupusho, mkukuriko, mgutuko, mgusano, mgurumo, mgawo, mfuatizo, mfinyango, mfazaisho, mfananisho, mdhihirisho, mchukuano, mbabaiko etc It is not self-evident that all derived nouns of the form **m+stem+o** should be given as entries But it is hard to see why some are there and some others are not

In the noun **class 1/2 (m/wa)** there are also a few missing nouns, for example: *mwenzake/wenzake* (26), *wenziwe* (11), *wenzao* (4), *mwenzetu* (2), *mfanyakazi/wafanyakazi* (11), *mpiga* (6), *mwakilishi* (6), *mwanangu* (4), *mwanaharamu* (4), *mzaa* (3), *mwendeshaji* (3) Other less commonly found missing terms are: *mwuguzi, mtupa, mfanyi, mwanajeshi, mwanajamii, mwanaadam (in this form), mtoa, mtia, mtenda, mshona, mpa, mkuza, mkosa, mfuma* and *mcheza* Here again, there are derived forms of the type **m+root+a**, which may or may not be as entries in a dictionary

The noun **class 9/10 (n/n)** is the biggest noun group in Swahili, and this is reflected also in this analysis Missing words are such as: *nyama* (99), *kuni* (24), *manga* (10), *samawati* (6), *chapati* (6), *treni* (4), *tochi* (3), *skurubu* (3), *sentimita/sentimeta* (6), *sarufi* (3), *pondo* (3), *njuzi* (3), and *lori* (3) Other less commonly occurring words of this class are: *tunguu, tanuru, supu, slingi, sensuri, pembetatu, paredi, chororo, chengelele, azma, slipa, sketi, piano, petali, penseli, nomino, ngurumo, netiboli, lita, lensi, kriketi*, and *chagizo*

The noun **class 5/6 (ji/ma)** has such wordforms as: *maneno* (279), *maswali* (9), *matofali* (7), *mapaa* (7), *mapokezi* (6), *mapande* (5), *masimulizi* (4), *makusudio* (4), *maudhi* (3), *makelele* (3), *maingiliano* (3), *maendelezo* (3), *lori* (3), *kwato* (3), and *kombora* (3) Less common forms are: *matayarisho, mapendeleo, makabiliano, mahesabu, mafizi, maelekezo, madawa, machaguzi, mauaji, matepe, matando, mataka, mashirikiano, mashangilio, masemo, mapito, mapango, mapando, maongozi, maombaji, makutano, makaribisho, makaratasi, majibu, majibizano, maghala, maelekeo, machimbuko, machakura, mabaraza, maaskari, maanguko*, and *maamkiano* As can be seen in the list, many words are there because of the inadequate information in KKS on the plural forms of the nouns The classes 5/6 and 9/10 are particularly problematic in that a number of words may take a plural form according to either of them, and this is not indicated in KKS

The noun **class 7/8 (ki/vi)** has the following missing wordforms: *vishimo* (10), *vitimbi* (6), *vitawi* (6), *vidudu* (6), *kikonyo/vikonyo* (10), *kikaango* (6), *kijanja* (5), *kidudu* (5), *kijungu* (4), *vitumba* (3), *vilima* (3), and *vigogo* (3) Less common wordforms in this group are: *vijumba, vijidudu, vifurushi, vidimbwi, kikahawia, kijisanduku, kifupa, kifereji, visomo, visehemu, vipingiti, vikanyagio, vikamba, vijoka, vijisababu, vijani, vifito, vidoa, vidaraja, kizuio, kizuia, kizidishio, kiwamba, kipashio, kijiutando, kijibarua, kujibanzi, kifutio, kifupisho*, and *kifananisho* A number of these words are classified also to some other noun class, and therefore may be found in KKS in some other place Several of them are diminutives, and as

such very productive forms. It is not necessary to include them all in a dictionary, but some more common ones perhaps should.

The noun **class 11** is quite problematic, because the nouns belonging to it may have different plural forms, or none at all. Part of the following list may be explained by this fact: Missing words of this group include: *nywele* (76), *ujuzi* (50), *uongozi* (8), *uendeshaji* (6), *ufupisho* (5), *ufa/nyufa* (5), *utumishi* (4), *uelewano* (4), *utandu* (3), *unyong'onyevu* (3), *unyevunyevu* (3), and *umakanika* (3) There are also a number of less common ones: *uwezekano, uwendawazimu, utwezo, utumizi, utibabu, umilikaji, ulanzi, ukosekanaji, uhakikisho, uchukivu, ubaba, uzoroteshaji, uwekevu, uwajibikio, uvutio, uvamizi, utuuzima, utunzo, utunzi, utulizo, utukuzo, utenganisho, utengano, utatanishi, utanguzi, utambulisho, usoshalisti, usomeshaji, ushindani, usemeaji, upunguzo, upokezi, upekee, upatano, upadre, uonyesho, ukubaliano, ukopeshaji, uketuzi, ukawio, ukaribiano, ukabiliano, uhakika, ufuniko, uduni, uchelewevu, uchakavu, ubishani, ubinafsi*. Here again, there are words which may be classified under a different noun group. There are also a number of derived wordforms which are not necessary in a dictionary.

In the group of **verbs** the number of missing entries is rather small. There are some, however, such as: timka, timsha, tawaza, vurumisha, vungaza, vugumiza, topasa, toharisha, titimsha, titimka, sondea, pamua, pamaza, onyeza, nyong'onyeza, rendea, nyenyereka, kikisa, dabua, chusia, bonga, binikiza, and kura. Some of these are here due to their exceptional derived form. Some of them are given in KKS as a cross reference, but no entry for them is found.

In the group of **adverbials** some common words as *mbalimbali* (108) and *waziwazi* (10) are missing. Another obvious mistake is the omission of *kwani*.

The group of **adjectives** contains some important wordforms, e.g. *kike* (49), *kiowevu/viowevu* (26), *kienyeji* (17), *ungaunga* (15), *maarufu* (12), *makumi* (10), *machungu* (10), *chembechembe* (7), *mchungu* (4), and *onyevu* (3). Other less frequent forms are: *mchoyo, kipuuzi, kijinga, vumilivu, vilemavu, tiifu, motomoto, kubalifu, fahamikivu*, and *kitropiki*.

3. Modern prose text and KKS

The adequacy of KKS was also tested extensively with the help of the Helsinki Corpus of Swahili (HCT),[5] which contains presently 972 160 words of Standard Swahili prose text. The corpus is divided into two parts, the first one containing fiction texts (excerpts from books), and the second one texts from newspapers. Among the writers of these texts are Shaaban Robert, Julius Nyerere, Said A. Mohamed, Mohamed S. Mohamed, E. Kezilahabi and others. Also scientific texts, mainly from the field of linguistics, are included. Newspaper texts are

[5] In fact one could speak also about computerized text archives, because the contents of the corpus is not (yet) fixed, but it is being expanded all the time according to resources. Selected parts of these archives can be treated as a corpus and used as representative text material for research purposes. Other parts are less useful for this purpose, such as new technical terms in Standard Swahili, wordlists of various languages and dialects, and transcriptions of taperecorded oral materials in Swahili dialects.

from Uhuru, Mzalendo, Mfanyakazi, Kiongozi and Lengo, and they cover the period of 1988-1993. In the selection of newspaper texts attention was paid to the wide coverage of different types of texts and topics.[6]

Tests were made so that all such words which are not included into KKS as entries were extracted from the corpus. This includes also the detection of often complicated verbforms, which was made possible by the morphological parser SWATWOL. To avoid the handling of some commonly occurring words (nevertheless missing in KKS), the program was made to ignore them. Such words, some of which were discussed above, are: *ni, na, si, mbalimbali, mzigo, nyama, nywele, ujuzi, kike, kuni, mwenendo,* and different forms of the root *-ingi*. The findings are summarized in table 1.

Table 1. Swahili words in the Helsinki Corpus of Standard Swahili not found in KKS.

	Book text words	*% of total*	*Newspaper text words*	*% of total*
Occurs at least once	*1444*		*735*	
Occurs more than 2 times	*1066*		*457*	
Occurs more than 5 times	*382*		*203*	
Occurs more than 10 times	*183*		*115*	
Occurs more than 20 times	*93*		*57*	
Occurs more than 50 times	*38*		*13*	
Total number of different wordforms	*2180*		*830*	
Total number of wordform occurrences	*9385*	*1.32*	*5819*	*2.24*

One has to note that not all filtered wordforms have been represented in the above table. For example, there was a total of 9385 unrecognized wordforms found in book texts, and only 2180 of them are considered for inclusion into KKS. In newspaper texts there was a total of 5819 unrecognized wordforms, but only 830 could be accepted as candidates for inclusion into KKS. The total number of wordforms (2180 in book texts and 830 in newspaper texts) is bigger than the number of actual words (1444 in book texts and 735 in newspaper texts), because, for example, verbs occurring in different verbforms are counted only once in word count. On the other hand, the number of words in the above table does not represent strictly

[6] It would be also useful to test the coverage of KKS with colloquial Swahili. There is quite an extensive selection of spoken Swahili texts in the Archives of Swahili Dialects, which was compiled as a joint effort between the Institute of Kiswahili Research (University of Dar-es-Salaam) and the Department of Asian and African Studies (University of Helsinki) in 1989-1992. These texts are not, however, ideal for this kind of testing, because they contain substantially non-Standard Swahili

the number of lexical entries, because some of the nouns are represented twice, in singular and in plural Some of the wordforms are there for the reason that they have a plural form which is not indicated in KKS A number of nouns may take alternative prefixes according to the noun class 5/6 (ji/ma) or 9/10 (n/n) At least frequent occurrences of such nouns in the class 5/6 should be indicated in KKS

The percentage of unrecognized wordform occurrences has been calculated only from the total number of wordforms in the corpus The percentage in book texts is 1 32, which means that in the average every 75th word in text is unrecognized In newspaper texts the percentage is much higher (2 24), which is at least partly due to the fact that newspaper texts include large numbers of names, which are not, and should not be, included in KKS On the other hand newspaper texts represent the latest development of standard written text and contain such new words which are not in the dictionary

On the basis of the resulting word lists, which are too extensive to be reproduced here, we may make the following general observations Book texts and newspaper texts are treated together

1 There are more than 300 such nouns which have a plural according to the **class 6 (ma)**, although not indicated in KKS Some are quite frequent, as: *maswali* (180), *madawa* (173), *masuala* (154) *mageuzi* (137), *madevu* (129), *makokoto* (93), *maamuzi* (82), *magonjwa* (71), *majibu* (68), *maudhui* (63), *mandevu* (55), *mapendeleo* (38), *maandalizi* (36), *makafara* (34), *mapozi* (32), *masimulizi* (28), *majohari* (28), *mauaji* (26), *matope* (25), *majuzi* (23), *mashambulizi* (22), *makumbusho* (21), *makaratasi* (21), *mashule* (21), *maghala* (19), *maanguko* (18), *maelekezo* (18), *matofali* (16), *mapokezi* (15), *matayarisho* (15), *masaa* (14), *maudhi* (13), *mathibitisho* (13) Others are less frequent in the corpus, but equally important in the dictionary

2 More than 250 words of the **class 11 (u)** were filtered out from the corpus Among the most frequent ones are the following: *uongozi* (273), *umuhimu* (196), *utunzi* (99), *utumishi* (86), *ujumla* (78), *uwezekano* (71), *ukimwi* (64), *uhakika* (66), *upuzi*[7] (47), *uendeshaji* (33), *ushindani* (31), *ukimya* (29), *ukabila* (25), *utapiamlo* (28), *uharabu* (20), *uhamisho* (19), *umasikini* (19), *ugumba* (18), *ukwenzi* (18), *uasili* (16), *ubinafsi* (16), *usambe* (16), *utowezi* (16), *uwekevu* (16), *urais* (15), *ubaba* (14), *ujanajike* (14), *uzinduzi* (14), *ulanguzi* (13), *utaalam* (13), *ufedhuli* (12), *uonyeshaji* (12), *utibabu* (12), *uhusika* (11), *uhandisi* (10), *usoshalisti* (10), *utatanishi* (10)

Part of the words of this group are found in **class 9/10 (n/n)**, like *hakika, jumla, kabila, asili, rais,* and *baba,* but there is no indication that they may take also the u-prefix and have a different meaning

3 Among the words of **class 9/10 (n/n)** filtered out from the corpus are: *tamthiliya* (175), *demokrasia* (122), *litamu* (65), *drama* (71), *rasilmali/raslimali* (57) *pingamizi* (52), *baisikeli* (45), *khanga* (38), *teknolojia* (35), *chapati* (34), *posho* (34), *sekta* (30), *changamoto* (28), *taxi*

---

[7] Note that the form 'upuuzi' is found in KKS, but not the form 'upuzi'

(28), *taksi* (14), *ligi* (26), *milki* (26), *ngurumo* (25), *supu* (25), *shutuma* (24), *awamu* (23), *lori* (22), *pumziko* (22), *theluthi* (20), *ajira* (19), *menejimenti* (19), *akaunti* (18), *lishe* (18), *fundikira* (17), *burudani* (16), *dayosisi* (16), *radio* (15), *sarungi* (15), *dudu* (12), *operesheni* (12), *pigano* (12), *kumbukizi* (12), *lita* (11), *olimpiki* (11), *sarufi* (11), *shurti* (11), *fedhuli* (10), *kapu* (10), *kontena* (10), *samawati* (10). Part of these words may take plural forms according to the class 6 (ma)

4. A few words of the noun **class 7/8 (ki/vi)** should also be mentioned: *kilabu* (21), *kipaumbele* (13), *kitandawili* (8), *kiinua* (5), *kianzio* (3), *kitega* (3), *kizinga* (2).

5. As in the internal test of KKS (see above), it was found that derived nouns of the type **m+root+a** frequent also in the corpus. Such words belong to the noun **class 1/2 (m/wa)**, and here are some commonly occurring filtered nouns: *mfanya* (61), *mwenda* (28), *mpenda* (27), *mpiga* (27), *mhusika* (22), *muhusika* (15), *mfanyi* (15), *mtenda* (15), *mpanda* (10), *mzaa* (10), *mcheza* (6). Also derived forms of the type **m+root+aji/zi** are quite frequent: *mwongozi* (25), *mgombeaji* (11), *mjenzi* (5), *mchambuzi* (4), *muuzaji* (4), *mmilikaji* (3), *muigizaji, muungaji, muumbaji, mwendeshaji, mwonyeshaji, mwombolezaji, mwonaji*. Other words of this noun class include: *mfanyabiashara* (73), *mwakilishi* (43), *mfanyi* (27).

6. The compound word construction using the productive root *'mwana-'* is rather poorly represented in KKS. There are many more common constructions, such as: *mwanakijiji/wanavijiji* (192), *mwanakisole* (37), *mwanakisomo* (37), *mwanakamati* (24), *mwanasiasa* (49), *mwanamichezo* (21), *mwanajeshi* (18), *mwanandege* (16), *mwanariadha* (16), *mwanamuziki* (13), *mwanakwaya* (11), *mwanakwetu* (10), *mwanamgambo* (10), *mwanasheria* (10), *mwanavijiji* (10) *mwanasayansi* (7), *mwanataaluma* (7). Other less commonly occurring constructions are: *mwanaelimu, mwanafalsafa, mwanafasihi, mwanahalmashauri, mwanaharamu, mwanahistoria, mwanaisimu, mwanakikundi, mwanalugha, mwanamapokeo, mwanamji, mwanamwari, mwanasaikolojia, mwanaserere*

7. The noun **class 3/4 (m/mi)** is represented in the list by e.g. the following examples: *mwongozo* (61), *mkakati* (50), *mpangilio* (49), *mwelekeo* (46), *mvutano* (39), *msukumo* (38), *mshikamano* (33), *msisitizo* (30), *mswada* (27), *michuano* (26), *mwamko* (12), *mkanganyo* (10).

8. There are some more rather common words, not belonging to any of the above categories, which deserve to be mentioned here. Such are: *maalum* (205), *maarufu* (145), *mojawapo* (81), *haramu* (75), *mwishowe* (66), *polepole* (45), *majuzi* (33), *makwao* (31), *kamala* (10), *kemkem* (10). Also compound forms such as *mwanangu, mwanawe, mwanao, babangu, mamangu* etc. are missing. As they can be derived according to the known rules, it is not selfevident that they should be included. But it is difficult to understand such an omission that all the weekdays except for *Alhamisi* and *Ijumaa* are missing.

9. Verbs are the most difficult group to handle, because each verb may take so many different forms. Derived forms also complicate the search, and it is not always clear whether the derived forms given in KKS really cover the actual use of each verb. This would be another question to investigate in more detail. Here are some of the verbs which are not in KKS: *chekelea, dekeza, ekea, ekwa, emesha, emewa, epukika, faharisha, fanikisha, fuatiliza,*

fyeruka, gagamia, garagaza, gharamia, ghumiwa, goteza, gugubia, harakia, kakamua, epukika, kumbukika, kurubia, lega, lembusha, lundikia, ombeza, ondokozwa, pembeza, refuka, ripuka, roa, singiza, telemkia, teremkia, tengamaa, tengemea, vyagaza, zoeleka, zongoa. Some of the verbs are in the list because of a different spelling from that given in KKS. Common variation in such phenomena should also be indicated.

4. The acquisition of new lexical data in context

The system described above allows the retrieval of lexical materials from the corpus in context. By user-defined means it is possible to retrieve contents in various quantities and include background information into the findings, such as the author, page number etc. Probably the most useful and convenient unit is the sentence. It is possible to mark the findings in the way one wishes and also to sort the data according to the keyword. Such an automatic sorting does not, however, bring always satisfactory results, because many words, particularly verbs, have many kinds of prefixes preceding the stem. A desired result may be obtained either through preceding each finding with a keyword in a stem form, or through a program which identifies the boundary between the stem and prefixes.

Tests with modern prose material show that KKS covers quite well the normal text found in fiction books and newspaper texts. There are, however, a number of words, which have become or are becoming common in modern Swahili, so that they should be considered to be included in KKS. Due to limitations of space, in the following I will give only some examples extracted from different texts. These examples are given in context to show how they are used.

Source: Said A. Mohamed, **Tata za Asumini**. Nairobi: Longman

Bila ya shaka baba'ako <<alidunguka>> sana kwa kuondoka kwako na hasa aliyokuwa akiyasikia.

Asumini bila ya shaka <<aliemewa>> na kuubana kimya.

Alizidi <<kuemewa>> Zaina.

Asumini aliizuia na Mkejel alipoona kazuiwa alitoa pumzi moto na kunywea kama <<bofu>> lililodungwa sindano.

Lakini <<bwanaarusi>> mwenyewe nd'o yupi?

Aliwatazama wanawake wenzake jinsi walivyojipamba kwa <<fensi>> na vituko.

Alinikuta sina <<khanga>> ya kujitanda, kanipiga vibaya vibaya.

Aliwahi kwenda kwenye <<kliniki>> ile mara nyingi na kukosa dawa.

Umejaa, umefura na <<kufufurika>> kupita kiasi.

Minazi na mivinje iliyosimama pembezoni <<kuhemkwa>> na bahari, pepo zilipovuma na kupoa nayo pepo ziliposita, ilikuwa, mwisho wa jicho lake, kama minazi na mivinje ya bandia na ya kuchezea.

Asumini alichora <<mfunda>> ardhini na wote walijipanga kwenye mfunda huo.

Sheikh Mkejel kavaa suruali ya <<mfyuro>> iliyomfika hadi magotini.

Hakuvua hata ule <<mkanzu>> wake

Alitangaza Bi Feruzi dhahiri, "Mimi nawaonea <<mushkeli>> waalimu wa aina hiyo

Asumini alibaki kukodoa macho alipoona kifua cha <<mwanandege>> wa kike kinaparazaparaza mwili wa yule barubaru ambaye yeye kamkimbia mbingu na ardhi

Sewa aliwaza juu ya maisha ya kundi kupiga <<mwijiku>> na kuzamia tena;

Asumini alimwuliza Sewa kwa taathira ileile ambayo Sewa aliitumia kuuliza <<suali>> lake

Iko wapi <<tochi>> yako ya kumulikia akili?

Na bila ya shaka wakati mwingine alicheza <<umeta>> ulipositawi uwanjani

Source: M M Mulokozi and K K Kahigi, **Kunga za ushairi na diwani yetu**:

Kizungumzwa ndicho kitu halisi ambacho kinazungumziwa na <<kifananishi>> kwa mafumbo

Nani atayasikiya, mwema hanywi wazi wazi, Bia hunywewa ubiya, lango hutiwa <<komezi>> Wanywaji hugugumiya, nani atawamaizi?

Tofauti kati ya <<konsonansi>> na takriri-konsonanti ni kwamba konsonansi ni marudiorudio ya sauti za kikonsonanti katika maneno mbalimbali

Na <<marenja>> yanabomoa njia zetu za udongo!

Oneni hirizi na <<midali>> za uwongo wa jana na leo

Katika shairi lake la "Kufua Moyo" Shaaban Robert haombolezi bali anashangilia matendo au matukio ya mashujaa ya kujitoa <<muhanga>> vitani

Lugha bila <<ridhimu>> ingekuwa ni mfuatano wa sauti tu usio na maana yoyote

Usambamba ni <<takriri>> ya sentensi au vifungu vya maneno vyenye kufanana kimaana au kimuundo

Sitiari ni <<tamathali>> ambayo athari yake hutegemea uhamishaji wa maana na hisi kutoka katika kitu au dhana moja hadi kitu au dhana nyingine tofauti

It can be seen from the examples given above that the words not recognized by KKS are of quite a different kind depending on the type of text. The words of Kunga za ushairi are from the field of literature studies, and it is not necessary to include all these words into KKS, because many of them do not belong to the basic vocabulary of the language. The words caught from Tata za Asumini are somewhat writer-specific and dependent on the subject dealt with in the book concerned.

5. Summary

The above discussion shows the power and accuracy of computerized information retrieval, particularly when language-specific tools can be utilized. I have described ways how one presently relevant task, updating Kamusi ya Kiswahili Sanifu, can benefit from recent advances in computational linguistics. The system helps in detecting weaknesses in the existing dictionary as well as in retrieving new lexical material from relevant texts. The examples of missing words given above give a superficial picture of the whole truth. Only the most often occurring examples were given. And it is not at all clear that those are the only words, or even the most important ones, to be considered for inclusion into KKS. From the viewpoint of the user, often the more rare words are the ones which one is likely to search from the dictionary.

And if they are missing, the dictionary does not fulfil its task properly. It is not practical, however, to list all those words here. Rather they are matarial for the databank, preferably with sufficient amount of context included.

Although I have discussed only tests made with KKS, the system described above may be applied to dictionaries of any language. KKS is not a very difficult case, because its lexical entries are single words. The work becomes more complicated, if we have to deal with collocations, phrases, and other phenomena, which are represented in the lexicon with more than one word. But also such a task is just a further challenge for developing programs to deal with such phenomena.

References

Brodda, Benny. 1990. "Corpus work with PC Beta." *Papers presented to the 13th International Conference on Computational Linguistics*. Vol 3, ed. by Hans Karlgren, p. 405-409.

Calzolari, Nicoletta and Bindi, Remo. 1990. "Acquisition of Lexical Information from a Large Textual Italian Corpus." *Papers presented to the 13th International Conference on Computational Linguistics* Vol 3, ed. by Hans Karlgren, p. 54-59.

Chuwa, Albina. 1987. "Uingizaji wa methali katika Kamusi ya Kiswahili Sanifu." *Kiswahili* 54,1-2: 202-214.

Hurskainen, Arvi. 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, Arvi. 1992. "Computer Archives of Swahili Language and Folklore - What is it?" *Nordic Journal of African Studies* 1,1:123-127.

Taasisi ya Uchunguzi wa Kiswahili. 1981. *Kamusi ya Kiswahili Sanifu*. Arusha: Oxford University Press.

Khamisi, Abdu M. 1987. "Trends in Swahili lexicography." *Kiswahili* 54,1-2: 192-201.

Koskenniemi, Kimmo. 1983. "Two-level Morphology: A General Computational Model for Word-Form Recognition and Production." Department of General Linguistics. University of Helsinki. Publication No. 11.

Mohamed, Said A. 1990. *Tata za Asumini*. Nairobi: Longman.

Mulokozi M.M. and Kahigi, K.K. 1979. *Kunga za ushairi na diwani yetu*. Dar-es-Salaam: Tanzania Publishing House.

Taasisi ya Uchunguzi wa Kiswahili. 1989. *Makala za Mkutano wa Kimataifa wa Usanifishaji wa Istilahi za Kiswahili*. Dar es Salaam: Chuo Kikuu cha Dar es Salaam.

Tumbo-Massabo, Zubeida N. 1989. "Jinsi ya Kuanzisha Benki ya Data ya Istilahi/Msamiati." *Makala za Mkutano wa Kimataifa wa Usanifishaji wa Istilahi za Kiswahili* p. 64-66. Dar es Salaam: Taasisi ya Uchunguzi wa Kiswahili, Chuo Kikuu cha Dar es Salaam.