Investigationes Linguisticae, vol. XI, Poznań, December 2004

Unifying Electronic Dictionaries of Swahili into the DICT Format

Beata Wójtowicz

Warsaw University
ul. Krakowskie Przedmieście 26/28, 00-927 Warszawa

wierzchob@wp.pl

Abstract

Electronic Swahili dictionaries differ in their form, format and content. Due to legal aspects, it may be impossible to consolidate them into a single one, however it is feasible to access them in a uniform way. In the present paper, the use of TEI and DICT format for this purpose is discussed.

1. The Language

Swahili is the most widely spoken African language, with nearly 80 million speakers in East and Central Africa. The language has ancient roots in the Bantu language family. It is spoken particularly in Tanzania and Kenya, where it serves not only as national, but also as an official language. Swahili also acts as a vehicular[1] language throughout wide areas of Africa.

Spoken by quite large populations, Swahili is of considerable importance in the global context. Therefore, it is taught in main universities all over the world.

2. Swahili Language Resources

Apart from printed resources such as grammars, books and dictionaries,[2] Swahili also has some resources available on the Internet, which constitute a great variety of available knowledge about the language. It is possible to find several electronic Swahili dictionaries that were presented in detail in [5]. Two of them will be described more closely in the next section.

The main archives of Swahili, known as Helsinki Corpus of Swahili, are available on the Helsinki University Language Corpus Server[3]. The corpus consists of approximately 15 million words and contains a large collection of documents such as Swahili dialects, prose text from books in standard Swahili, religious texts, newspapers and radio broadcasts. The collection is still extending as new texts from the Internet are added. The access to the corpus is by individual agreement. The exploitation of those archives is facilitated by a number of specific tools. Since recently SALAMA[4] (Swahili Language Manager) is available. It makes use of language analysis

[1] I.e., it is used for non-official communication between members of different language groups.
[2] See, e.g., http://www.vessella.it/biblioswa.htm and http://www.hit.uib.no/AcoHum/nel/NEL-chapter-final.html
[3] http://www.aakkl.helsinki.fi/cameel/corpus/intro.htm
[4] http://www.aakkl.helsinki.fi/cameel/corpus/salamainfo.htm

that allows for more accurate and comprehensive search than it was possible with standard string search.

As an example of freely available Internet resources we can mention newspapers such as Nipashe (http://nipashe.netfirms.com), Majira (http://bcstimes.com/majira), Alasiri (http://ippmedia.com/alasiri.htm), and radio broadcasts such as Voice of America (http://www.voa.gov/swahili/index.html), BBC (http://www.bbc.co.uk/swahili), and Deutsche Welle (http://www.dwelle.de/kiswahili). Furthermore also some pieces of Swahili literature and translations of religious texts are accessible (http://www.pscw.uva.nl/lpca).

Work carried on in the field of computational linguistics enriched Swahili resources by software like AINI - morphological parser for Swahili[1], SWATWOL - morphological analyser, SWACGP[2] – morphological disambiguator and syntactic mapper, and a simple analyser presented in [1]. Simple analyser is also included as an integrated part of one of the dictionaries that will be described below.

2.1. Electronic Dictionaries

In this section two electronic dictionaries of Swahili will be presented. The interesting issue is to what extent it would be possible to integrate them. In the attempt of unification technical as well as legal problems have to be taken into consideration.

The following comparison has been inspired by [4].

1.1.1. Swahili-English Fried Dictionary

Swahili-English Dictionary available at the site of the Dict Development Group[3] was created in 1997 by Morris Fried from Lincoln University as a *Swahili-Kiswahili[4] to English Translation Program*.

It is a free MS Windows program designed for offline use and distributed under GNU General Public License. Accompanying documentation is rudimentary. The dictionary consists of 1428 entries. Due to build-in morphological analyser, we can search also for inflected forms, which is a great advantage. The only source language is Swahili, as a search result we get English equivalent with additional grammatical information, part of speech and e.g. about mood or tense for verbs:

```
pata
get, Get!
verb - root, imperative

amepata
he, she has gotten
verb - perfect tense
subject: third-person singular

mlango
door
noun
```

The dictionary has been created in 1997 and since then it hasn't been updated.

[1] AINI was created at the Department of African Linguistics, State University at Leiden, The Netherlands.
[2] Both SWATWOL and SWACGP have been developed at the Institute for Asian and African Studies, Helsinki University, Finland.
[3] http://www.dict.org/links.html
[4] *Kiswahili* is a native name for the language.

2.1.1. Ergane/FreeDict Dictionary

Ergane[1] is a multilingual dictionary programme that uses Esperanto as an auxiliary language to translate single words and short expressions from one language to another. The software is available for free, under GNU General Public License, for offline use in MS Windows environment. It consists of set of dictionaries and Esperanto is used as an intermediate language for translating from one natural language to another. A user manual *Ergane Help* is available together with some additional information about specific language database. Ergane supports many languages, among others English and Swahili and on those basis we get English-Swahili-English dictionary. Swahili part consists of 665 entries (without verbs) which compose a *Traveller's Dictionary*.

The program offers a dictionary and a practice utility. Search interface enables the user to search words, but they must be entered in the exact form, i.e., identical to that in the database. As a result we get two equivalents – an Esperanto and a target language translation without neither additional information nor examples:

```
mlango
  pordo
door
```

This dictionary, with many others, has been used by the FreeDict Project[2] and has been converted in 2000 by Horst Eyermann into Swahili-English dictionary in the so-called DICT format. DICT stands for Dictionary Server Protocol[3] and is an Internet protocol that allows a client to access dictionary definitions from a set of natural language dictionary databases. DICT was developed as a format for accessing monolingual dictionaries of English but later many bilingual dictionaries have been also converted into it. FreeDict dictionaries are maintained with CVS[4] (Concurrent Versions System). Dictionaries in this format are additionally distributed as packages of various Linux distributions, and the Swahili-English dictionary is available as a package of latest Debian unstable distribution.

3. The DICT Format

Available Swahili databases are not accessible via uniform interface, and they are not accessible from a single site. Some of them are small and incomplete individually, but would collectively provide an interesting and useful database of Swahili words. The solution to the dictionary database problem is offered, among others, by a dictionary server protocol DICT.

As it is stated in the RFC2229[5] the DICT protocol is designed to provide access to multiple databases. Word definitions can be requested, the word index can be searched (using an easily extended set of algorithms), information about the server (e.g., which index search strategies are supported, or which databases are available), and information about a database can be provided (e.g., copyright, source, or distribution information). Further, the DICT protocol has extensions that can be used to restrict access to some or all of the databases.

The reasons for choosing DICT as a competitive dictionary format may be as follows. The use of DICT offers many advantages. It is relatively easy, from the technical point of view, to create and maintain a database in that format. It gives the possibility of continuous development of a dictionary hosted on a server and, at the same time, having users working offline with its snapshots on their personal computers. It also enables dual access to resources, online or offline, thanks to various client software. Client software is able to search among several available

[1] http://travlang.com/Ergane
[2] http://www.freedict.de
[3] http://www.dict.org
[4] CVS is an open standard for version control.
[5] http://www.rfc-editor.org

dictionaries at the same time. Depending on client software, different query possibilities are available.

The DICT client software is, for example, a *Lookup* Emacs agent. It supports looking up definitions and searching for a match with a nice interface within Emacs environment. It enables looking up word definitions in all available dictionaries at one query, easy selection of dictionary and search strategy. The available query possibilities when searching for matching word are, e.g., exact, by prefix, suffix, substring, regular expression, keyword, text. The backward moving through the visited definitions is available. Additionally in the latest versions of GNU Emacs and XEmacs one can get support for popup menus[1], in GNU Emacs 21 and XEmacs 21 one can lookup words by simply pointing the mouse cursor to them (tool-tips). It also supports dictionaries, which are not encoded as UTF-8.

An interesting, already mentioned FreeDict is a project that supports the use of the bilingual databases with the DICT dictionary server. First databases were derived from Ergane, but since then many others have been added. At present, SourceForge[2] hosts most resources. The data are kept in XML format, complying with the TEI DTD[3] described in [3]. It enables including features such as phonetics, part of speech and etymology information in a project independent format.

4. Towards Unified Dictionary

Both of described Swahili dictionaries, Swahili-English Fried and FreeDict Swahili-English dictionary, are available under GNU General Public License what makes it possible to make changes to their content. Joining them into one requires, however, conversions into one specified format, as the formats of both dictionaries differ. The decision was to convert Swahili-English Fried Dictionary into the FreeDict dictionary DICT format.

After the conversion of the Swahili-English Dictionary to the source format of DICT dictionaries, entries from both dictionaries have merged. It resulted in a new Swahili-English dictionary with 1569 entries. Then it has been reverted into an English-Swahili part forming new *Swahili-English and English-Swahili xFried Dictionary* (*xFried* stands for *extended Fried*). Additionally part of speech information has been retained from the Swahili-English Fried Dictionary and added to the FreeDict Dictionary entries.

The program for formatting DICT dictionary databases, *dictfmt*, creates two files. One file, with the extension *dict*, contains the dictionary entries. The second is an index file. Together, they form an indexed database of headwords and translations.

The following is a sample from Swahili-EnglishxFried.dict file:

```
00-database-url
  http://www.mimuw.edu.pl/~jsbien/BW/Swa-Eng-xFried/
00-database-short
  Swahili-EnglishxFriedDictionary
00-database-info
  This file was converted from the original database on:
    Sat Feb 21 15:39:53 2004

The original data is available from:
  http://www.mimuw.edu.pl/~jsbien/BW/Swa-Eng-xFried/

This Swahili-English Dictionary is based on Swahili-Kiswahili to English
Translation Program by Morris Fried (www.dict.org/links.html), which has
been supplemented by entries from Freedict Swahili-English Dictionary
created by Horst Eyermann  (http://www.freedict.de) from the Swahili-
Esperanto and Esperanto-English Ergane dictionaries
(http://www.travlang.com).
```

[1] Described in e.g. [2].
[2] http://www.sourceforge.net
[3] http://www.tei-c.org

```
The conversion of Fried's dictionary to DICT format and merging it with
Eyermann's Ergane-based dictionary has been done, with some help of
Computer Science colleagues, by Beata Wójtowicz (wierzchob@wp.pl) in
2004.
This program is free software; you can redistribute it and/or modifyit
under the terms of the GNU General Public License as published by the
Free Software Foundation.

Swahili-English xFried Dictionary

Allah (n)
    God
abiria (n)
    passenger(s)
adhuhuri (n)
    midday
adui (n)
    enemy(-ies)
afisa (n)
    officer
afisi (n)
    bureau
Afrika (n)
    Africa
```

The sample from Swahili-EnglishxFried.index file is following:

```
00databasealphabet Kwk   4
00databaseinfo     B6    Iz
00databaseshort    BD    3
00databaseurl      A     BD
a                  Ixx   n
a huzuni           n1    0
a kenda            op    b
a kigeni           pE    u
a kumi             py    a
a kutosha          qM    q
a kuume            q2    m
a kwanza           rc    c
a moto             r4    Y
```

The DICT format has the drawback that the entry structure is not represented explicitly. In order to render the structure adequately while working with more complicated entries we can use an encoding standard. Using a markup scheme as TEI Guidelines secures the possibility of data migration to new, more sophisticated versions in the future.

The conversion of the *dict* file can be partialy done with a python script, *dict2tei.py*, available from the FreeDict servers at Sourceforge. The TEI header and part of speech tags need to be added.

The following example presents a sample from the Swahili-EnglishxFried.xml file:

```
<teiHeader>
 <fileDesc>
   <titleStmt>
     <title>Swahili-English xFried/FreeDict Dictionary</title>
       <respStmt>
         <resp>compiled by
          <name>Beata Wojtowicz, wierzchob@wp.pl</name></resp>
       </respStmt>
   </titleStmt>
   <publicationStmt>
```

```
      <publisher>FreeDict</publisher>
        <availability>
         <p>Available under the terms of the GNU General Public
            Licence.</p>
        </availability>
        <date>2004</date>
    </publicationStmt>
    <sourceDesc>
      <p>This file was compiled from other electronic  documents.
       It is based on Swahili-Kiswahili to English Translation
       Program  by  Morris  Fried  available  from
       http://www.dict.org/links.html,which has  been supplemented
       by entries from Freedict Swahili-English Dictionary created
       by Horst Eyermann (http://www.freedict.de) from the
       Swahili-Esperanto and Esperanto-English Ergane dictionaries,
       available from http://www.travlang.com .</p>
    </sourceDesc>
 </fileDesc>
    <encodingDesc>
      <projectDesc>
       <p>This dictionary comes to you trough nice people making it
        available for free and for good. It is part of the FreeDict
        project, http://www.freedict.de / http://freedict.org .</p>
       </projectDesc>
    </encodingDesc>
</teiHeader>

<text>
  <body>
    <div0 type='dictionary'>
<!-- BEGINNING OF A DICTIONARY -->

<entry>
  <form><orth>Allah</orth></form>
  <gramGrp> <pos>n</pos> </gramGrp>
  <def>God</def>
</entry>
<entry>
  <form><orth>abiria</orth></form>
  <gramGrp> <pos>n</pos> </gramGrp>
  <def>passenger(s)</def>
</entry>
<entry>
  <form><orth>adabu</orth></form>
  <gramGrp> <pos>n</pos> </gramGrp>
  <def>manners</def>
</entry>
```

TEI DTD complying XML files have been submitted recently to the FreeDict Project. Converted to the DICT format, they are now available from the Sourceforge site[1] for a free download and use with a DICT server.

The entries from the Swahili-EnglishxFried.dict file after conversion from the XML file are represented like following:

```
Allah <n.>
    God
abiria <n.>
    passenger(s)
adabu <n.>
```

[1] http://sourceforge.net/project/freedict

```
    manners
adhuhuri <n.>
    midday
```

All of the work was carried under Linux environment but tested under Microsoft Windows what in consequence resulted in a multiplatform compatible package.

4.1. Using Swahili-English and English-Swahili xFried/FreeDict Dictionary with Emacs

As mentioned, a *Lookup* Emacs agent is one of DICT clients. That means that multiple DICT dictionaries can be accessed in a unified way while working in Emacs environment. In this section different query possibilities available in *Lookup* will be presented.

In a *Lookup* window a list of available dictionaries is displayed, as in my example where three are available:

```
Type `m' to select, `u' to unselect, `?' for help.
% Identfier  Title                                      Method
- ---------  -----                                      ------
*eng-sw     English-Swahili xFried/FreeDict Dictionary  =<>-r@
*foldoc     The Free On-line Dictionary of Computing    =<>-r@
*sw-eng     Swahili-English xFried/FreeDict Dictionary  =<>-r@
```

After selecting a dictionary or dictionaries that we want to search through a different query possibilities are available under an *f* key on our keyboard. Below some different query examples are presented together with an output of the search.

```
1. 'word' - exact string search
'amini'
Swahili-English xFried/FreeDict Dictionary amini
amini <v.>
     believe

2. '*word' - sufix search
'*abu'
Swahili-English xFried/FreeDict Dictionary adabu
Swahili-English xFried/FreeDict Dictionary dhahabu
Swahili-English xFried/FreeDict Dictionary hesabu
Swahili-English xFried/FreeDict Dictionary kitabu
Swahili-English xFried/FreeDict Dictionary kwa sababu
Swahili-English xFried/FreeDict Dictionary mwarabu
Swahili-English xFried/FreeDict Dictionary sababu
Swahili-English xFried/FreeDict Dictionary uarabu
Swahili-English xFried/FreeDict Dictionary ulaya wa waarabu
adabu <n.>
     manners

3. 'word*' - prefix search
'maa*'
Swahili-English xFried/FreeDict Dictionary maalumu
Swahili-English xFried/FreeDict Dictionary maana
Swahili-English xFried/FreeDict Dictionary maarifa
maalumu <adj.>
     special

4. '*word*' - infix search
'*ghari*'
Swahili-English xFried/FreeDict Dictionary magharibi
magharibi <n.>
     west
```

```
5. @word - query inside entries
'@sham'
Swahili-English xFried/FreeDict Dictionary bahari ya sham
Swahili-English xFried/FreeDict Dictionary sham
```

Below, for comparison, is an exact string search:

```
'sham'
Swahili-English xFried/FreeDict Dictionary sham
```

5. Further Possibilities

Another freely available dictionary is a Student Swahili-Polish Dictionary,[1] created with the help of students from the Department of African Languages and Cultures at Warsaw University. Converting it into the DICT format would considerably expand the number of entries that could be searched with a single query.

Swahili texts available on the Internet can serve as Swahili corpus of the most temporary language but, of course, the copyright aspects should be taken into consideration. Nevertheless such resources could significantly help to improve available dictionaries, the choice of entries, equivalents and examples.

6. References

[1] Derzhanski, I., Nenova, I. 1997. *Prolog for linguists (in Bulgarian)*. Sofia: Intela.

[2] Schryver, G. M. 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, vol. 16, no 2, pp. 143-199.

[3] Sperberg-McQueen, C. M., Burnard, L. (eds.) 2002. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen

[4] Underwood, N., Navarretta, C. 1998. *A Draft Manual for the Validation of Lexica*. *Final Report*. Copenhagen: Center for Sprogteknologi.

[5] Wójtowicz, B. 2003. Dictionaries on the Web, a Case of Swahili. *Studies of the Department of African Languages and Cultures*, no 34, pp. 59-82.

[1] http://www.mimuw.edu.pl/~jsbien/BW/SSSP/SSSP.pdf