Corpus of text in indigenious African scripts

Solomija Buk & Andrij Rovenchak

Ivan Franko National University of Lviv - Ukraine

We report the launch of the project on compiling a multilingual corpus composed of texts written in indigenous African scripts. The corpus is intended as a research tool for the studies on these writing systems and will cover all kinds of script specimens: books, newspapers, manuscripts, private records, tax receipts, etc.

The issue of an adequate representation of indigenous writing systems is solved as follows. Three scripts are represented in the Unicode standard: N'ko, Vai, and Somali Osmanya. It is likely that the Bamum script will join this list soon. For most scripts (Bassa, Mende, Kpelle, Loma, etc.), which are non-Unicode ones, the encoding is arbitrary they can be allocated in the Private Use Area for Unicode compatibility. For the corpus, a system is developed ensuring a unique name for every symbol in each script.

The text division in the corpus can be ***logical*** *(parts, sections, paragraphs, sentences, words, etc.) and* ***physical*** *(sheets, pages, columns, lines, etc.). The mark-up of the latter is considered urgent for the presentation of old indigenous texts.*

One of the forms of presentation for this research tool is an online concordance. A trial version for the Vai script is already functioning at the following link: http://www.ktf.franko.lviv.ua/~andrij/vai-concord.html.

The first stage of the project implementation is the preparation of sample subcorpora in several scripts based primarily on text illustrations in printed sources as well as information available in the Web. This should confirm the functionality and applicability of the proposed approach and reveal some possible necessity of improvements. Further collection of texts must be done in a close cooperation with scientists working in this domain.

In the epoch of globalization, indigenous writing systems are on the wedge of extinction and oblivion. Creation of a research tool to study them can preserve this unique heritage before it is lost forever.