Kernerman Dictionary Research Grants

The Creation of an Innovative Kiswahili-English Online Dictionary

Gilles-Maurice de Schryver and Sarah Hillewaert

AFRILEX has allocated a two-year grant for 2003 and 2004 to Gilles-Maurice de Schryver (principal investigator) and Sarah Hillewaert (co-investigator) for the project 'The Creation of an Innovative Kiswahili-English Online Dictionary', the outline of which appears herewith.

The Kernerman Dictionary Research Grants are allocated by independent Assessment Committees administered by AFRILEX, ASIALEX and EURALEX.

Funds are still available for 2004 from ASIALEX and EURALEX for research projects in the following fields:
- The study of the dictionary-using behavior of language learners at the elementary-to-high-school levels, and the design of these dictionaries.
- Specialized corpora for foreign-language learners.
- The function of lexicography in vocabulary acquisition.
- Trilingual and multilingual lexicography.
- Lexicography in the service of language preservation.

More details are available in *Kernerman Dictionary News*, Number 10, July 2002 (http://kdictionaries/newsletter/kdn10-11.html). Applications should be made to the relevant committees directly.

Kiswahili (Swahili) is one of Africa's major languages, spoken throughout East Africa as a lingua franca by tens of millions of people. Despite its official status, substantial number of speakers and relatively long dictionary tradition, the state of lexicographical research as well as the availability of modern and up-to-date dictionaries for Kiswahili is far from satisfactory. The numerous Kiswahili monolingual and bilingual dictionaries compiled for over a century are largely based on Western compilation principles, and until now the most commonly used dictionaries remain rooted in lexica originally derived by missionaries.

Kiswahili is an *agglutinating* language, meaning that morphemes are juxtaposed to form linguistic words. In all current dictionaries, 'orthographic words' are decomposed into their formatives, with only the latter being lemmatised. As a result, not all native speakers of Kiswahili can look up 'words' in their own language – as this implies being able to cut off prefixes and suffixes – and even trained scholars often need more than one look-up round before they hit on what they are looking for (since vocal changes between formatives are not always predictable).

This research project attempts to deal with all these problems simultaneously. The aim is to create the first corpus-based Kiswahili dictionary that is also intuitive in nature, and to research the feasibility of this approach in real time. Instead of lemmatising stems as in traditional dictionaries, the idea is to lemmatise full orthographic words (in addition to stems), and to provide full translations for these strings. In order to sensibly limit the number of items one can physically treat, the items will be selected from a frequency list derived from a large corpus. Concordance lines will be called up for each frequent orthographic word, and the various translations will be recorded in order of frequency. A user will thus be able to look up words directly, as they are spoken or written, and the translations will be arranged from most likely to least likely. An English search index will additionally enable searches in the reverse direction. Since, obviously, such an approach will require much more 'space' than in a traditional stem-based dictionary, the dictionary will be developed and made available in an electronic environment right from the start, primarily on the Internet, where it is also possible to keep a log of all searches. Analyzing these log files will enable further research on whether or not this hybrid approach is feasible and to amend the approach if need be.

Given the intuitive lemmatisation approach, native speakers and learners at the elementary and intermediate levels will for the first time be able to effectively look up words, and find meanings of 'real' words, which should help to develop a dictionary culture. Furthermore, the log files will be utilised to full potential by tracking each individual dictionary-use behaviour, including vocabulary retention. For the first time, truly unobtrusive data will be collected and true look-up behaviour in an electronic environment will be recorded. Finally, this project will also ensure that Kiswahili, an increasingly popular language on the Internet, is also kept alive in a modern online reference work based on sound lexicographical principles.

Mr de Schryver has participated in building a 13-million-word Kiswahili corpus, which will be queried in this project. He is the main compiler of a recent online dictionary for Sesotho sa Leboa (http://africanlanguages.com/sdp), from which valuable information will be drawn. Ms Hillewaert will be responsible for the compilation and editing of the Kiswahili-English translations. The two have worked together on various Kiswahili projects over the past few years, and produced a lexicon for Sheng (a language that is largely based on Kiswahili and spoken by the youth in Nairobi).

It is expected that a first lexicon, consisting of translations for the top 1500 orthographic words, could be uploaded following three months of work, and batches of 1500 items could be added to the online dictionary every three months. The project is planned to run for at least three years (and thus reach 18,000 items) before the great majority of the searches will be successful.