

LREC 2014 Workshop

CCURL 2014

**Collaboration and Computing for Under-Resourced Languages
in the Linked Open Data Era**

Reykjavík (Iceland), 26th May 2014

Poster Presentation

Title

Hungarian-Somali-English Online Dictionary and Taxonomy

Author

István Endrédy

Abstract

Background. The number of Somalis coming to Europe has increased substantially in recent years. Most of them do not speak any foreign language, only Somali, but a few of them speak English as well. Aims. A simple and useful online dictionary would help Somalis in everyday life. It should be online (with easy access from anywhere) and it has to handle billions of word forms, as Hungarian is heavily agglutinative. It should handle typos as the users are not advanced speakers of the foreign languages of the dictionary. It should pronounce words, as these languages have different phonetic sets. It should be fast with good precision because users do not like to wait. And last but not least, it should support an overview of the vocabulary of a given topic. Method. A vocabulary (2000 entries) and a taxonomy (200 nodes) was created by a team (an editor and a native Somali speaker) in an Excel table. This content was converted into a relational database (mysql), and it got an online user interface based on php and jqueryui. Stemmer and text-to-speech modules were included and implemented as a web service. Typos were handled with query extension. Results. Although the dictionary lookup process does stemming with a web service and makes a query extension process, it is very fast (100-300ms per query). It can pronounce every Hungarian word and expression owing to the text-to-speech web service. Conclusion. This dictionary was opened to the public in October, 2013. (<http://qaamuus.rmk.hu/en>) The next step is the creation of a user interface optimised for mobile devices.