FreeDict: an Open Source repository of TEI-encoded bilingual dictionaries

Piotr Bański

Institute of English Studies,
University of Warsaw

Beata Wójtowicz

Dept of African Languages and Cultures,
University of Warsaw;

Institute of Computer Science,
Polish Academy of Sciences

Support from the Standing Committee for the Humanities (SCH) of the European Science Foundation made this paper possible (http://www.esf.org/human).

TEI-MM, Ann Arbor, November 2009

How it all began

- Beata's Ph.D. diss on Swahili lexicography, outlining the structure of a modern electronic Swahili-Polish dictionary,
- before that, Beata edits or compiles two little dictionaries (swa-pol, swa-eng), one of them for the Freedict project (swa-eng v. 0.2),
- in early 2007 (?) Piotr talks Beata into trying to apply for a grant to build her dream swa-pol-swa dictionary, we prepare a proposal, and thus get some material to present and discuss at conferences,
- … and so it begins, we talk and read and listen, and re-evaluate and modify, fail to get the grant, reapply... twice...,
- in the meantime, we work on the swa-eng Freedict dictionary in P5, because we like the project and want to popularize it; Piotr joins the Freedict admin team, sets up the wiki and just can't stop meddling with the tools (actually, Adam Przepiórkowski stops him for 10 months of work on the National Corpus of Polish),
- finally, the reviewers get tired and possibly scared of having to re-read our proposals, and we get the grant, officially starting now, in November,
- that forces us to define our stance wrt Freedict, license-wise; in the worst case, we will keep donating pieces of our closed-source dictionary to Freedict.

FreeDict(.org)

Started in 2000 by Horst Eyermann, as a sister project to DICT.

Many databases were adapted from Ergane (travelang.com) by concatenating (= crossing) their dictionaries,with Esperanto as the middle language

(Dictionary crossing: L1→L2 *crossed with* L2→L3 or L3→L2 *becomes* L1→L3)

- advantage: mass production of glossaries
- disadvantage: translation errors bound to crop up

High hopes, lots of work by Michael Bunk, conversion from a plain text format into TEI P4, a TEI P4 dictionary editor for Linux, then... licensing problems (Ergane changed their licenses, many databases had to be withdrawn), Real Life and eventually stagnation.

Revival in late 2007, things started to move again. New features at Sourceforge, newer documentation, a gradual move towards P5 was initiated: the conversion tools had to be adapted, now the XSLT path from P5 to DICT is open.

DICT(.org)

Dictionary Server Protocol (Rickard Faith and Bret Martin, NWG RFC #2229, 1997)

- TCP-based query/response protocol that allows access to dictionary databases,
- data in textual form, an option to provide MIME-encoded content,
- currently maintained and developed at SourceForge.net;
- more than one way to query a DICT database: you can search the definitions and the headwords, using regex-based criteria,
- clients can be free-standing desktop applications or they can be integrated into editors or web browsers; DICT web gateways also exist.

DICT is one of the strengths of FreeDict, as it allows for easy publication of dictionaries. FreeDict, in turn, groups bilingual dictionaries that DICT can use.

Controlled tag abuse; conformance levels

We want:

- mass production of Bad-but-Big dictionaries/glossaries with minimal markup,
- semi-automated production of better dictionaries with richer markup,
- lexicographic works with 'perfect' markup; hand-crafted.

This means:

- 3 ODDs (eventually), each of them progressively restrictive and with tightening semantics,
- allowing for tag abuse at the lower levels (e.g. use <def> for translation equivalents, use <note> for practically everything else, including grammatical information); the tag abuse is controlled by the appropriate ODD,
- allowing for low machine-readability at the BbB level (equivalents separated by commas and semicolons; notes in brackets);
- the three ODDs should not be independent (Laurent, help!); we thought of externally expressed constraints, but ODD inheritance would be much better!

Basic restrictions

No mess – an external demand, by the tools (no entryFree or dictScrap, <sense> and gramGrp/pos obligatory, even if <pos/> ends up empty – support for empty <pos/> implemented right before TEI-MM, see eng-ell, ca. 20.000 entries)

Three levels of conformance for dictionary encoding (an old concept, possibly due to the CES):

- level 1 for mass encoding, <def> with a single string inside
- level 2 for some more nifty microstructure (gramGrp, iterated <form> for pronunciation, iterated <def>s, <note> instead of brackets),
- level 3: the new ultracool cit/quote structure where e.g. <def> means a real definition, should one appear in the entry, <note>s are specialized (<usg>, etc.)

The ODDs should ideally be in a subsumption relation – ODD inheritance would be nice to have... otherwise it's a maintenance nightmare.

Let us illustrate the above now.

Level 1

```
<sense>
   <def>equivalent.1a, (sense restriction) equivalent.1b;
      equivalent.2 (definitional comment)</def>
</sense>
```

Level 2

```
<sense>
   <def>equivalent.1a, <note type="hint">sense restriction</note>
            equivalent.1b</def>
   <def>equivalent.2</def>
   <note type="def">definitional comment</note>
</sense>
```

Level 3

```
<sense n="1" xml:lang="L2">
    <cit>
        <quote>equivalent.1a</quote>
    </cit>
    <cit>
        <note type="restr" xml:lang="L1">sense restriction</note>
        <quote>equivalent.1b</quote>
    </cit>
</sense>
<sense n="2" xml:lang="L2">
    <cit>
        <quote>equivalent.2</quote>
        <def>definitional comment</def>
    </cit>
</sense>
```

Local vs. project-wide markup

- language-specific, where various linguistic features pertaining to the given language pair are encoded in detail

<form type="N"> <orth>alasiri</orth> </form> → alasiri [sg = pl]

- uniform pre-conversion format, which is translated into the DICT format by a single project-wide set of stylesheets.

The mapping between the language-specific format and the pre-conversion format is performed by language-specific XSLT stylesheets.

XSLT is also helpful in producing new entries, language-specifically (we automatically produce entries for plural nouns from the singulars – in Swahili, plurals are marked by prefixes; illustration in the appendix)

What we have

- 70 databases of varied quality (most need maintenance) and size (the largest have 140-150 thousand entries),
- regular format (eventually exposing the translation equivalents in cit/quote markup),
- a way to publish almost immediately, via Sourceforge, Freedict's own Debian package series, and via DICT,

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

- officer

- easy access (web, desktop; small local servers easy to install),
- plans for the future...

Plans for the future

- Make all databases P5-conformant,
- In doing so, establish what exactly is needed for our internal three-level conformance and freeze the ODDs,
- Find a DWS for the developers (lossless transformation into/from the data format of e.g. FLEx or Glossword),
- Establish a reasonably reliable concatenation method to produce more glossaries or dictionaries,
- Perform dictionary reversal for those databases where the equivalents are suitably exposed in cit/quote.

What we need

- A way to express the relatedness of ODDs, preferably by inheritance (cf. Laurent's talk!) – we used to think of an outside set of pointers to express this;
- Dictionary developers – anyone here? got any clever M.A.-level students interested in lexicography?
- Data available under free licenses;
- Tools (or reliable methods of lossless translation of data formats); how likely is it to have a Dictionary Writing System using the TEI as its format? It would have to understand the ODDs.

Possible extensions

Wójtowicz and Bański: Swahili-Polish-Swahili dictionary project (previously, it informed the Freedict swa-eng dictionary; it may be based on it now, pending licensing questions)

Bański and Moszczyński (LREC 2008), “Enhancing an English-Polish Electronic Dictionary for Multiword Expression Research” – an RNG grammar for describing MWEs as regular expressions, applied to dictionary entries; dormant, will be woken up around January.

Conversion between TEI and other dictionary-oriented XML formats (OLIF, LIFT)

Bański, Habilitazionsschrift on applying the GOLD ontology to dictionary markup and content – Piotr has around a year for that, not longer (or he gets a kick out of the Uni).

+ whatever you can think about with regard to FreeDict in your own research – we'll be happy to assist on the administrative/SF side.

Thanks for your attention!

http://freedict.org/

Appendix: our story (I)

```
<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>
```

Appendix: our story (II)

```
<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>
```

Appendix: our story (III)

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

Appendix: automatic enrichment of the dictionary (I)

The singular 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>
```

Appendix: automatic enrichment of the dictionary (II)

The plural entry:

```
<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>
```