Hungarian-Somali-English online dictionary and taxonomy

István Endrédy
istvan.endredy@gmail.com

Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
MTA-PPKE Hungarian Language Technology Research Group
1083 Budapest, Práter u. 50/a

BACKGROUND

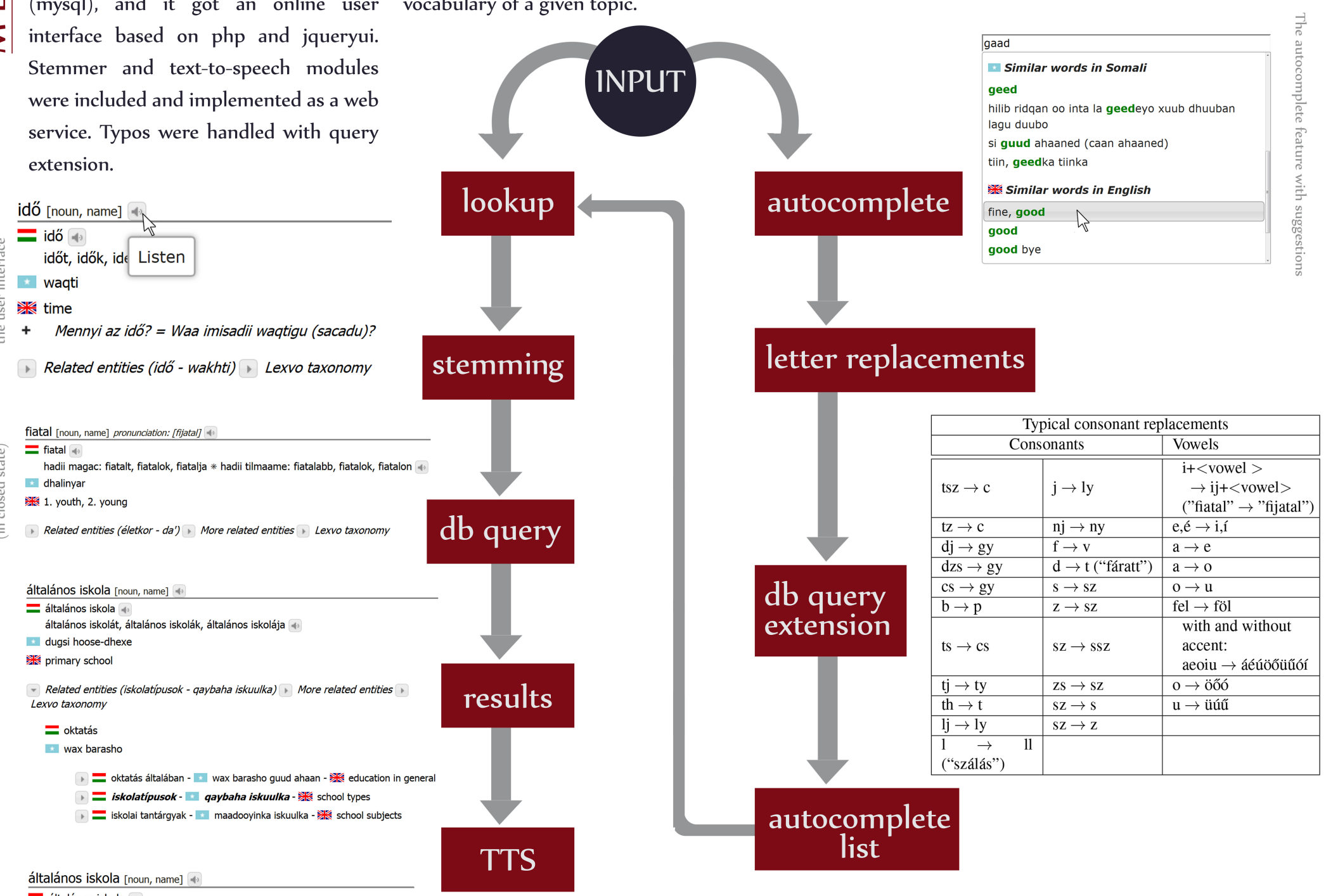
The number of Somalis coming to Europe has increased substantially in recent years. Most of them do not speak any foreign language, only Somali, but a few of them speak English as well.

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.

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.



Typical consonant replacements

Consonants		Vowels
tsz → c	j → ly	i+<vowel> → ij+<vowel> (”fiatal” → ”fijatal”)
tz → c	nj → ny	e,é → i,í
dj → gy	f → v	a → e
dzs → gy	d → t (“fáratt”)	a → o
cs → gy	s → sz	o → u
b → p	z → sz	fel → föl
ts → cs	sz → ssz	with and without accent: aeoiu → áéúöőüűóí
tj → ty	zs → sz	o → öőó
th → t	sz → s	u → üúű
lj → ly	sz → z	
l → ll (“szálás”)		

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

[1] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg. Springer-Verlag.

[2] de Melo, G. and Weikum, G. (2008). Language as a foundation of the Semantic Web. In Bizer, C. and Joshi, A.,editors, Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008), volume 401 of CEUR WS, Karlsruhe, Germany. CEUR.

[3] Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, C.,and Gordos, G. (2000). Profivox — A Hungarian text-to-speech system for telecommunications applications.International Journal of Speech Technology, 3(3-4):201–215.

[4] Oravecz, C. and Dienes, P. (2002). Efficient stochastic part-of-Speech tagging for Hungarian. In Proceedings of the Third International Conference on Language Resources and Evaluation, LREC2002, pages 710–717, LasPalmas.

[5] Prószéky, G. and Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In Dale, R. and Church, K. W., editors, ACL. ACL.

[6] Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. Web Semant., 6(3):203–217, September.

This work was partially supported by TÁMOP – 4.2.1.B – 11/2/KMR-2011-0002 and TÁMOP – 4.2.2/B – 10/1–2010–0014.