How Can Language Technology Fight Against Language Death?

Ivett Benyeda
Research Institute for Linguistics, Hungarian Academy of Sciences
benyeda.ivett@nytud.mta.hu

Eszter Simon
Research Institute for Linguistics, Hungarian Academy of Sciences
simon.eszter@nytud.mta.hu

Péter Koczka
Research Institute for Linguistics, Hungarian Academy of Sciences
koczka.peter@nytud.mta.hu

According to the UNESCO Atlas of the World's Languages in Danger (Moseley 2010) there are 646 definitely endangered, 528 severely endangered and 576 critically endangered languages. The European Union places great emphasis on the preservation of linguistic diversity, which means that it is a common aim to support endangered languages and prevent language death.

The digital revolution of our era has a dramatic impact on nearly all aspects of society. Language communities are most sensitive to and therefore most affected by new paradigms in communication technology (Simon et al. 2012). According to Kornai (2013), a language is digitally viable only to the extent it produces new, publicly available digital material. Language death implies loss of function, entailing the loss of prestige, and ultimately the loss of competence. In this context, language technology aspires to become an enabler technology that helps people to collaborate, conduct business and share knowledge regardless of language barriers (Simon et al. 2012). However, cutting-edge technologies are typically available only for widely-spoken ('thriving') languages (Kornai 2013).

In this presentation, our aim is to present an ongoing project whose objective is to produce digital material for the following endangered Finno-Ugric (FU) languages: Komi-Zyrian, Komi-Permyak, Udmurt, Meadow and Hill Mari and Northern Sami, helping them in the process of revitalisation.

To achieve our goals, we collect parallel, comparable and monolingual texts for the mentioned small FU languages and for thriving languages that are of interest to the FU community: English, Russian, Finnish and Hungarian. We generate proto-dictionaries for the FU–thriving language pairs and will deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary. See the workflow in Figure 1: the green-coloured steps have already been

conducted, the orange ones are under development, while the red one indicates the final step which will be taken in the last phase of the project.

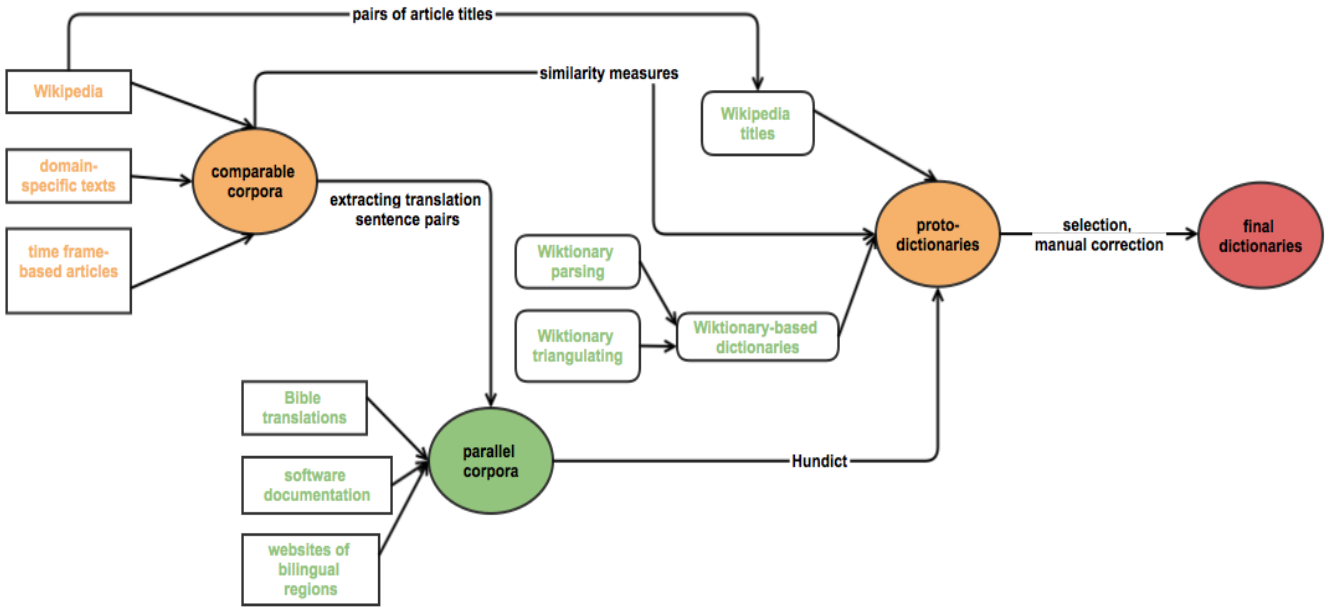

Figure 1: The main steps of the workflow of creating dictionaries

First, we created proto-dictionaries from already existing, digitally available dictionaries, from Wiktionary by using parsing and triangulating methods (Ács 2013), and by extracting title pairs from Wikipedia.

For extracting translation candidates from parallel corpora, we use the HunDict[1] tool. We are experimenting with methods to extract real parallel sentences from comparable data, which can then be used as input material for generating new word pairs. Several similarity measures can also be used for extracting translation candidates directly from comparable corpora. The difficulty with this task is the lack of text processing tools for these small languages. We will try to use language-independent tools and methods for extracting as much translation candidates as possible.

Having the proto-dictionaries, they will be merged, then translation candidates with the highest confidence measure will be chosen. As the last step, each entry will be manually checked by native speakers and uploaded to Wiktionary.

While the expansion of Wiktionary is targeting the direct support of the mentioned language communities, other materials (dictionaries, corpora, models) will enable the development of additional tools as all of them will be publicly available after the end of the project.

The research reported in the paper is conducted with the support of the Hungarian Scientific Research Fund (OTKA) grant #107885.

[1] https://github.com/zseder/hundict

References

Moseley, C. (eds.).(2010) *Atlas of the World's Languages in Danger*. 3rd ed. Paris, UNESCO Publishing. Online version: **http://www.unesco.org/culture/en/endangeredlanguages/atlas**

Simon, E.; Lendvai, P.; Németh, G.; Olaszy, G.; and Vicsi, K. (2*012) A magyar nyelv a digitális korban – The Hungarian Language in the Digital A*ge. Georg Rehm and Hans Uszkoreit (Series Editors): META-NET White Paper Series. Springer

Kornai, A. (2013). *Digital Language Death*. PLoS ONE, 8(10).

Ács, J. (2014) *Pivot-based multilingual dictionary building using Wiktionary*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14), Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).