Nordic Journal of African Studies 4(2): 81-92 (1995)

Information Retrieval and Two-directional Word Formation[1]

ARVI HURSKAINEN
University of Helsinki, Finland

ABSTRACT

This paper presents some applications of SWATWOL, a morphological parser of Swahili, for information retrieval. It presents a solution to the problem of retrieving accurate linguistic information in a language, where word formation branches out from the lemma to both directions. After discussing technical problems and their solution, some research tasks that have been carried out, or which are in process, are described. Among those are: testing dictionaries, retrieval of loan words from running text, word-form distribution of lemmas, and total distribution of words (counted according to lemmas) in different types of text. The obviously most important application is the production of a database for dictionary compilers, with a selection of contexts of use for each lemma.

INTRODUCTION

In retrieving information from an untagged running text, agglutinating languages pose a problem in that words may be found in several forms. For certain tasks it is not therefore possible, or at least it is not convenient, to use a direct string search method. By utilizing possibilities offered by regular expressions we may construct quite complicated and accurate search keys, which already are vastly more efficient than concrete surface search strings. Even this is not enough in some cases.

Let us take an example of Swahili monosyllabic verbs. For example, in a verb-form *wa-li-po-kwisha-wa-l-ish-a* (when they were finished with feeding them) 'l' (to eat) is the only constant element, the verb root. If the task is to retrieve monosyllabic verbs from running text, the only viable solution is to analyze the word-form first and then use the information thus obtained in further processing.

[1] A preliminary version of this paper was read at NODALIDA-95, the 10th Nordic Conference of Computational Linguistics, Helsinki 29-30 May 1995.

When we analyze the above word-form with SWATWOL[2] we get the following result:

walipokwishawalisha : wa-li-po-kwisha-wa-l-ish-a
"<walipokwishawalisha>"
"la" 1/2-SP-PL3 PAST 16-REL-SG COMPL 1/2-OBJ-PL3 VB-MONOSLB CAUS
"la" 1/2-SP-PL3 PAST 16-REL-SG COMPL 1/2-OBJ-PL2 VB-MONOSLB CAUS

Here we have two readings, because the object marker 'wa' is identical in PL2 and PL3. Nevertheless, in both of them we have the full morphological information as a source of further analysis.

The above example also shows that verbs in Swahili are constructed of a number of slots, which may be filled with morphemes appropriate for each particular slot. It is also typical that pre-root slots may be filled by several alternative morphemes. The structure of Swahili verb is as follows:

Figure 1.

Verb structure	*Example*
1. Pre-initial negative prefix	
2. Subject prefix	wa-
3. Negative prefix	
4. Time/aspect prefix	-li-
5. Relative prefix	-po-
6. Prefix for completed action	-kwisha-
7. Object prefix	-wa-
8. Infinitive marker	
9. Verb root	-l-
10. Derivational suffix (applicative, causative, stative, neutro-passive, reciprocal form, passive etc.)	-ish-

[2] SWATWOL is a morphological analyzer designed for analyzing Standard Swahili language. It is based on the two-level formalism, where each character has a lexical and surface representation, and where morphophonological inconsistencies are dealt with by two-level rules. For more detailed information on the formalism in general see Koskenniemi (1983), and for Swahili see Hurskainen (1992). For applications of SWATWOL for research tasks see Hurskainen (1994a, 1994b, 1995a, 1995b). The first morphological analyzer for Swahili, AINI, was designed by Schadeberg and Elias (1989). It has been applied for information retrieval by Toscano (1991/92, 1994). To my knowledge, no extensive research results have been published or carried out with AINI.

11. Derivational suffixes in combination with one or more of the above
12. Verb-final vowel -a
13. Post-final relative suffix

In two-level formalism, words are constructed from a consecutive series of morphemes, from left to right. The lexicon system consists of a series of sub-lexicons, where continuation from one sub-lexicon to another is precisely defined. In an ideal case, the lexicon system is so assembled, that each word-form starts from a common root lexicon and branches out so that the lexicon system forms a tree-like construction. Unfortunately all word-forms, e.g. verb-forms in Swahili, do not follow this ideal pattern. In implementing an analyzer, problematic are restrictions which occur before the verb root, but which are triggered by a morpheme occurring after the root.

Particularly problematic is the slot 13 above, the post-final relative marker, which prevents the occurrence of slots 1, 3-6, and 8. The location of the post-final relative marker at the very end of the verb-form, i.e. after the large root lexicon, is what causes problems in building the lexicon system. When such a suffix is encountered at the very end of the word construction process, the system must be able to make restrictions in what already was accepted earlier in building grammatically correct forms. In the present formalism, these restrictions are being taken care of by rules, which state that if a feature (in this case post-final relative marker) is encountered in word-building process, certain slots (here 1, 3-6 and 8) must be empty. Through such a rule system it is possible to 'mimic' backtracking, although a finite state automaton does not literally do it. In each two exclusive slots, a special diacritic is needed for enabling the rules to identify such problematic slots.

1. Applications of SWATWOL for Retrieving Quantitative Data

I have integrated the morphological analyzer SWATWOL for information retrieval from unrestricted and uncoded Standard Swahili text (Hurskainen 1995a). This facility improves greatly the accuracy of search, and hence enlarges the area of viable research tasks. Although SWATWOL is a morphological parser, it covers also many such features which in isolating languages are part of syntax. Particularly verb morphology in Bantu languages extends to the area often belonging to the domain of syntax.

1.1 TEST WITH MONOSYLLABIC WORDS

Above was given an example of a monosyllabic verb to demonstrate that it is impossible to use the verb root as a search key for finding monosyllabic verbs. To show that the system really works, I have retrieved all monosyllabic verbs from a corpus of 1.2 million words. Table 1 shows two kinds of search results: (1) the number of forms each monosyllabic verb has in the corpus, and (2) the total number of occurrences of each monosyllabic verb. There are seven monosyllabic verbs in Swahili: *w* (to be), *p* (to give), *f* (to die), *l* (to eat), *j* (to come), *ny* (to drop like a rain), *ch* (to rise (of sun)).

Table 1. Occurrence of monosyllabic verbs in Standard Swahili Corpus.

Root	*Gloss*	*Occurences*	*Number of forms*
w	to be	33,112	889
p	to give	1,282	615
f	to die	1,835	507
j	to come	1,507	289
l	to eat	1,097	280
ny	to trickle	447	106
ch	to rise	99	53

There are big differences concerning the distribution of these verbs in text. I have no aim in this paper to analyze these differences. I only want to make the general observation that these differences are what could be expected, given the meaning and functions of these verbs. The verb *w* (to be) is the most common one, as expected. It would be even more frequent if Swahili would use this verb also in present tense constructions. Instead it uses the uninflecting particle *li*, which occurs in corpus 10.914 times.

The only problem in retrieving monosyllabic stems is related to ambiguity of word-forms. Some of the forms are as well forms of some other verbs, and in a few cases of nouns. Both prefixes and suffixes contribute to the ambiguity, and only word-external information can solve such problems.

1.2 RETRIEVING LOANWORDS FROM RUNNING TEXT

One of the implementations is the version which identifies loanwords from running text, showing the language of origin of each lexeme.[3] In the case of Swahili, such information has also wider interest related to language policy. Among part of the researchers and especially politicians in East Africa there is an interest to show that Swahili is not an 'offshoot' of Arabic, but a genuine Bantu language. Serious researchers never doubted this, but the debate still continues. On the other hand, on the islands of Indian Ocean and in the coastal Swahili settlements there is a trend to introduce more Arabic loans into the language than in the interior of the continent, where the influence of Islam and Arabic is less significant. The morphological analyzer is useful in finding out the use of such Arabic loanwords in different types of texts, from different times, written by writers of different origin and from different areas. With the aid of such an analyzer it is also possible to compare, how loanwords are represented in a dictionary, and how they are actually used in different types of texts. Table 2 summarizes results of selected analyzed texts.

Table 2. Arabic loans in selected Swahili texts. (NEWS = Newspaper texts; Nyerere = Julius Nyerere, Ujamaa; Shaaban = Selected works of Shaaban Robert; Mohamed_nyo = Mohamed S. Mohamed; Nyozi za Usiku; Mohamed_tat = Said A. Mohamed, Tata za Asumini; Tippu = Tippu Tip, Maisha ya Tippu Tip; Fasihi = Fasihi, Makala za Semina ya Kimataifa ya Waandishi wa Kiswahili, Dar-es-Salaam)

Source	*Total*	*Arabic loans*		*Unique*	*Arabic loans*	
	word-forms	*all*	*%*	*word-forms*	*unique*	*%*
NEWS	288,528	59,879	20.8	31,080	6,287	20.2
Nyerere	49,566	12,262	24.7	6,705	1,664	24.8
Shaaban	106,784	28,708	26.0	17,921	4,599	25.7
Mohamed_nyo	38,144	7,187	18.8	9,499	1,765	18.6
Mohamed_tat	45,721	9,633	21.1	11,164	2,017	18.1
Tippu	26,067	4,957	19.0	5,540	847	15.3
Fasihi	78,818	19,163	24.3	13,958	3,006	21.5

Seven different types of text were chosen for analysis. Each of them was analyzed in terms of total representation of Arabic loans, where also multiple occurrences of the same form were counted, and also in terms of unique occurrences. NEWS represent standard newspaper prose. Nyerere is an example of an up-country Swahili political text. Shaaban represents an older standard prose. The writers of

[3] Substantial work in identifying Arabic loans was made by Johanna Kestilä, and Faruk Abu-Chacra and Haseeb Shehadeh helped in identifying problematic cases. All of them deserve special thanks for their contribution.

Mohamed_nyo and Mohamed_tat are Muslim native speakers of Swahili. Tippu is an autobiography of an Arab merchant from last century, and Fasihi consists of scientific text of literature research.

At first glance results are surprising. Non-native speaker Swahili contains Arabic loans more than the texts of Muslim native speakers. Another interesting phenomenon is that the proportion of Arabic loans is almost the same in both types of analysis (total and unique). A third interesting result is that some of the percentages are almost the same as the proportion of Arabic loans in the latest monosyllabic dictionary, Kamusi ya Kiswahili Sanifu (24%). These quantitative results give hints to many kinds of hypotheses, which can be further studied and tested. For example, in BOOKS out of 500 most common lemmas a total of 118 were verbs. Out of these verbs 18 were of Arabic origin. It gives only 16 % of the total, which seems surprisingly low. This hints to the possibility that the most generally used verbs are of Bantu origin and Arabic loans are among the less frequently occurring verbs.

1.3 LEMMA-BASED RETRIEVAL

As can be guessed from the above, the capability of SWATWOL to identify lemmas also in quite controversial cases has a number of applications. It enhances an accurate research of vocabulary also in morphologically complex languages. Verbs in Bantu languages are particularly complex. The large number of morpheme slots in a verb-form is not the only factor contributing to this complexity. Perhaps still more important is the fact that most of the pre-root slots can be occupied by more than ten different morphemes. For example, slots 2, 5, and 7 may be occupied by any of the 15 different morphemes referring to the noun class concerned. The fact that some of these prefixes have an identical form, although they refer to different noun classes, causes ambiguity in analysis, and this can be resolved only on the basis of extra-word information.

1.3.1 WORD-FORM VARIATIONS IN TEXT

Below are results of two search tasks, where the aim was to find out, in how many different forms each lexical lemma appeared in two types of corpus text. The first one (NEWS) consists of contemporary newspaper texts from 1988-1994, containing texts from several Swahili newspapers (a total of 288.528 word-forms). The second one (BOOKS) is a text collection of 35 prose books or booklets (a total of 673.848 word-forms).

In NEWS, a total of 6020 word lemmas were found. Note that only those word-forms were counted that SWATWOL was able to analyze. The word-forms discarded were non-Swahili words, misspellings, and a small number of genuine

Swahili words, which are considered too rare or transient to be included into the lexicon. As could be expected, words with the largest number of word-forms were verbs. A total of 357 verbs had more forms than any word of the other word classes. The word with the largest number of forms in NEWS is fanya (to do), with 621 different forms. On top of the word-form frequency list are the lemmas given in Table 3.

Except for verbs, words with several possible forms are some adjectives (excluding non-inflecting ones) and numerals with Bantu origin (1-5 and 8). In order of frequency in NEWS they are: *ingi* (many, 12), *enye* (who has, which has, 12), *kuu* (big, 11), *ote* (all, 10), *dogo* (small, 10), *zito* (heavy, 9), *geni* (stranger, 9), *wili* (two, 8), *tatu* (three, 8), *tano* (five, 8), *chache* (few, 8), baya (bad, 7), *enyewe* (oneself, 7), *pya* (new, 7), *refu* (long, 6), *bovu* (bad, rotten, 6), *chafu* (dirty, 6), *eupe* (white, 6), *eusi* (black, 6), *zuri* (good,6), *moja* (one, 6), *nne* (four, 5), *nane* (eight, 5), *bichi* (raw, 5), *epesi* (light, 5), *fupi* (short, 5), *gumu* (hard, difficult, 5), *kali* (hard, angry, 5), *tamu* (sweet, 5), and *zee* (old, 5).

In the other part of the corpus, i.e. BOOKS, the total number of words (lemmas) was 10,049, considerably more than in newspaper texts (see above). The lemma with the largest number of word-forms in this text is *ona* (to see) with 959 different forms. The top list of 20 is given below in Table 3. Also the combined list of word-form diversity in NEWS and BOOKS is given in Table 3. The 20 verbs with biggest number of word-forms are presented in the order of grand total diversity (NEWS and BOOKS combined).

Table 3. The 20 verbs with the largest number of verb-forms in NEWS and BOOKS.

Lemma		NEWS	BOOKS	NEWS+BOOKS
fanya	to do	470	849	1.087
ona	to see	199	959	1.038
pata	to get	356	837	996
wa	to be	384	803	889
toka	to leave	232	618	699
taka	to want	261	507	660
weza	to be able	262	578	656
0pa	to give	215	523	615
acha	to leave	158	506	606
tumia	to use	248	436	529
sema	to say	171	466	518
fa	to die	157	439	507
ita	to call	122	405	463
anza	to begin	240	337	445
pita	to pass	118	369	438

weka	to put	171	309	413
eleza	to explain	179	291	401
kuta	to meet	119	324	389
toa	to give	155	278	347
saidia	to help	121	250	319
omba	to ask for	110	217	284

The statistics show that the lemma with the largest number of word-forms in Swahili is *fanya* (to do, 1.087 forms). The number is not absolute, of course. The more text we have in corpus, the more likely it is that the number of word-forms increases. This is testified also by the statistics of NEWS and BOOKS, which both have fewer word-forms than both together. There are also some unexpected results in the list of NEWS. While BOOKS follows generally the order of the combined list, in NEWS there are such words as *ona, acha, tumia* and *anza,* which deserve more research.

1.3.2 Word-form Variation and the Total Occurrence of Word-forms

Table 4 below shows the number of word-forms, which the 20 most branching lemmas have in Swahili, as was in Table 3. For comparative reasons, also the number of total occurrences of these lemmas is given. By comparing the relation of the total occurrence with unique occurrences, an average frequency of the same word-form of each lemma can be calculated.

The statistics show that some of the verbs have a high frequency of identical verb-forms (up to 40.6 by *wa*), while some others have less than two identical verb-forms in the average. The general tendency is that the most frequent verbs are also those with the largest number of identical word-forms, but this rule has exceptions, as is shown in Table 4.

Table 4. The 20 verbs with the largest number of verb-forms in NEWS and BOOKS.
Total number of occurrences of each lemma also shown.

Lemma	*NEWS*			*BOOKS*		
	Number	*Total occ.*	*Mean*	*Number*	*Total occ.*	*Mean*
fanya	470	1,503	3.2	849	3,306	3.9
ona	199	466	2.3	959	4,429	4.6
pata	356	1,027	2.9	837	2,996	3.6
wa	384	15,599	40.6	803	33,112	41.2
kuwa		(4,444)			(13,997)	
toka	232	1,289	5.6	618	2,008	3.2
taka	261	732	2.8	507	2,352	4.6
weza	262	1,265	4.8	578	4,021	7.0
pa	215	285	1.3	523	1,282	2.5
acha	158	369	2.3	473	1,077	2.3
tumia	248	557	2.2	436	1,128	2.6
sema	171	1,960	11.5	466	4,035	8.7
fa	157	397	2.5	439	1,835	4.2
ita	122	201	1.6	405	1,001	2.5
anza	240	885	3.7	337	2,659	7.9
pita	118	470	4.0	369	1,112	3.0
weka	171	430	2.5	309	770	2.5
eleza	179	511	2.9	291	822	2.8
kuta	119	243	2.0	324	1,017	3.1
toa	155	725	4.7	278	1,051	3.8
saidia	121	297	2.5	250	647	2.5
omba	110	205	1.9	217	540	2.5

1.4 Frequency of Lemmas in Text

SWATWOL also makes possible to make frequency analysis of lemmas in various kinds of texts. While above was an example of an analysis of the number of forms which each lemma has in a given text corpus, below is a demonstration of frequency count of lemmas in the same texts.

The list of the most frequently occurring words (lemmas) of NEWS shows that it is quite different than the list of multiple word-forms. On top are uninflecting words, such as *na* (and), *ya* (gen. connector), *kwa* (with), *katika* (in), *za* (gen. connector), *la* (gen. connector). The most frequent noun is *nchi* (country, land), and the most frequent verb is *sema* (say). An exception is the monosyllabic verb *w* (to be) which is number one in frequency. Its infinitive form *kuwa* is used also as a

copula in sub-ordinating clauses in the sense 'that', 'so that'. But also as a verb it is the most frequent word in Swahili.

In BOOKS the list of most common lemmas is as follows: *na* (and), *ya* (gen. connector), *kwa* (with), *katika* (in), *mtu* (man, human being), *kama* (when, if), *yake* (poss. pronoun), *za* (gen connector), *ona* (see), *la* (eat), *lakini* (but), *sema* (say), *weza* (be able). Also here the most common one is, however, the verb *w* (to be).

2. SWATWOL IN TESTING DICTIONARIES AND IN RETRIEVING LEXICAL DATA

The filtering mode of SWATWOL can be used for listing such words which are not in a given dictionary. This requires such a lexicon which contains only those words that are as entries in a dictionary. I have tested Kamusi ya Kiswahili Sanifu, a monolingual dictionary of Swahili, and found out a number of inconsistencies in it (Hurskainen 1994a). For example, more than 600 such words were used in explanations that were not as entries in the dictionary. Missing were also a number of words which were in rather common use in newspaper texts and in prose books. Such a facility can be designed and applied for testing any dictionary, and it can be used as a tester in compiling new dictionaries.

The capability of identifying the lemma of all word-forms is a necessity in compiling a lemma-in-context type of data-base for dictionary compilation. The new Swahili-English Standard Dictionary will be based on the data-base compiled in this way.

3. CONCLUSION

The application of a morphological analyzer for information retrieval has proved a powerful method, particularly in languages with two-directional word formation. Morphological information is far more reliable as a key for information retrieval than direct string search. The work with Swahili will continue in the field of disambiguation (Voutilainen & Tapanainen 1993) and the problems of syntactic and part-of-speech parsing (Karlsson 1990, Karlsson et al 1993; Koskenniemi et al 1992). One can note with satisfaction that while the working environments are becoming increasingly powerful, the time needed for processing is becoming insignificant, even in analyzing large corpora. The prospects in this field look most promising.

REFERENCES

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

1992b *Computer Archives of Swahili Language and Folklore - What is it?* **Nordic Journal of African Studies** 1(1): 123-127.

1994a *Kamusi ya Kiswahili Sanifu in test: A computer system for analyzing dictionaries and for retrieving lexical data.* **Swahili Forum 1** (AAP 37): 169-179.

1994b *Quantitative Analysis of Swahili Noun Classes.* Working Papers in Linguistics 22, University of Trondheim. Pp. 1-16.

1995a (forthcoming): *A Language Sensitive Approach in Information Management and Retrieval. A Case of Swahili.* Paper presented at the First World Congress of African Linguistics (WCAL) 18-22.7. 1994, University of Witwatersrand - University of Swaziland.

1995b (forthcoming): *Affirmative and Negative Tense/Aspect marking in Swahili.* Working Papers in Linguistics, University of Trondheim.

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

Karlsson, F. 1990.
Constraint Grammar as a Framework for Parsing Running Text. In: Papers presented to the 13th International Conference on Computational Linguistics, H. Karlgren (ed.). Helsinki. Vol. 3:168-73.

Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. 1993.
Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text. (In print).

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

Koskenniemi, K., Tapanainen, P. and Voutilainen, A. (1992).
Compiling and using finite-state syntactic rules. In: Proceedings of the fifteenth International Conference on Computational Linguistics. COLING-92. Vol I, pp. 156-162, Nantes, France. 1992.

Schadeberg, T.C. and Elias, P.S.E. 1989.
AINI: A Morphological Parser for Kiswahili. Leiden: Department of African Linguistics (State University at Leiden).

Toscano, M. 1990/91.
Manuale per l'analisi morfologica computerizzata di testi swahili. Serie Didattica 2. Napoli.

1994 *From text to Dictionary: Steps for a Computerised Process.* **Swahili Forum 1** (AAP 37): 181-195.

Voutilainen, A. and Tapanainen, P. 1993.
Ambiguity resolution in a reductionist parser. In: Proceedings of EACL-93. Utrecht.