Technical Reports on Language Technology
Report No 61, 20020
http://www.njas.helsinki.fi/salama

Msimulizi as corpus for accurate search[1]

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

In Technical Report 60[2], I described the process of converting printed text into machine-readable form. This report is an extension to it, and here I will go into more detail in describing and demonsrating the capabilities of the search system based on analysed text. All the material on Msimulizi (years 1888-1896) that is available on SOAS web page was processed into machine-readable form, including manual editing of the whole text. The second round of editing was done on the basis of computational analysis, which points out the remaining scanning mistakes. The clean text was then converted into an analysed format, which is optimal for information retrieval. The report demonstrates especially such search tasks, which are hardly possible using conventional string search, due to the complex word structure of Swahili.

Key Words: *information retrieval, morphological analysis.*

1 Introduction

Msimulizi is the first printed periodical in Swahili language. It started to appear in 1888 with intervals of two months. Its aim was to provide a communication channel between various mission stations of the Universities, Mission to Central Africa. The stations are located in the area of the current Tanzania and Malawi. In addition, news in the area of current Uganda are also reported. The mission agency had also work there, which is why there was great interest to hear news fom that area also.

The aim of the periodical was to have news reports from each station in each issue. Only seldom this fully realized, due to various reasons. The periodical includes also general news on current themes, such as conflicts between colonial powers (England, Germany, Portugal) and the local Arab administration.

The Universities' Mission to Central Africa was established as response to the alarming news on ruthless slave trade of Arabs in eastern and central Africa. Especially the reports of Livingstone gave impetus to action. *Msimulizi* has many reports on how slaves were still raided and sold to Arabs, and how mission agencies tried to get slaves to mission schools and give them Christian education. Former slaves played important roles in mission schools.

---

[2] http://www.njas.helsinki.fi/salama/printed-text-into-machine-readable-form.pdf

According to *Msimulizi* texts, mission work was very much concentrated on education. Medical care was also part of the work, but it is referred to only when someone badly ill is moved to another station for treatment.

Training continued to up to six years of education, and examinations were arranged at certain intervals. *Msimulizi* reports on these examinations and mentions even the names of those who passed on each class.

Msimulizi reveals that Africa was dangerous to whites. Astonishingly many of the missionaries died during those years. There is no diagnosis of deaths, but fever is usually reported to be the reaon of death. Obviously, malaria was the main cause of deaths.

Msimulizi follows contemporary historical events. In addition to Livingstone, who already had died that time, the periodical tells about Stanley, who travelled across the continent. Also, the whereabouts of Emin Pasha in Uganda were reported on.

2 Language

The Swahili language of the periodical is not even. The writing style and even the orthography varies depending on the writer. It is not known how much the reports were edited before printing. Yet different styles are apparent.

Considering the time of writing - about 140 years ago - the language has changed remarkably little since then. There are underdeveloped elements, such as the lack of month names (English forms are used) and the Arabic use of some numbers. Also, the naming of the week days is searching for its form.

Most of the *Msimulizi* text was written by advanced students, selected as reporters in each mission station. Reporting is detailed in relation to time. Usually the day of the month, and often also the week day, and sometimes also the time of the day, are mentioned in reporting on events. The events themselves were usually related to travelling and local festivities, which are usually Christian, but also sometimes traditional.

From the viewpoint of constructing an accurate information retrieval system for *Msimulizi*, it is important to be able to treat also the non-standard forms properly. The morphological analyzers normally recognize the standard forms only, and non-standard forms are treated with a guesser. This method is very unreliable, especially if the non-standard forms are verbs. Therefore, I have included the non-standard forms of *Msimulizi* into the morphological lexicon, so that they can be retrieved as accurately as the standard forms.

3 The structure of the enriched text

The accurate retrieval system described here uses the so-called enriched text as search target, in this case the *Msimulizi* text. The enriched text is produced so that the normal text is passed through the analysis and morphological disambiguation, and the result contains morphological interpretation for each morpheme of the word. In the case of verbs, the interpretation may be quite complex due to several consecutive morphemes.

In the final enriched text, most morphological information is removed. Only the word lemma and its POS class code are retained, and all the rest is removed. As a result, the enriched text is clearly structured, because after each word form comes the lemma of that

word and its POS tag. Such text form is also easy to convert into more readable form in information retrieving phase.

The original text in *Msimulizi* is located under various headings. Most headings include the name of the mission station, such asd HABARI ZA MISOZWE (News from Misozwe), but also topic names, such as HABARI ZA VITA (War news) are included. In the edited and modified version, each sentence is on its own line. In the beginning of the sentence, there is an identification code area of eight spaces. The identification code has the form 'NYA-88-', where the first part means NJASSA and th last part the year 1888.

The identification code has two functions. First, it helps in retrieving material on a certain subject (in this case news from Njassa area), or from a certain year. Second, the code can be produced in each search result, if needed.

The list of topic names and their abbreviations are in (1).

(1)

BOO - BOONDE
chi - CHISUMULU
CHI - CHITANGALI
CHU - CHUAKA
HAD - HADITHI
KIC - KICHELWE
KIL - KILIMANI
KIU - KIUNGANI
KOR - KOROGWE
KOT - KOTAKOTA
KUF - KUFARIKI
LIK - LIKOMA
MAC - MACHEMBA
MAG - MAGILA
MAS - MASASI
MBW - MBWENI
MIS - MISOZWE
MIW - MIWA
MKU - MKUNAZINI
mku - MKUZI
MLI - MLIMA WA MLINGA
CHI - MLOLELA
MSA - MSARAKA
MSU - MSUMBA
MWI - MWITI
NEW - NEWALA
NAN - NANYANGA
NYA - NYASSA
PAC - PACHIA
PWA - PWANI
SAF - SAFARI
SHA - SHAMBA LA MBWENI
UGA - UGANDA
UNA - UNANGU
VIT - VITA
vit - VITENDAWILI
WAR - WARENO
YER - YERUSALEMI

In case the section of *Msimulizi* does not fall in any of those categories, the category name MSI (MSIMULIZI) is used. Therefore, each sentence has the two-part identification code.

An extract from the enriched *Msimulizi* text is in (2).

(2)

```
KIU-88- Lakini {lakini_CONJ} zamani {zamani_ADV} hakutaka {taka_V} hatta
{hatta_ADV} kidogo {kidogo_AD-ADJ} , {,_COMMA} sasa {sasa_ADV} anataka
{taka_V} sana {sana_AD-ADJ} .
KIU-88- Katika {katika_PREP} Oct. {Oct._PROPN} 4 {4_NUM} tulimsafirisha
{safirisha_V} Christopher {Christopher_PROPN} Kiwapo {wapo_V} mwalimu
{mwalimu_N} wa {wa_GEN-CON} Kiungani {kiunga_N} kwenda {kwenda_V} Newala
{Newala_N} kufundisha {fundisha_N} kule {kule_PRON} .
```

```
KIU-88- Amefuatana {fuatana_V} na {na_PREP} Charles {Charles_PROPN} W.
{W._N} Rehani {rehani_N} , {,_COMMA} mpishi {mpishi_N} , {,_COMMA}
aliyemwoa {oa_V} siku_hizi {siku_hizi_ADV} Hope {Hope_PROPN} Chela
{Chela_PROPN} , {,_COMMA} Mbweni {mbwe_N} .
KIU-88- Twasikia {sikia_V} wameshuka {shuka_V} salama {salama_ADV} Lindi
{Lindi_N} .
KIU-88- Tunatumaini {tumaini_V} tutaingia {ingia_V} Kanisa {kanisa_N}
letu {etu_PRON} jipya {pya_ADJ} mwezi {mwezi_N} ukiandama {andama_V} .
KIU-88- Bwana {Bwana_N} Askofu {askofu_N} ameleta {leta_V} simu {simu_N}
kusema {sema_V} atafika {fika_V} Unguja {Unguja_N} Oct. {Oct._PROPN} 31
{31_NUM} .
```

The conversion of the *Msimulizi* text into sentence-per-line format was challenging, because the punctuation was non-systematic, and often the normal full stop at the end of the sentence was replaced with a comma or semicolon, or even with an empty space. Full stops after month days also disturbed the process, and a lot of post editing was needed for correcting wrong sentence splitting. But this was due to the non-conventional use of punctuation marks, and this should not be found in most texts.

4 Search methods

The enriched form of text makes it possible to direct search on the surface form or the analyzed form of the word. Both search methods are needed in practice. However, considering the complex word structures in Swahili, the possibility of using the analyzed form as a search key is useful.

Consider the search task, where you want to find information on deaths in *Msimulizi*. Obvious search keys are words, which mean death or dieing. In Swahili, such words are the verbs *kufariki* and *kufa*, as well as the noun *kifo* (pl. *vifo*). It appears that the noun *kifo* does not appear at all. The verb *kufa* appears 164 times, and the verb *kufariki* 90 times.

Especially the verb *kufa* is problematic, because its only constant element is the stem *f* and all the rest is inflection. We can see all the inflected forms of this verb in (3).

(3)

```
 2 Alikufa [fa_V]
 1 Alipokufa [fa_V]
 1 Kikafa [fa_V]
 3 afe [fa_V]
25 akafa [fa_V]
 2 akifa [fa_V]
11 alikufa [fa_V]
 2 alipokufa [fa_V]
 6 aliyekufa [fa_V]
34 amekufa [fa_V]
 1 angalikufa [fa_V]
 1 angekufa [fa_V]
 5 atakufa [fa_V]
 2 hafi [fa_V]
 1 hajafa [fa_V]
 3 hakufa [fa_V]
 1 hawakufa [fa_V]
 1 hufa [fa_V]
 1 iliokufa [fa_V]
 1 imekufa [fa_V]
 6 kafa [fa_V]
22 kufa [fa_V]
 5 nife [fa_V]
 1 nikifa [fa_V]
 1 ningekufa [fa_V]
 1 tukifa [fa_V]
 1 tulikufa [fa_V]
 2 tunakufa [fa_V]
 1 ufe [fa_V]
 8 vifao [fa_V]
 2 wafe [fa_V]
 9 wakafa [fa_V]
 3 waliokufa [fa_V]
10 wamekufa [fa_V]
```

```
 3 wanakufa [fa_V]
 1 wangekufa [fa_V]
 1 wasife [fa_V]
```

The list shows that without analysed text form the search task would be impossible. We can produce a similar list of the occurrences of the verb *kufariki* (4).

(4)

```
 2 Alifariki [fariki_V]
 1 Alipofariki [fariki_V]
 1 Amefariki [fariki_V]
23 akafariki [fariki_V]
 8 alifariki [fariki_V]
 1 aliofariki [fariki_V]
 1 alipofariki [fariki_V]
 1 alivyofariki [fariki_V]
 5 aliyefariki [fariki_V]
35 amefariki [fariki_V]
 5 kafariki [fariki_V]
 2 kufariki [fariki_V]
 6 waliofariki [fariki_V]
 3 wamefariki [fariki_V]
```

We see that the list for *kufariki* is much shorter, which means that less forms are constructed with this stem. This may be partly due to the fact that the basic verb is *kufa*, while the verb *kufariki* is a loan word from Arabic and its connotation is more solemn. Yet one could consider using the normal string search for finding the occurrences, because the stem is *fariki*, including also the final wovel, as loan word verbs do in Swahili.

We can retrieve also other monosyllabic verbs from *Msimulizi*. Such verbs include: *kula* (to eat), *kuwa* (to be), *kunya* (to rain), *kucha* (to rise), *kuja* (to come), and *kupa* (to give). These verbs have also extended forms, but we will concentrate here on base forms (5).

(5) kula - to eat

```
 2 Chala [la_V]
 2 Kapala [la_V]
 1 Kaule [la_V]
 2 Mpale [la_V]
 1 Mtakula [la_V]
17 Mtaula [la_V]
 1 Pala [la_V]
 1 Wakala [la_V]
 1 Zaila [la_V]
 5 akala [la_V]
 2 akila [la_V]
 4 alao [la_V]
 1 alikula [la_V]
 1 amekula [la_V]
 2 anakula [la_V]
 1 anawala [la_V]
 1 asile [la_V]
 4 hali [la_V]
 1 hamli [la_V]
 1 hatukula [la_V]
 1 hawakula [la_V]
 2 hawali [la_V]
 1 hazikula [la_V]
 6 hula [la_V]
 1 kuila [la_V]
35 kula [la_V]
 1 kuleni [la_V]
 2 kumla [la_V]
```

```
 2 kutawala [la_V]
 1 lala [la_V]
 1 laleni [la_V]
 1 nimekula [la_V]
 1 nitazila [la_V]
47 tukala [la_V]
 1 tukamla [la_V]
 1 tule [la_V]
 1 tulipokula [la_V]
 1 tumekula [la_V]
 1 tunakula [la_V]
 2 tutakula [la_V]
 1 twalikula [la_V]
 2 ulalo [la_V]
 2 unakula [la_V]
 1 vyakula [la_V]
 1 waile [la_V]
 1 wakaila [la_V]
15 wakala [la_V]
 1 wakila [la_V]
 1 walapo [la_V]
 3 wale [la_V]
 2 walikula [la_V]
 1 waliokula [la_V]
 2 wamekula [la_V]
 1 wanile [la_V]
 1 watakula [la_V]
 6 yale [la_V]
 1 zikala [la_V]
```

All the forms are theoretically legal forms of the verb *kula*, but there are some words, which are in fact proper names in text. Disambiguation had failed in those cases. Also such words as *wale*, *yale* and *lala* are obviously not forms of the verb *kula*. However, the wast majority of hits are correct.

(6)
kuwa - to be

```
 2 Akawa [wa_V]
 7 Alikuwa [wa_V]
 1 Alipokuwa [wa_V]
 1 Asiwe [wa_V]
 1 Hakuwa [wa_V]
 1 Hapakuwa [wa_V]
20 Ikawa [wa_V]
60 Ilikuwa [wa_V]
 3 Imekuwa [wa_V]
 2 Itakuwa [wa_V]
 1 Itakuwaje [wa_V]
 5 Kiwayo [wa_V]
 1 Lilikuwa [wa_V]
 1 Mwe [wa_V]
 2 Nalikuwa [wa_V]
 1 Niliokuwa [wa_V]
 1 Nilipokuwa [wa_V]
 6 Palikuwa [wa_V]
 4 Tukawa [wa_V]
 2 Tukiwa [wa_V]
 2 Tulikuwa [wa_V]
 1 Tulipokuwa [wa_V]
 2 Tuwe [wa_V]
 2 Ulikuwa [wa_V]
 1 Ulipokuwa [wa_V]
 2 Uwe [wa_V]
 3 Walikuwa [wa_V]
 3 Walipokuwa [wa_V]
 1 Yakuwa [wa_V]
 1 Yalikuwa [wa_V]
 1 ahilikuwa [wa_V]
 1 ahtulikuwa [wa_V]
48 akawa [wa_V]
24 akiwa [wa_V]
 2 alikokuwa [wa_V]
210 alikuwa [wa_V]
 2 alimokuwa [wa_V]
 5 aliokuwa [wa_V]
17 alipokuwa [wa_V]
 1 alivyokuwa [wa_V]
39 aliyekuwa [wa_V]
15 amekuwa [wa_V]
 2 anakuwa [wa_V]
 2 angalikuwa [wa_V]
 2 angekuwa [wa_V]
 2 asiwe [wa_V]
 1 atakayekuwa [wa_V]
 9 atakuwa [wa_V]
 9 awe [wa_V]
 3 haijawa [wa_V]
 3 haikuwa [wa_V]
 1 hajawa [wa_V]
```

```
  3 hakuwa [wa_V]
  2 halikuwa [wa_V]
  6 hapakuwa [wa_V]
  2 hatukuwa [wa_V]
  1 hatuwi [wa_V]
  1 hawakuwa [wa_V]
  1 hawatakuwa [wa_V]
  2 hayakuwa [wa_V]
  2 huwa [wa_V]
  2 ijapokuwa [wa_V]
  1 ikauwa [wa_V]
161 ikawa [wa_V]
  1 ikawaje [wa_V]
  1 ilikokuwa [wa_V]
292 ilikuwa [wa_V]
  1 iliokuwa [wa_V]
  8 ilipokuwa [wa_V]
  4 ilivyokuwa [wa_V]
  5 iliyokuwa [wa_V]
 20 imekuwa [wa_V]
  3 inakuwa [wa_V]
  2 ingalikuwa [wa_V]
  2 isiwe [wa_V]
  1 itakavyokuwa [wa_V]
 20 itakuwa [wa_V]
  2 itakuwaje [wa_V]
  1 iwayo [wa_V]
 12 iwe [wa_V]
  1 kawa [wa_V]
  2 kilichokuwa [wa_V]
  1 kilikuwa [wa_V]
  1 kilivyokuwa [wa_V]
  1 kimekuwa [wa_V]
  1 kinakuwa [wa_V]
  7 kulikuwa [wa_V]
  1 kulipokuwa [wa_V]
  5 kumekuwa [wa_V]
  1 kunakuwa [wa_V]
 50 kuwa [wa_V]
  1 likawa [wa_V]
  2 likiwa [wa_V]
 16 lilikuwa [wa_V]
  1 lilipokuwa [wa_V]
  4 limekuwa [wa_V]
  1 litakuwa [wa_V]
  1 liwalo [wa_V]
  1 liwe [wa_V]
  2 mkiwa [wa_V]
  1 mliokuwa [wa_V]
  1 mlivyokuwa [wa_V]
  1 mlizokuwa [wa_V]
  1 mtakuwa [wa_V]
  6 mwe [wa_V]
  7 nalikuwa [wa_V]
  8 nikawa [wa_V]
  2 nikiwa [wa_V]
  3 nilikuwa [wa_V]
  7 nilipokuwa [wa_V]
  1 niliyokuwa [wa_V]
  1 nimekuwa [wa_V]
  2 nitakuwa [wa_V]
  2 niwe [wa_V]
  1 pakawa [wa_V]
 29 palikuwa [wa_V]
  3 palipokuwa [wa_V]
  1 palivyokuwa [wa_V]
  2 pamekuwa [wa_V]
  1 pasiwe [wa_V]
  1 patakuwa [wa_V]
  2 sijawa [wa_V]
  1 sikuwa [wa_V]
  1 tukakuwa [wa_V]
 40 tukawa [wa_V]
  6 tukiwa [wa_V]
 36 tulikuwa [wa_V]
  2 tuliokuwa [wa_V]
 14 tulipokuwa [wa_V]
  2 tulivyokuwa [wa_V]
  2 tuliyokuwa [wa_V]
  4 tumekuwa [wa_V]
  2 tungalikuwa [wa_V]
  4 tutakuwa [wa_V]
  9 tuwe [wa_V]
 11 twalikuwa [wa_V]
 19 ukawa [wa_V]
  7 ukiwa [wa_V]
 37 ulikuwa [wa_V]
  4 uliokuwa [wa_V]
  3 ulivyokuwa [wa_V]
  1 umekuwa [wa_V]
  1 unakuwa [wa_V]
  1 utakavyokuwa [wa_V]
  1 utakuwa [wa_V]
 10 uwe [wa_V]
  1 uwepo [wa_V]
  2 vilikuwa [wa_V]
 14 wakawa [wa_V]
  9 wakiwa [wa_V]
 92 walikuwa [wa_V]
 25 waliokuwa [wa_V]
 12 walipokuwa [wa_V]
  4 walivyokuwa [wa_V]
```

```
 1 walizokuwa [wa_V]
12 wamekuwa [wa_V]
 1 wanakuwa [wa_V]
 1 watakaokuwa [wa_V]
 1 watakiwa [wa_V]
 7 watakuwa [wa_V]
 7 wawe [wa_V]
 8 yakawa [wa_V]
 5 yakuwa [wa_V]
21 yalikuwa [wa_V]
 1 yaliokuwa [wa_V]
 1 yalipokuwa [wa_V]
 1 yalivyokuwa [wa_V]
 2 yamekuwa [wa_V]
 2 yamekuwaje [wa_V]
 2 yatakavyokuwa [wa_V]
 1 yatakuwa [wa_V]
 2 yatakuwaje [wa_V]
 7 yawe [wa_V]
 5 zikawa [wa_V]
 1 zikiwa [wa_V]
16 zilikuwa [wa_V]
 2 zimekuwa [wa_V]
 1 zitakuwa [wa_V]
 1 ziwazo [wa_V]
 2 ziwe [wa_V]
```

(7)
kunya - to rain

```
 6 ikanya [nya_V]
 2 ikinya [nya_V]
 5 ilikunya [nya_V]
 2 ilipokunya [nya_V]
 1 imekunya [nya_V]
 1 inakunya [nya_V]
 1 inye [nya_V]
 1 itakunya [nya_V]
 2 kunya [nya_V]
 1 unyayo [nya_V]
 1 watawanya [nya_V]
 2 yanya [nya_V]
```

(8) kucha - to rise

```
 3 Kakucha [cha_V]
 1 Kukicha [cha_V]
 1 Mkacha [cha_V]
 1 acha [cha_V]
 2 ache [cha_V]
 3 kucha [cha_V]
 2 kulipokucha [cha_V]
 1 kunakucha [cha_V]
 1 mwacha [cha_V]
 4 wacha [cha_V]
 6 wache [cha_V]
 1 wacheni [cha_V]
```

(9)
kuja - to come

```
15 Akaja [ja_V]
 2 Akija [ja_V]
 7 Alikuja [ja_V]
 1 Alipokuja [ja_V]
 1 Amekuja [ja_V]
 1 Anakuja [ja_V]
 1 Ikaja [ja_V]
 2 Ilikuja [ja_V]
 3 Njoo [ja_V]
 1 Njooni [ja_V]
```

```
  1 Tukaja [ja_V]
  1 Tulikuja [ja_V]
  1 Tutakuja [ja_V]
  1 Ukija [ja_V]
  1 Ulikuja [ja_V]
  1 Vikija [ja_V]
  1 Wajao [ja_V]
  1 Waje [ja_V]
  6 Wakaja [ja_V]
  1 Wakija [ja_V]
  7 Walikuja [ja_V]
  1 Waliokuja [ja_V]
  2 Walipokuja [ja_V]
  1 Wamekuja [ja_V]
  6 aja [ja_V]
  8 ajapo [ja_V]
  8 ajaye [ja_V]
 27 aje [ja_V]
172 akaja [ja_V]
 18 akija [ja_V]
 58 alikuja [ja_V]
  2 aliokuja [ja_V]
 14 alipokuja [ja_V]
  9 aliyekuja [ja_V]
 56 amekuja [ja_V]
 30 anakuja [ja_V]
  2 anayekuja [ja_V]
  1 angekuja [ja_V]
  8 asije [ja_V]
  1 asiyekuja [ja_V]
  1 atakayekuja [ja_V]
 18 atakuja [ja_V]
  1 bwanatwaja [ja_V]
  2 haijaja [ja_V]
  1 haiji [ja_V]
  3 haikuja [ja_V]
  2 haiwaji [ja_V]
  5 hajaja [ja_V]
  5 hakuja [ja_V]
  1 hamji [ja_V]
  1 hamngekuja [ja_V]
  1 hatukuja [ja_V]
  1 havikuja [ja_V]
  4 hawajaja [ja_V]
  4 hawaji [ja_V]
 14 hawakuja [ja_V]
  1 hawangalikuja [ja_V]
  1 hazijaja [ja_V]
 34 huja [ja_V]
  1 hukuja [ja_V]
  2 ijapo [ja_V]
  2 ijayo [ja_V]
  2 ije [ja_V]
 17 ikaja [ja_V]
  3 ikija [ja_V]
 15 ilikuja [ja_V]
  5 ilipokuja [ja_V]
  1 iliyokuja [ja_V]
 12 imekuja [ja_V]
 16 inakuja [ja_V]
  2 isije [ja_V]
  1 itakuja [ja_V]
 13 kaja [ja_V]
  1 kijacho [ja_V]
  3 kikaja [ja_V]
  1 kinakuja [ja_V]
128 kuja [ja_V]
  3 kumekuja [ja_V]
  1 kunakuja [ja_V]
  5 likaja [ja_V]
  2 linakuja [ja_V]
  1 linapokuja [ja_V]
  1 litakalokuja [ja_V]
  4 mje [ja_V]
  1 mlipokuja [ja_V]
  1 mlivyokuja [ja_V]
  2 mmekuja [ja_V]
  1 msije [ja_V]
  1 mwaja [ja_V]
  7 naja [ja_V]
  2 nakuja [ja_V]
  1 nalikuja [ja_V]
  2 nije [ja_V]
  5 nikaja [ja_V]
  1 nikija [ja_V]
  4 nimekuja [ja_V]
  2 ninakuja [ja_V]
  1 ninaokuja [ja_V]
  2 nitakuja [ja_V]
  6 njoo [ja_V]
  4 njooni [ja_V]
  3 palikuja [ja_V]
  2 pamekuja [ja_V]
  1 sikuja [ja_V]
  2 tuje [ja_V]
 21 tukaja [ja_V]
  1 tukaje [ja_V]
  1 tukija [ja_V]
  1 tulikuja [ja_V]
  1 tulipokuja [ja_V]
 10 tumekuja [ja_V]
  4 tunakuja [ja_V]
```

```
  2 tutakuja [ja_V]
  1 twaja [ja_V]
  3 ujao [ja_V]
  3 uje [ja_V]
  4 ukaja [ja_V]
  3 ukija [ja_V]
  1 ulikuja [ja_V]
  2 ulipokuja [ja_V]
  2 umekuja [ja_V]
  1 unakuja [ja_V]
  1 unapokuja [ja_V]
  1 ungekuja [ja_V]
  1 utakuja [ja_V]
  2 vikaja [ja_V]
  1 vikija [ja_V]
  1 vilikuja [ja_V]
 23 wajao [ja_V]
  3 wajapo [ja_V]
 22 waje [ja_V]
157 wakaja [ja_V]
 19 wakija [ja_V]
 95 walikuja [ja_V]
 26 waliokuja [ja_V]
  5 walipokuja [ja_V]
 34 wamekuja [ja_V]
 26 wanakuja [ja_V]
  6 wanaokuja [ja_V]
  1 wanavyokuja [ja_V]
  1 wangekuja [ja_V]
  4 wasije [ja_V]
  1 wasipokuja [ja_V]
  1 watakaokuja [ja_V]
  9 watakuja [ja_V]
  3 yaja [ja_V]
  1 yajayo [ja_V]
  1 yaje [ja_V]
  1 yakaja [ja_V]
  1 zije [ja_V]
  4 zikaja [ja_V]
  1 zilikuja [ja_V]
  1 zilipokuja [ja_V]
  3 zimekuja [ja_V]
  1 zinakuja [ja_V]
```

Note that the list also includes such exceptional cases as the imperative forms *njoo* (come, sg) and *njooni* (come, pl).

(11)
kupa - to give

```
  2 Akampa [pa_V]
  1 Akanipa [pa_V]
  2 Nikawapa [pa_V]
  2 Nipe [pa_V]
  1 Nipeni [pa_V]
  2 Tupe [pa_V]
  1 Tupeni [pa_V]
  1 Wakawapa [pa_V]
  1 Walimpa [pa_V]
 17 akampa [pa_V]
  6 akanipa [pa_V]
 16 akawapa [pa_V]
  3 alimpa [pa_V]
  1 alinipa [pa_V]
  2 aliwapa [pa_V]
  1 aliyempa [pa_V]
  1 aliyenipa [pa_V]
  1 aliyewapa [pa_V]
  1 alizompa [pa_V]
  1 amempa [pa_V]
  1 amewapa [pa_V]
  4 anipe [pa_V]
  1 atawapa [pa_V]
  1 atupaye [pa_V]
  1 avipaye [pa_V]
  1 awape [pa_V]
  1 haikumpa [pa_V]
  1 hakuwapa [pa_V]
  1 hamnipi [pa_V]
  1 hapi [pa_V]
  1 hauwapi [pa_V]
  1 hawakunipa [pa_V]
  1 hawawapi [pa_V]
  1 hukunipa [pa_V]
  4 humpa [pa_V]
  1 hunipi [pa_V]
  2 huwapa [pa_V]
  1 kampa [pa_V]
  1 kawapa [pa_V]
  2 kukupa [pa_V]
  1 kulipa [pa_V]
 10 kumpa [pa_V]
  1 kunipa [pa_V]
  1 kutupa [pa_V]
  7 kuwapa [pa_V]
  1 kuwapeni [pa_V]
```

```
1 mkinipa [pa_V]
2 mnipe [pa_V]
1 mpe [pa_V]
1 mpeni [pa_V]
1 mtampa [pa_V]
1 napa [pa_V]
1 nikakupeni [pa_V]
5 nikampa [pa_V]
4 nikawapa [pa_V]
1 nikiwapa [pa_V]
1 nikupe [pa_V]
2 nimekupa [pa_V]
2 nipe [pa_V]
6 nitakupa [pa_V]
1 nitakupeni [pa_V]
1 niwapazo [pa_V]
2 niwape [pa_V]
2 sikupeni [pa_V]
1 simpi [pa_V]
1 tukajipa [pa_V]
1 tukampa [pa_V]
2 tukiwapeni [pa_V]
1 tulipe [pa_V]
1 tumempa [pa_V]
2 tumpe [pa_V]
1 tupe [pa_V]
1 tupeni [pa_V]
1 tutampa [pa_V]
1 tutawapa [pa_V]
1 twakupa [pa_V]
3 unipe [pa_V]
1 utupe [pa_V]
8 wakampa [pa_V]
2 wakanipa [pa_V]
2 wakawapa [pa_V]
2 wakimpa [pa_V]
2 walimpa [pa_V]
1 walipe [pa_V]
1 walizompa [pa_V]
1 wamenipa [pa_V]
1 wampa [pa_V]
3 wampe [pa_V]
1 watakupa [pa_V]
1 watanipa [pa_V]
1 wawape [pa_V]
```

Above we have examples of the monosyllabic verbs in base form. They have also several extended forms, such as applicative, causative, reciprocal and passive forms. Below are some examples of these.

(12)
kufia - die on behalf of

```
2 Alifia [fia_V]
1 afie [fia_V]
2 akafia [fia_V]
2 alikofia [fia_V]
1 aliyefia [fia_V]
1 atawafia [fia_V]
1 hakufia [fia_V]
1 kufia [fia_V]
1 walifia [fia_V]
1 wamewafia [fia_V]
```

(13)
kupeana - to give to each other

```
2 kupeana [peana_V]
1 wakipeana [peana_V]
1 wapeane [peana_V]
```

(14)
kupewa - to be given (that is, to get)

```
1 Akapewa [pewa_V]
1 Apewa [pewa_V]
```

```
 1 Apewaje [pewa_V]
 5 Tukapewa [pewa_V]
 1 Tulipewa [pewa_V]
 1 Wakapewa [pewa_V]
 1 Walipewa [pewa_V]
 1 Wamepewa [pewa_V]
17 akapewa [pewa_V]
 2 akipewa [pewa_V]
 1 alilopewa [pewa_V]
 1 aliopewa [pewa_V]
 8 alipewa [pewa_V]
 1 alipopewa [pewa_V]
 6 amepewa [pewa_V]
 1 anapewa [pewa_V]
 1 apewa [pewa_V]
 1 apewazo [pewa_V]
 1 asipopewa [pewa_V]
 3 atapewa [pewa_V]
 1 hawajapewa [pewa_V]
 1 hawakupewa [pewa_V]
 1 hawapewi [pewa_V]
 2 hupewa [pewa_V]
 2 kapewa [pewa_V]
16 kupewa [pewa_V]
 1 likapewa [pewa_V]
 1 lipewe [pewa_V]
 1 mmepewa [pewa_V]
 1 napewa [pewa_V]
 2 nikapewa [pewa_V]
 1 nililopewa [pewa_V]
 3 nimepewa [pewa_V]
 1 sikupewa [pewa_V]
34 tukapewa [pewa_V]
 2 tuliopewa [pewa_V]
13 tulipewa [pewa_V]
 1 tulivyopewa [pewa_V]
 1 tuliyopewa [pewa_V]
 2 tumepewa [pewa_V]
 2 tupewayo [pewa_V]
 1 ukipewa [pewa_V]
33 wakapewa [pewa_V]
 2 wakipewa [pewa_V]
 4 waliopewa [pewa_V]
18 walipewa [pewa_V]
 1 walipopewa [pewa_V]
 1 walizopewa [pewa_V]
13 wamepewa [pewa_V]
 2 wapewe [pewa_V]
 1 wasipewe [pewa_V]
```

(15)
kulisha - to make eat, that is: to feed

```
 1 Kalisha [lisha_V]
 1 hatukulishi [lisha_V]
 2 kuwalisha [lisha_V]
 1 wakatulisha [lisha_V]
 2 wanaonilisha [lisha_V]
 1 watanilisha [lisha_V]
```

5 Searching on the basis of the surface form

Traditionally, the surface string search was the only method in information retrieval. In isolating languages such as English, it still is a viable method. The situation becomes much more complex with heavily inflecting languages such as Swahili or Finnish. The above examples demonstrate this. The direct string search is viable also in Swahili, if we want to search on the basis of surface strings, be they whole words or part of words.

With the search method demonstrated here, the search key may be the beginning or end part of the word, or any part inside the word, or the whole word. The system finds the matches and surrounds the found key section of the word with square brackets. Examples are below.

(16)
Key word is: aliwa

```
NEW-96- Dec. 21 , Mwalimu G. Yohana Mpalila [aliwa]sili hapa toka Mwiti .
NEW-96- Dec. 22 , Mwalimu Augustino [aliwa]sili akitoka Miwa .
NEW-96- Assubuhi ya Uz[aliwa] tuliimba Carols tukizunguka mjini tukaingia
Kanisani saa a kwanza mpaka saa 4 kasoro dakika kumi .
NEW-96- Siku ile ya Uz[aliwa] watoto waliambiwa marks V zao wakapokea
haki zao .
NEW-96- Dec. 26 , Mwalimu Filipo [aliwa]leta watoto wake wapokee zawadi
zao walioshinda na wasioshinda pia .
NEW-96- Dec. 28 , Archdeacon Farler [aliwa]sili hapa pamoja_na Mwalimu
James Chigulu kwani yee ni mgonjwa wa_macho .
NEW-96- Jioni yake watu walikusanyika nyumbani kuonyeshwa sanamu za taa
Takatifu za Uz[aliwa] .
MAS-96- Bassi barua ilipofika pwani ndipo akatokea Bwana mkubwa wa Lindi
kuja kutaka mali yote waliotwaa kwa watu wa safari , na mali
yaliyotw[aliwa] katika vita hii ndio pembe , watumwa , na tumbako .
MAS-96- Dec. 23 , Padre Hugh na watu wake walio Wamasihiya na wanafunzi
wakaja kujiweka tayari kwa_siku kuu ya Uz[aliwa] .
MAS-96- Penyi Nangoo nikapotea njia , kwani wenzangu n[aliwa]acha mitini
wakila matunda , nikafika mpaka ndani huko kwenyi mashamba ya akina Satia
.
NYA-96- Bassi tukajaribu kuvunja rupia , ache tusumbuke , killa tuendapo
inakat[aliwa] kwa_sababu rupia zitumikazo huku zimepigwa_chapa , bassi
tukajaribu killa duka mwisho kwa_bahati tukavunja .
NYA-96- Tena karibu wazungu wetu wote walikata_tamaa walitaka kurudi kwao
Ulaya , illa Bwana Faulkner [aliwa]tuliza , akisema "Haithuru twendeni
kuko_huko Nyasa .
```

The result can be arranged also so that the key word is aligned, and a chosen amount of context on each side can be printed. In example (17), the context size is defined on the basis of the number of character slots. In this case the left context is 20 slots and the right context 30 slots.

(17)

```
NEW-96- u G. Yohana Mpalila [aliwa]sili hapa toka Mwiti .
NEW-96- , Mwalimu Augustino [aliwa]sili akitoka Miwa .
NEW-96-       Assubuhi ya Uz[aliwa] tuliimba Carols tukizu
NEW-96-       Siku ile ya Uz[aliwa] watoto waliambiwa mark
NEW-96- 26 , Mwalimu Filipo [aliwa]leta watoto wake wapoke
NEW-96- , Archdeacon Farler [aliwa]sili hapa pamoja_na Mwa
NEW-96- a taa Takatifu za Uz[aliwa] .
MAS-96- i , na mali yaliyotw[aliwa] katika vita hii ndio p
MAS-96- i kwa_siku kuu ya Uz[aliwa] .
MAS-96- a , kwani wenzangu n[aliwa]acha mitini wakila matu
NYA-96- illa tuendapo inakat[aliwa] kwa_sababu rupia zitum
NYA-96- illa Bwana Faulkner [aliwa]tuliza , akisema "Haith
```

The context can also be defined on the basis of the number of words on each side. In (18) the context is three words on each size. Also, separate punctuation marks are treated as words.

(18)
```
NEW-96-             G. Yohana Mpalila [aliwa]sili hapa toka Mwiti
NEW-96-           , Mwalimu Augustino [aliwa]sili akitoka Miwa .
NEW-96-                Assubuhi ya Uz[aliwa] tuliimba Carols tukizunguka
NEW-96-                Siku ile ya Uz[aliwa] watoto waliambiwa marks
NEW-96-              , Mwalimu Filipo [aliwa]leta watoto wake wapokee
NEW-96-           , Archdeacon Farler [aliwa]sili hapa pamoja_na
NEW-96-             taa Takatifu za Uz[aliwa] .
MAS-96-             , na mali yaliyotw[aliwa] katika vita hii
MAS-96-             kwa_siku kuu ya Uz[aliwa] .
MAS-96-             , kwani wenzangu n[aliwa]acha mitini wakila matu
NYA-96-         , killa tuendapo inakat[aliwa] kwa_sababu rupia
NYA-96-            illa Bwana Faulkner [aliwa]tuliza , akisema "Haithuru
```

We see in the examples above that the direct string search method is not ideal in this case, because hits include verb prefix combinations, verb stems or part of them, and parts of proper names. The examples serve merely as demonstration on how results can be presented in various ways.

We get more precise results when we use the analysed forms as search key.

6 Searching on the basis of analysis result

Keeping the context on both sides of the hit as three words, we test with two stems, *uzaliwa_N* and *zaliwa_V* (19-20).

(19)
Search key: uzaliwa_N (birth, that is Christmas)
```
CHU-89-      Alhamisi baada_ya Uzaliwa [uzaliwa_N] sisi Waalimu wa
MAS-89-     kuingia tunangojea Uzaliwa [uzaliwa_N] .
SHA-89-                kuu za Uzaliwa [uzaliwa_N] tunaona hapana pah
YER-89-S              i kuu ya Uzaliwa [uzaliwa_N] , 1888 .
YER-89-N           ana zote za Uzaliwa [uzaliwa_N] mnazoimba leo Mbwe
CHI-90-                hii ya Uzaliwa [uzaliwa_N] hatukupata kustare
CHI-90-         ndio karibu_na Uzaliwa [uzaliwa_N] naona Dec. 24
MAG-91-               yetu ya Uzaliwa [uzaliwa_N] ilistawi sana .
MIS-91-                kuu ya Uzaliwa [uzaliwa_N] tulikuwapo wote La
MAG-92-               yetu ya Uzaliwa [uzaliwa_N] ilitufurahisha san
MAG-92-                kuu ya Uzaliwa [uzaliwa_N] .
KOR-92-                kuu ya Uzaliwa [uzaliwa_N] wa Bwana wetu
KIU-93-          mosi mbele_ya Uzaliwa [uzaliwa_N] ikawa shughuli yet
CHI-94-                kuu ya Uzaliwa [uzaliwa_N] vyakula vimeisha m
CHI-94-                kuu ya Uzaliwa [uzaliwa_N] .
KIU-94-                kuu ya Uzaliwa [uzaliwa_N] tukapamba nyumba y
KIU-94-             nyimbo za Uzaliwa [uzaliwa_N] ( Christmas Carols
KIU-94-               Siku ya Uzaliwa [uzaliwa_N] alitusayidia kwa u
NEW-95-                kuu ya Uzaliwa [uzaliwa_N] ilikuwa sala kubwa
mku-95-             huku mpaka Uzaliwa [uzaliwa_N] , nami nalipata
MBW-96-              nyimbo za Uzaliwa [uzaliwa_N] .
MBW-96-                siku ya Uzaliwa [uzaliwa_N] wa Bwana wetu
KOR-96-             sikukuu ya Uzaliwa [uzaliwa_N] wa Bwana tukahamis
MWI-96-          'holidays' za Uzaliwa [uzaliwa_N] .
NEW-96-            Assubuhi ya Uzaliwa [uzaliwa_N] tuliimba Carols tu
```

```
NEW-96-                ile ya Uzaliwa [uzaliwa_N] watoto waliambiwa
NEW-96-           Takatifu za Uzaliwa [uzaliwa_N] .
MAS-96-                kuu ya Uzaliwa [uzaliwa_N] .
```

(20)
Search key: zaliwa_V (to be born)

```
YER-89-         usiku ule alipozaliwa [zaliwa_V] Mwana wa Daudi
YER-89-  huonyesha mahali alipozaliwa [zaliwa_V] Isa Masiya na
MBW-89-             , mtoto amezaliwa [zaliwa_V] Sept. 10 ,
MKU-90-      wa watumwa watakaozaliwa [zaliwa_V] sasa watakuwa huru
HAD-90-      watu wangine waliozaliwa [zaliwa_V] katika ustaarabu ,
SHA-91-            Alhamisi alizaliwa [zaliwa_V] mtoto kwa Alfred
CHI-92-        ndiyo siku aliyozaliwa [zaliwa_V] Sultani wao Madachi
NAN-92-     mtoto mchanga aliyezaliwa [zaliwa_V] kama siku themintas
MSI-93-         yakawa kama amezaliwa [zaliwa_V] nayo .
CHI-94-               27 , walizaliwa [zaliwa_V] watoto wawili Yusuf
NEW-95-        nzige hawa waliozaliwa [zaliwa_V] hapa , kwani
MBW-95-             Denys , alizaliwa [zaliwa_V] Feb. 18 ,
```

Note that when we use the base form as search key, the hit will be produced immediately after the surface form of the word.

7 Choosing source text

The identification code at the beginning of each sentence gives a possibility to define the search material. For example, we can extract all the sentences that deal with war in 1888 (21).

(21)

```
VIT-88- Wadoicha {mdoicha_N} wanaona {ona_V} fetheha {fetheha_N} sana
{sana_AD-ADJ} kama {kama_ADV} mashauri {shauri_N} yao {ao_PRON} yote
{ote_PRON} yamefahitika V .
VIT-88-  Sasa  {sasa_ADV}  wameshawishi  {shawishi_V}  Waingreza
{Waingreza_PROPN} huko {huko_ADV} Ulaya {Ulaya_N} kushariki {shariki_V}
kazi {kazi_N} ya {ya_GEN-CON} kuzuia {zuia_N} na {na_CC} kuwaathibu
{athibu_N} watu {mtu_N} wote {ote_PRON} wanaochukua {chukua_V} watumwa
{tumwa_V} baharini {bahari_N} , {,_COMMA} na {na_CC} wanaochukua
{chukua_V} bunduki {bunduki_N} ao {ao_N} baruti {baruti_N} .
VIT-88- Bassi {bassi_ADV} sasa {sasa_ADV} wanaanza {anza_V} kuvizia
{zia_V} pwani {pwani_ADV} yote {ote_PRON} , {,_COMMA} Wadoicha
{mdoicha_N} wanavizia {zia_V} tangu {tangu_PREP} Tanga {Tanga_N} hatta
{hatta_ADV} kisiwa {kisiwa_N} Mafya {jifya_N} na {na_CC} Waingreza
{waingreza_N} tangu {tangu_PREP} Mafya {jifya_N} hatta {hatta_ADV} Rovuma
{Rovuma_PROPN} , {,_COMMA} na {na_CC} tangu {tangu_PREP} Tanga {Tanga_N}
hatta {hatta_ADV} Lamu {Lamu_N} .
VIT-88- Merikebu {merikebu_N} nyingi {ingi_PRON} zinavizia {zia_V} ,
{,_COMMA} na {na_CC} merikebu {merikebu_N} moja {moja_NUM} kubwa
{kubwa_ADJ} sana {sana_AD-ADJ} za {za_GEN-CON} chuma {chuma_N} tupu
{tupu_ADJ} , {,_COMMA} nene {nene_ADJ} kabisa {kabisa_ADV} , {,_COMMA}
inakaa {kaa_V} sikuzote {sikuzote_ADV} bandarini {bandari_N} hapa
{hapa_ADV} Unguja {Unguja_N} ina {ina_V} ngome {ngome_N} mbili
{mbili_NUM} za {za_GEN-CON} chuma {chuma_N} tupu {tupu_ADJ} juu_yake
{juu_ake_ADV} , {,_COMMA} mnamo {mnamo_PREP} mizinga {mzinga_N} ya
{ya_GEN-CON} ajabu {ajabu_N} mno {mno_AD-ADJ} ya {ya_GEN-CON} kupigia
```

```
{pigia_N} adui {adui_N} akikaa {kaa_V} mbali {mbali_ADV} mayili
{mayili_N} tano {tano_NUM} .
```

The number of reports on war varies greatly between years. In (22) are statistics on the number of sentences each year.

(22)
1888 - 4
1889 - 93
1890 - 3
1891 - 0
1892 - 22
1893 - 0
1894 - 0
1895 - 0
1896 - 0

However, these statistics do not give the full picture of the situation, because the reports take into account only those cases, where a specific title on war is present. War reports are also in many other reports. For this reason, we must take another approach. We serch for information using keys based on actual words in text. Obvious key words are *vita* (war), *kupigana* (to fight), *bunduki* (gun), *mkuki* (spear), *mshale* (arrow). In order to save space, I will give only statistics on each word in *Msimulizi* (23).

(23)
Key word: *vita_N* (war) - 273
pigana_V (fight - 93
bunduki_N (gun) - 170
mkuki_N (spear) - 9
mshale_N (arrow) - 7
uthia_N (threat) - 120

The words *vita*, *pigana* and *uthia* appear quite often each year. This gives a more reliable picture of the situation than the picture given by the titles alone. The noun *bunduki* is not a very reliable indicator of war, because *bunduki* was very much used as salute to welcome an honoured guest. *Buduki* was so recklessly used that death accidents in welcome ceremonies were reported.

We can direct the search to certain material. For example, we can retrieve the war-related material from Uganda (24). For saving space, context is contracted.

(24)

```
UGA-89-   Bassi Uganda sasa vita [vita_N] tupu wala_mwisho
UGA-90- shua ngine , walete vita [vita_N] huko Sese , laki
UGA-94-      Hapana hofu ya vita [vita_N] siku_hizi .
UGA-90- u , wakautwaa wakapigana [pigana_V] sana wakatwaa
UGA-90- saidia sana , tumepigana [pigana_V] na wale adui z
UGA-94- wa jamaa zetu tutapigana [pigana_V] marra ; kwani
UGA-94- ka inchi yetu , tupigane [pigana_V] tena wapate ku
```

```
UGA-89- enda barazani na bunduki [bunduki_N] zao tayari ,
UGA-89- alimpigia mfalme bunduki [bunduki_N] asimpate , la
UGA-90- karibia ikapigwa bunduki [bunduki_N] ya pwani ikap
UGA-89- ua wale wawili kwa mkuki [mkuki_N] wake , lakini m
UGA-89- awili aliowaua kwa mkuki [mkuki_N] wake , mmoja nd
UGA-89-        Bassi marra uthia [uthia_N] mtupu na machaf
UGA-90- esi Uganda , kwani uthia [uthia_N] uko wa Kamega ,
```

8 Use of non-standard words in Msimulizi

The analysed version of Msimulizi makes it also possible to study the use of non-standard words. A word is considered non-standard, when it is no longer used in standard Swahili, and the word is intentionally written in the form as it was printed. The non-standard element may be in the form of the word stem or in inflection. *Msimulizi* contains also clear typing errors, and such errors were corrected in editing phase. If the word was not in standard form and it was repeatedly written in the same way, it was considered a non-standard form of the word. All such non-standard words were included into the morphological analyzer, so that they could be retrieved in any word form.

In (25) is a list of non-standard words of *Msimulizi*. Only such words are listed, where the non-standard element is in the word stem.

(25)

```
     48 "adhuhuri" N
     20 "afathali" ADV
      1 "ahadia" V
     10 "alasili" N
      2 "angawa" CONJ
     29 "anjili" N
     17 "asharini" NUM
      1 "ashirini" NUM
     44 "asikari" N
      2 "athabu" N
      3 "athibiwa" V
      3 "athibu" V
      1 "athimishwa" V
      1 "athimiwa" V
     28 "balli" CONJ
     88 "balyozi" N
      3 "bandera" ADV
     43 "bandera" N
     29 "barozi" N
     49 "barra" N
     34 "bass" ADV
   1680 "bassi" ADV
     23 "billa" PREP
      1 "biskwiti" N
      4 "bithaa" N
     16 "boonde" N
      2 "borra" ADJ
     67 "buddi" N
     41 "burre" ADV
      2 "bwato" N
     10 "chayi" N
      2 "chukuwa" V
      9 "dayima" ADV
      2 "doktari" N
      1 "edaashara" NUM
    361 "enyi" POSS-PRON
      2 "faraga" N
      4 "fathaika" V
      5 "fathili" N
      2 "fathili" V
      1 "fathiliwa" V
     16 "fayida" N
      5 "fayidi" V
      2 "fayidia" V
      2 "fayidika" V
      2 "fayidiwa" V
     29 "fetha" N
      2 "fetheha" N
      1 "fithuli" N
      6 "forotha" N
     14 "fortha" N
     27 "frasi" N
      9 "fullani" ADJ
      2 "fungasa" V
      2 "ghasiya" N
      6 "ghathabu" N
```

```
 125 "ginsi" N
   3 "hassa" ADV
   3 "hatiya" N
1474 "hatta" ADV
   6 "hayi" ADJ
   3 "hifathi" V
   4 "ilimu" N
 327 "illa" CONJ
   8 "illakini" CONJ
  55 "illi" CONJ
 137 "ingine" ADJ
   1 "itlafu" N
   2 "jaisha" V
   4 "jihathari" V
   3 "juwa" V
   2 "kabithi" V
   2 "kabithiwa" V
   1 "kasikazini" ADV
   2 "kassa" ADV
   9 "kassi" ADV
   5 "kassorobo" ADV
   2 "kassrobo" ADV
   1 "katha" ADJ
  19 "kathalika" ADV
  14 "kathawakatha" ADJ
   4 "kathi" N
   6 "khutubu" V
   5 "kidoicha" N
 438 "killa" ADJ-PRE
   8 "kinwa" N
  34 "kitwa" N
   2 "kuaheri" ADV
   2 "kuaherini" ADV
   5 "kwiba" V
  16 "kwitwa" V
   6 "magaribi" ADV
   4 "malkiya" N
   4 "mangaribi" N
  32 "mangine" N
  23 "manovari" N
 527 "marra" ADV
  10 "mathabahu" N
  18 "mathbahu" N
  18 "mayili" N
   2 "mayiti" N
  59 "mdoicha" N
   1 "mdoichi" N
  17 "mgine" PRON
  25 "mhadi" N
   1 "mharamia" N
   1 "mjermani" N
   2 "mkohani" N
   6 "mngine" ADJ
   3 "mosikiti" N
   4 "moskiti" N
  47 "mpila" N
   2 "msayidia" N
   6 "mshujaa" N
   6 "msikiaji" N
   1 "mudda" N
   2 "mwaanafunzi" N
  14 "nahotha" N
  16 "nakhotha" N
   1 "nassi" ADV
  78 "naswi" ADV
   2 "nduwi" N
   9 "nuss" ADJ-PRE
   1 "nussra" ADJ-PRE
 102 "nussu" ADJ-PRE
   1 "ogolea" V
   4 "pilao" N
   1 "polomoka" V
   4 "raiya" N
   2 "rassi" N
  37 "rathi" N
   6 "rayia" N
   1 "rayiya" N
   2 "rhotuba" N
   1 "ridio" N
   3 "rithi" V
   1 "rithiwa" V
   4 "saalamu" N
   2 "sabwini" NUM
   3 "salaama" N
   6 "salam" N
   1 "sayidia" N
 142 "sayidia" V
   3 "sayidiana" V
  12 "sayidiwa" V
  24 "selaha" N
  15 "sermala" N
   6 "settini" NUM
  48 "shangwi" N
   5 "sharbati" N
   1 "shariki" N
  37 "shariki" V
   4 "sharikisha" V
   1 "sharikishwa" V
   6 "shauko" N
   2 "sherbeti" N
  63 "shidda" N
   1 "sigileti" N
```

```
  1 "siswi" PRON
  1 "sitaashara" NUM
  1 "sittini" NUM
 11 "sumulia" V
  2 "suruwali" N
  7 "swafi" ADJ
  4 "tafathali" ADV
  3 "tafathalini" ADV
  6 "taraja" V
 12 "thahabu" N
 34 "thaifu" ADJ
  1 "thalathini" NUM
  6 "thambi" N
  1 "thaminia" V
120 "thani" V
 15 "thania" V
  6 "thaniwa" V
  5 "tharau" V
  3 "tharuba" N
 11 "thawabu" N
  2 "thehebu" N
  4 "thelitashara" NUM
  4 "thoofika" V
  2 "thoofisha" V
  2 "thoruba" N
  1 "thulumiwa" V
  4 "thulumu" V
 25 "thuru" V
 26 "tissa" NUM
  1 "tissini" NUM
  1 "torumpeta" N
  1 "towa" V
  5 "tufano" N
  1 "tufanu" N
  1 "uhayi" N
 11 "unuss" ADJ-PRE
123 "unussu" ADJ-PRE
  1 "uswafi" N
  3 "uthaifu" N
  1 "uthalimu" N
 10 "uthi" V
120 "uthia" V
 13 "uthika" V
  1 "uthiwa" V
  2 "uwa" V
  1 "vunjilia" V
  6 "walla" ADV
 83 "wallakini" CONJ
419 "wangine" N
  8 "wangineo" N
  7 "yayi" N
138 "zayidi" AD-ADJ
  1 "zowea" V
  1 "zowelea" V
  3 "zuwia" V
```

Many words are classified as non-standard because of the orthography. Typical examples are such consonants that are written with digraphs, such as *dh* > *th* (e.g. thoruba). Between two vowels is often added the half-consonant w (e.g. *ua* > *uwa*, *saidia* > *sayidia*). Doubling of the consonant occurs after syllables with stress (*basi* > *bassi*, *bali* > *balli*, *ila* > *illa*, *nusu* > *nussu*, *tisa* > *tissa*).

Non-standard morphemes, which do not feature in the above list, include such prefixes of verbs as *tuli* > *twali*, *mli* > *mwali*, *zili* > *zali*, *nita* > *nta*, *na* > *ha*. The prefixes include the subject marker and the TAM marker. Examples of verb forms are in (26).

(26)

```
[twali]penda
[twali]shinda
[twali]wapokea
[twali]watazamia
[twali]kaa
[twali]kusanyika
[twali]cheza
[twali]karibishwa
[mwali]nitendeaje
[zali]ungua
[zali]fanya
```

```
[zali]batilika
[zali]tujia
[zali]tazamiwa
[nta]jaribu
[nta]sahau
[nta]uvunjilia
[nta]chagua
```

9 Discussion

Archives contain lots of printed material, which is immensely valuable for many kinds of purposes. Computers offer means for permanent preservation of this material. The material is normally preserved in bitmap form, which itself is merely another way to make the material readable for human beings. Digital form would allow kwick, accurate and comprehensive information retrieval, if text would be machine-readable. The OCR is a step forward, but unfortunately it works properly only, if the original text is clean. Therefore, in practice the OCR-read text must be corrected manually. Even this task is not easy, if the original text is difficult to read.

It is possible to construct post-editing programs, which do corrections automatically. However, only part of corrections can be done in this way without access to context. The only fully reliable method is to go the whole text through manually.

Manual correcting is often considered expensive, which is why cheaper methods have been sought for. We must consider the benefits of proper editing and weight them against costs. If the text to be edited, such as *Msimulizi*, has permanent value as source material, manual editing is not too expensive. Therefore, I have done it for *Msimulizi*. There are many more historical materials, which would deserve similar treatment as Msimulizi.

When the text is edited, it can be converted into the form suitable for computational treatment without much effort. This text form can then be processed into rich text format. Searches can then be made to this text format. The method enhances vastly more accurate and comprehensive search methods than traditional surface text search.