# A New Corpus-based Swahili-Polish Dictionary

Beata WÓJTOWICZ
Department of African Languages and Cultures
University of Warsaw
Warsaw, Poland
b.wojtowicz@uw.edu.pl

Over the last several years, we have observed a rising interest in African studies at the University of Warsaw in Poland, especially in the field of Swahili language learning. Due to the shortage of educational material, a new project, involving the creation of a dictionary, was proposed by the Department of African Languages and Cultures, and the idea was then approved by the Polish Ministry of Science (for the years 2009-2012).

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), which has been out of print for a number of years. While students can manage with various English-language textbooks available for the teaching of Swahili, they find it difficult without access to a bilingual dictionary which would be in their own native tongue, especially in the early years of their studies. Presently, Polish students are forced to use non-Polish dictionaries, including various resources available on the Internet. Two electronic Swahili-English dictionaries are especially popular and widely-used: the *Internet Living Swahili Dictionary* (Kamusi Project, 2012) – the largest such dictionary and an on-line community-based initiative, as well as the *TshwaneDJe Swahili-English Dictionary* (Hillewaert *et al.*, 2012) – the first, and so far the only, corpus-driven electronic Swahili dictionary.

The paper describes the progress of the project, which endeavours to create a new dictionary that would be constructed in accordance with recent lexicographic practices. The final text of the dictionary will be based on corpus-driven data, published in an electronic form, encoded in XML, and accessible via the Internet, with possibly a printed version published as well. The focus within this article is on the Swahili-Polish section, as the reverse-language part will be initiated in the form of a structured index – a standard in this type of resource. The primary target audience for the dictionary are learners of Swahili; therefore, the dictionary will be of the descriptive-translational type, with additional grammatical information provided beyond the scope of what is minimally necessary. The data included within the dictionary will also be presented in a user-friendly manner.

The skeleton of the dictionary has been derived from an electronic resource – the largest available and the only annotated corpus of Swahili, namely the Helsinki Corpus of Swahili (HCS). This material has already served as the source for the very successful Swahili-Finnish-Swahili dictionary by Abdulla *et al.* (2002). The selection of headwords for the Polish-Swahili dictionary (ultimately 10,000 entries, published incrementally) has been made primarily on the basis of a frequency list derived from the above-mentioned corpus, and further updated with information available in various Swahili textbooks and other material.

The primary data included in the dictionary has been derived from a tagged corpus. All entries have annotations with POS information and some additional data, depending on the category of a given word, such as class numbers and animacy categorization for nouns, types of derivation for verbs, as well as expanded references to their base/root or derivatives. Additionally, all entries are accompanied by their English equivalents, which have been further concatenated with electronic English-Polish dictionary data. Therefore, apart from the grammatical information, the lexicographer is also provided with the Polish equivalents.

In this presentation, I would like to concentrate on presenting the primary dictionary data, as well as the dictionary guidelines, aimed at producing a useful educational tool which would keep the students' needs in mind. Emphasis is placed on the presentation and possible visualisation of derivational families (which are quite extensive in Bantu languages), as this is regarded as an educational feature useful for the development of our students' linguistic skills.

**References**


Abdulla, A., Halme, R., Harjula, L. & Pesari-Pajunen, M. (2002). *Swahili-Suomi-Swahili sanakirja*. Helsinki: Suomalaisen Kirjallisuuden Seura.

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

Hillewaert, S., Joffe, P. & de Schryver, G. M. (2012). *Kamusi ya Kiswahili–Kiingereza Katika Mtandao*/*Online Swahili-English Dictionary.* Available from: http://africanlanguages.com/swahili/.

Kamusi Project. (2012). *The Internet Living Swahili Dictionary.* Available from: http://kamusi.org

Stopa, R. & Garlicki, B. (1966). ‘Mały słownik suahilijsko-polski i polsko-suahilijski.’ Warszawa: Wiedza Powszechna.