Updating UWAZO – A corpus based Kiswahili-Italian on-line lexical data base

Maddalena Toscano (Università degli studi di Napoli – L'Orientale), Massimilano Sorrentino (Università di Napoli – Federico II) and Tommaso Borrelli (Skylines s.r.l., Napoli)

UWAZO is a small project aimed at building a corpus-based on-line accessible Swahili-Italian lexical data base, built on the base of the TEI guidelines for dictionary editing. It is meant specially for Italian students of Swahili language course.

The first version, still under use, allows the user to structure information related to an entry by selecting groups and elements according to the type of content; all the information have to be inserted manually. Operations such as inserting, deleting, duplicating or moving groups and/or elements are possible but a bit awkward. Among the various types of content there are groups and elements which can store samples, with Italian translation, and eventually indication of the source text. The samples are extracted from a corpus of Swahili texts through the use of SWALEM, a basic Swahili lemmatizer for which an on line version is under preparation.

The updated version presented here integrates the lemmatizer into the data base. The result should be a software which performs the following tasks:

– take a Swahili text as input (SWALEM)

– produce a lemma list, with indication of basic POS tagging and a very general Italian glossa (both context independent) of the Swahili words contained in the text (SWALEM)

– import the list into the lexical data base (UWAZO)

– produce as an output a structured data base which includes fields with lemma list, basic POS tagging, Italian glossa, contexts (UWAZO)

The user can then proceed to further operations such as adding, deleting, duplicating, moving the groups and/or elements, plus selecting the relevant contexts and eventually adding the translation.

The paper aims at presenting the results of the software updating, with a language dependent version especially tailored for Swahili, and a language independent version (which does not include lemmatization) usable also for other languages which use Latin alphabet; both versions will be availbale to interest users.