RESEARCH INSTITUTE FOR LINGUISTICS
HUNGARIAN ACADEMY OF SCIENCES

Automatic creation of bilingual dictionaries for Finno-Ugric languages

IWCLUL2015
16 January 2015
Tromsø, Norway

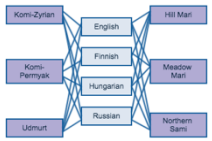
Eszter Simon, Ivett Benyeda, Péter Koczka, Zsófia Ludányi
{simon.eszter, benyeda.ivett, koczka.peter, ludanyi.zsofia}@nytud.mta.hu

Research Institute for Linguistics, Hungarian Academy of Sciences

Introduction

The objective of the project: to provide linguistically based support for several small Finno-Ugric (FU) digital communities in generating online content.
Our goal: to generate proto-dictionaries for 24 language pairs and deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary.
To achieve our goals: we collect parallel, comparable and monolingual text material for 6 small FU and 4 major languages. After processing and analysing them, we apply several automatic dictionary building methods.

Figure 1. Language pairs



Text collection

Parallel corpora
• Bible translations from the Parallel Bible Corpus, Bible.is, The Unbound Bible;
• software documentation from the OPUS corpus;
• translations from the websites of officially bilingual regions of Norway, Finland and Russia.
Comparable corpora
• interlanguage-linked article pairs from Wikipedia with some metadata and Wikidata IDs;
• domain-specific monolingual texts by specifying a keyword: documents on the Sami culture, education and society in English and Northern Sami;
• multilingual daily newspaper materials from the same time interval: articles from online newspapers in Finland.
Monolingual texts
• from several websites in various domains;
• serve as training data for the tokenizer and sentence splitter.

lang	mono	lang pairs	parallel		comparable	
			L1	L2	L1	L2
sme	1,364,254	sme–eng	691,260	724,750	253,930	1,754,968
		sme–fin	245,440	273,973	239,651	5,259,591
		sme–rus	173,179	220,790	212,332	233,748
		sme–hun	171,668	224,014	86,244	106,391
kpv	480,609	kpv–eng	121,108	174,742	89,580	183,602
		kpv–fin	121,120	133,715	88,507	80,797
		kpv–rus	117,903	125,085	108,013	141,369
		kpv–hun	121,319	134,344	68,179	74,274
koi	719,325	koi–eng	0	0	257,871	194,784
		koi–fin	0	0	137,578	77,696
		koi–rus	0	0	188,334	139,976
		koi–hun	0	0	95,120	64,794
mhr	1,335,457	mhr–eng	128,316	175,075	121,588	250,583
		mhr–fin	128,328	133,965	118,120	115,028
		mhr–rus	109,449	109,818	158,977	215,724
		mhr–hun	128,565	134,618	106,813	121,453
mrj	366,964	mrj–eng	0	0	137,088	306,465
		mrj–fin	0	0	85,134	93,622
		mrj–rus	0	0	124,289	187,687
		mrj–hun	0	0	77,855	90,168
udm	584,113	udm–eng	0	0	67,306	135,450
		udm–fin	0	0	56,222	49,961
		udm–rus	0	0	80,800	129,293
		udm–hun	0	0	41,883	48,736

Table 1. Number of tokens for monolingual, parallel and comparable corpora. We use the ISO 639-3 language codes: sme – Northern Sami, kpv – Komi-Zyrian, koi – Komi-Permyak, mhr – Meadow Mari, mrj – Hill Mari, udm – Udmurt.

Text processing

1. Text extraction: from HTML, PDF, etc.
2. Character normalization: conversion to plain text using standard Unicode characters in UTF-8 encoding.
3. Language discrimination: *Blacklist Classifier*
• for Komi-Zyrian and Komi-Permyak: 97.47% accuracy,
• for Meadow Mari and Hill Mari: 96.77% accuracy.
4. Filtering out non-FU language elements: *Langid*
• language models created using manually selected text samples;
• dates preserved for creating time frame-based comparable corpora.
5. Sentence segmentation and tokenization: *Apache OpenNLP*
• models built only for languages written in Cyrillic script
• 10k sentences for every language, manual correction, split into train and test sets (90%-10%)
• Russian abbreviation list for supporting sentence segmentation
• over 98% F-measure for all languages
6. Morphological analysis and disambiguation
• analysers available as online applications: Udmurt, Komi-Zyrian, Hill Mari
• *Giellatekno* source files for HFST-based morphological analysis: all small FU languages except Komi-Permyak
• for languages that lack any morphological analysers:
  • using semi-supervised or unsupervised morphological segmentation tools such as *Morfessor*
  • annotation transfer between closely related languages

Methods used for creating proto-dictionaries

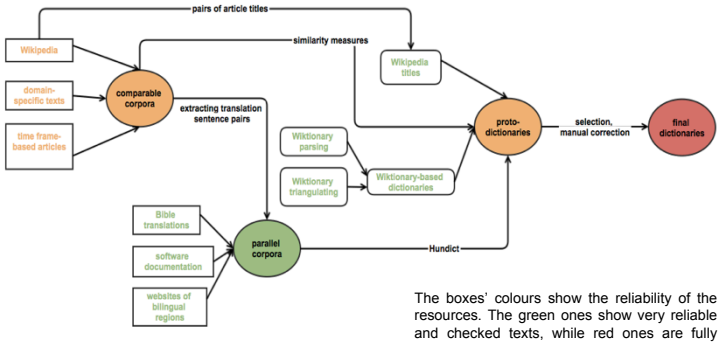
Wikipedia titles: bilingual dictionaries from Wikipedia title pairs using the interwiki links.
Wiktionary-based methods: *Wikt2dict*
• **Wiktionary parsing**: extracting translations from the translation tables in the English, Finnish, Russian and Hungarian editions of Wiktionary;
• **Wiktionary triangulation**: discovering previously non-existent links between translations.
•**HunDict:** extracting word pairs based on high co-occurence in corresponding text segments resulted in word pairs along with their confidence measures.

lang pair	no. of entries	lang pair	no. of entries
koi–eng	3,170	mrj–eng	4,264
koi–fin	1,633	mrj–fin	1,784
koi–hun	1,356	mrj–hun	1,408
koi–rus	1,999	mrj–rus	2,783
kpv–eng	3,636	sme–eng	18,529
kpv–fin	2,244	sme–fin	20,137
kpv–hun	1,782	sme–hun	12,791
kpv–rus	4,416	sme–rus	11,319
mhr–eng	5,348	udm–eng	5,891
mhr–fin	4,815	udm–fin	3,836
mhr–hun	3,312	udm–hun	3,270
mhr–rus	4,448	udm–rus	4,691

Table 2. Number of entries of raw, un-cleaned proto-dictionaries for each language pair.

The workflow of creating dictionaries

Figure 2. Main steps of the workflow of creating dictionaries



The boxes' colours show the reliability of the resources. The green ones show very reliable and checked texts, while red ones are fully automatically generated materials without any manual checking.

1. We create annotated parallel corpora from the downloaded text materials.
2. From comparable texts, we extract translation sentence pairs and create parallel corpora.
3. Some dictionary building methods based on the words' similarity measures can also be used directly on comparable corpora.
4. From all the annotated parallel corpora, we create proto-dictionaries with Hundict and extend these with the results of the triangulation method used on Wiktionary and the word pairs from corresponding Wikipedia article titles.
5. Having a few proto-dictionaries for each language pair, these will be heavily overlapping each other.
6. Combining the proto-dictionaries together and extending them with already existing electronic dictionaries we will be able to determine a threshold above which translation pairs are likely to be valid.
7. As a last step all candidates will be checked by native speakers for getting absolutely reliable translations.

Conclusion and future work

Collecting and processing text material in these under-resourced languages are challenging.
- Only a very few digital text material is available.
- There are no tokenizers and sentence splitters developed especially for these languages.
- There are no available morphologically annotated texts for training and testing.

In spite of the difficulties, we collected some text material for these languages and built proto-dictionaries by applying several methods.

Plans for the future:
- Enriching dictionary entries with morphological, etymological and lexico-semantic information and translation equivalents across languages.
- Creating Wiktionary input files as automatically as possible.
- Uploading the entries validated and corrected by native speakers to Wiktionary.

After cleaning the copyright issues, we will make all of the generated resources (corpora, dictionaries, models) publicly available.

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