DAOULAGAD – A CELTO-SLAVIC OCR DICTIONARY

DMITRI KHRAPOV

When the present writer started learning the Welsh language in the mid-1990s, there was no Welsh-Russian or Russian-Welsh dictionary either online or in print. In order to facilitate my studies, a Welsh↔Russian dictionary was created and put online at http://www.cymraeg.ru/geiriadur/. At the time of writing the dictionary contains more than 10,000 Russian and 6,000 Welsh words, more than 70% of the 5,000 most frequently used words are covered. It takes into account initial consonant mutations and supports 'Item and Arrangement' and 'Word and Paradigm' morphological models (Plungian 2003: 71). Spell-checking is implemented with special data structures called 'metric trees' (Burkhard-Keller's metric trees and vantage-point metric trees are employed) and via an open source software library called GNU aspell.

My gratitude goes to A. R. Muradova and to numerous volunteers who made it possible to add the Irish and Breton languages to the system. (As yet) untranslated Celtic words are looked up in English or French-language dictionaries, thus eliminating the need for multiple dictionaries.

There are plans to publish the Welsh↔Russian section, and it may well become the first Celtic-Russian dictionary in print.

The ever-growing nature of the dictionary dictated that it stays online. At the same time, reading printed Welsh and Breton books was (and is) one of its major uses, so a solution for convenient text input had to be found, especially for more uncommon characters like *ñ, ù* or *ŵ*. The inspiration came from a hand-held scanner like IRISPen or C-Pen.

As a result, Daoulagad [dɔwˈlaːgat], a mobile dictionary app, was written (in Common Lisp programming language). OCR (Optical Character Recognition) capabilities make it possible to use a smartphone camera as an input device. A user chooses languages, takes a picture and gets a translation. Hand-writing recognition and language auto-detection are not supported yet. As it is a dictionary, not a translator, it translates one word at a time (lest reading pleasure be spoiled).

Recognition errors are corrected using trigram frequencies calculated over an extensive corpus. The following OCR techniques (Cheriet 2007: 129) were tested: wavelets, Hough and Fourier transforms, ARG matching, Hausdorff metric, ANNs; L1 metric was chosen. Hidden Markov Models have not yet been tested.

Daoulagad works on Android, iOS and Windows Phone smartphones. More than 40 languages using Arabic, Armenian, Cyrillic, Devanāgarī, Greek,

Hebrew, Latin and Thai scripts are supported, as well as hanzi/kanji. It is possible to recognize a word in any of these languages and translate it into Welsh or any other supported language.

Independent scholar

References

Plungian, V., 2003, *Общая морфология* [General morphology], Moscow.
Cheriet, M., **Kharma**, N. et al., 2007, *Character recognition systems: a guide for students and practitioners*, Hoboken, New Jersey.