Swahili-Polish lexicography: A new chapter

Piotr Bański and Beata Wójtowicz (Warsaw University)

With nearly 80 million speakers in East and Central Africa, Swahili is the most widely spoken African language. It serves not only as official and national language but also as a *lingua franca* for many millions of African citizens. Although numerous bilingual Swahili dictionaries exist, the only publication of this kind targeting Polish is the small Swahili-Polish and Polish-Swahili dictionary by Stopa and Garlicki (1966), obsolete, heavily criticized and out of stock for years now.

The talk presents an ongoing project, aiming at building a new, electronic, XML-encoded Swahili-Polish-Swahili learner dictionary. The dictionary will contain 10,000 entries, with extended grammatical information, explicitly marked agreement and class prefixes, as well as examples of use.

At the moment, the dictionary exists as a test Swahili-English dictionary containing almost 300 entries (growing week by week) presenting test-cases for the most outstanding problems discussed in Swahili lexicographic literature. It also contains a fairly complete implementation of the pronominal system of Swahili, and thus, while being overall a 'toy' dictionary, it can easily be called a dictionary of Swahili pronouns (which, listed as roots as well as agreeing forms, take up nearly half of all entries; an example is presented in Figure 1).

This Swahili-English dictionary will be merged with a Swahili-English dictionary produced by SALAMA (Swahili Language Manager, cf. Hurskainen 2007) out of the 20 million-word Helsinki Corpus of Swahili (HCS 2004) and only then will it be turned into a Swahili-Polish dictionary proper, in a semi-automatic fashion: it will be concatenated with an electronic English-Polish dictionary and then cleaned up by the lexicographer team.

The test dictionary has already been concatenated with the dictionary by Saloni & Piotrowski (1997) and the results of this process will also be presented. Because the process of concatenation, with its obvious virtues and undeniable dangers, is a topic for a separate paper, the presentation of the "mini Swahili-Polish dictionary" will only serve as the conclusion of the talk.

We will present the linguistic features of Swahili that guided the architectural decisions concerning the dictionary and the possibilities in which the contents may be visualized for the user. The various visualization strategies and formats are made possible by a consistent encoding format, based on the Text Encoding Initiative (TEI) Guidelines (Sperberg-McQueen & Burnard 2007). The encoding format makes it possible to implement cross-entry hyperlinks that link to derivatives, derivational bases, inflected forms or synonyms. At the user's request, it is possible to restructure the dictionary on the fly, e.g. including all possible derivatives (or only some of them) in the entry for the root word, or splitting them across the entire dictionary as separate entries (the default case). By creating this dictionary in such a form, we believe to be opening a new chapter in Swahili-Polish lexicography.

Figure 1. A sample entry from the raw dictionary, before it is processed by XSLT scripts that create indexes, hyperlinks, and that may also be used to query the contents in various ways, e.g. by grammatical (sub) category (*kwako* is a possessive pronoun), by the type of agreement (with nouns of classes 15 and 17), by inflectional stem (*-ako*), etc.

```
<entry type="inflected" n="kwako">
  <form>
    <orth type="morph">kw|ako</orth>
    <orth>kwako</orth>
  </form>
  <gramGrp>
    <pos>pron</pos>
```

```
      <subc type="pron">poss</subc>
      <per>2</per>
      <number>sg</number>
      <gramGrp type="agr">
        <iType>15</iType>
        <iType>17</iType>
      </gramGrp>
    </gramGrp>
    <xr type="inflection_of">
      <ref target="#ako">-ako</ref>
    </xr>
    <sense>
      <trans>
        <tr>your</tr>
        <tr>yours</tr>
      </trans>
      <dicteg>
        <q>nyumbani <oRef/></q>
        <def>at your house</def>
      </dicteg>
    </sense>
  </entry>
```

References

HCS 2004. *Helsinki Corpus of Swahili. Compilers*. Institute for Asian and African Studies (University of Helsinki) and CSC – Scientific Computing Ltd.

Hurskainen, Arvi. 2007. SALAMA Dictionary Compiler. Ms., University of Helsinki. Available at http://www.transtechno.fi/dict-examples.pdf

Piotrowski, Tadeusz and Zygmunt Saloni (eds.). 1997. *Nowy słownik angielsko-polski i polsko-angielski*. Warszawa, Wilga. Converted into XML by Adam Mazur and Maciej Wojciechowski, 2006.

Sperberg-McQueen, C.M. and Lou Burnard (eds). 2007. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Available at http://www.tei-c.org/P5/

Stopa, Roman and Bolesław Garlicki. 1966. *Mały słownik suahilijsko-polski i polsko-suahilijski*. Warszawa, Wiedza Powszechna.