A repository of free lexical resources for African languages: the project and the method

Piotr Bański
Institute of English Studies,
University of Warsaw
E-mail: pkbanski@uw.edu.pl

Beata Wójtowicz
Dept. of African Languages and Cultures,
University of Warsaw
E-mail: b.wojtowicz@uw.edu.pl

Summary

- Our focus here is on FreeDict [http://www.freedict.org/], a project that has the potential to become home to, among others, free bilingual dictionaries for African languages. The project is part of SourceForge.net.
- The dictionaries can be usable even in their early versions, which can be subject to further supervised improvement as user feedback accumulates: "publish early, publish often", in the open-source way.
- We demonstrate a possible process of dictionary development on the example of one of FreeDict dictionaries – Swahili-English xFried/Freedict Dictionary – the first FreeDict dictionary encoded according to the TEI P5 XML standard.
- The final product can be accessed via desktop clients, via a Firefox add-on, or on the Web [http://dict.org].

DICT

- DICT (Dictionary Server Protocol; Faith and Martin 1997) is by now a well-established TCP-based query/response protocol that allows a client to access definitions from a set of various dictionary databases. It provides data in textual form, but it also has the potential of providing MIME-encoded content. The dictionary server software, dictd, is maintained and developed at SourceForge.
- The DICT format is a plain text format with an accompanying index file (an option of serving MIME content also exists).
- There is more than one way to query a DICT database: you can search the definitions and the headwords, using regex-based criteria.
- The clients can be free-standing desktop applications or they can be integrated into editors or web browsers. DICT web gateways also exist. The DICT project provides a list of clients and alternative servers.

viko[1] *n*
• Plural of kiko : pipe

viko[2] *cop loc unspec*
• are (in an unspecified place)
(agrees with cl. 8)

afisa (also **ofisa**) (pl: maafisa , maofisa) *n*
• officer

The screenshots above demonstrate the CSS "work view" of the dictionary (v. 0.4.1 of March 28th) and the way in which the Firefox add-on client presents query results (the mismatches are due to the incomplete support for TEI P5 in the FreeDict build system, originally designed for TEI P4; support for P5 got introduced only in mid-March).

Why Swahili-English?

- Just because we happen to be working on a Swahili-Polish-Swahili dictionary, and this is an offshoot of the testing phase of the project; we wanted to donate our test Swahili-English dictionary to FreeDict, and this is how the entire adventure began. This dictionary (in versions 0.3 and 0.4) replaced the earlier dictionary by the same name that Beata created from freely available GPL-ed sources.
- But our point is that any dictionary of any size can be submitted!

FreeDict

- FreeDict was founded in 2000 as an expression of the natural open-source synergy with DICT: DICT provided the platform for disseminating content of all kinds of dictionaries, while FreeDict grouped bilingual dictionaries that could be disseminated on this platform.
- Later on, FreeDict adopted the TEI P4 XML format. At the moment, it has also basic support for TEI P5 (in the CVS only; this is work in progress).

FreeDict is the nexus of the following:

- XML, with its potential for creating well-structured documents,
- TEI P5, an encoding standard taking advantage of this potential,
- the SourceForge repository as well as distribution and content-management network,
- the DICT distribution network: apart from being able to query DICT servers straight from the desktop, Firefox users can also take advantage of an add-on client that returns definitions for words highlighted on a web page (an example is shown to the left),
- FreeDict tools, as means to manipulate dictionaries and to create, among others, the DICT format (usable directly from DICT servers and by other dictionary-providing projects, e.g., StarDict or Open Dict); the build process provides targets for platforms other than DICT, e.g. the Evolutionary Dictionary or zbedic.

Additionally:

- Lexical resources submitted to FreeDict will be able to undergo further transformations, such as reversal or concatenation, which means that work put into developing a single resource may well be re-used in developing others.
- The project has its own distribution system, in the form of GNU/Linux packages.
- Content published by FreeDict is guaranteed to be free.

Below are the possible stages of development of an example entry; from the simplest glossary to something close to a machine-processable lexical database (we skip xml:lang attributes).

```
<entry>
   <form><orth>alasiri</orth></form>
   <def>afternoon</def>
</entry>
```

```
<entry xml:id="alasiri">
  <form><orth>alasiri</orth></form>
  <gramGrp><pos>n</pos></gramGrp>
  <sense>
      <def>afternoon (period between 3
           p.m. and 5 p.m.)</def>
  </sense>
</entry>
```

```
<entry xml:id="alasiri">
   <form type="N">
      <orth>alasiri</orth>
   </form>
   <gramGrp><pos>n</pos></gramGrp>
   <sense>
      <def>afternoon</def>
      <note type="def">period between
           3 p.m. and 5 p.m.</note>
   </sense>
</entry>
```

```
<entry xml:id="alasiri">
   <form type="N">
      <orth>alasiri</orth>
   </form>
   <gramGrp><pos>n</pos></gramGrp>
   <sense>
      <def>afternoon</def>
      <def>opponent <note type="hint">in
games or sports</note></def>
   </sense>
</entry>
```

Each of the above is transformable into a DICT-based dictionary, accessible locally or via the Internet.

alasiri [sg=pl] *n*
• afternoon
def: period between 3 p.m. and 5 p.m.

And at every stage, the dictionary content can be verified in the "work view", provided by a CSS stylesheet, as shown above.

On the right, we present an example of an entry in a dictionary with a somewhat detailed amount of information and granularity thereof.

The dictionary developer needs to fill in simple templates for the relevant parts of speech.

```
<entry xml:id="adui">
 <form>
  <orth>adui</orth>
 </form>
 <xr type="plural-form">
  <ref target="#maadui">maadui</ref>
 </xr>
 <gramGrp>
  <pos>n</pos>
 </gramGrp>
 <sense xml:id="adui.1" n="1">
  <cit type="trans">
   <quote>enemy</quote>
  </cit>
 </sense>
 <sense xml:id="adui.2" n="2">
 <cit type="trans">
  <quote>opponent</quote>
  <note type="hint">in games or sports</note>
 </sense>
</entry>
```

```
<entry>
 <form>
   <orth>adui</orth>
   <ref target="#maadui"/>
 </form>
 <gramGrp><pos>n</pos></gramGrp>
 <sense>
   <def>enemy</def>
 </sense>
 <sense>
  <def>opponent</def>
  <note type="hint">in games or
                    sports</note>
 </sense>
</entry>
```

Then, the predictable work is performed by XSLT scripts, which...

(a) add XML structure to the entry created by a developer

(b) create new entries (in this case, a template plural entry, containing a reference to the singular form)

```
<entry xml:id="maadui">
 <form>
  <orth>maadui</orth>
 </form>
 <gramGrp>
  <pos>n</pos>
 </gramGrp>
 <sense>
  <xr type="plural-sense">Plural of
   <ref target="#adui">adui</ref>
  </xr>
  <def>enemy</def>
  <def>opponent <note type="hint">in
  games or sports</note></def>
 </sense>
</entry>
```

Developments planned for the near future

- After we reach version 0.5 with the cit/quote markup, we plan to start experimenting with dictionary reversal and concatenation (crossing).
- The support for LIFT (Lexicon Interchange FormaT) is next on the agenda.
- More XML technology: tools for feeding dictionaries into, and querying their contents from, native XML databases.

Selected references

- Faith, Rik and Martin, Brett. 1997. A Dictionary Server Protocol. Request for Comments: 2229 (RFC #2229). Network Working Group. Available from ftp://ftp.isi.edu/in-notes/rfc2229.txt
- TEI Consortium, eds. 2007. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.2.0. Last updated on October 31st 2008. TEI Consortium. Available from http://www.tei-c.org/Guidelines/P5/