Language technology support for Finno-Ugric digital communities

Eszter Simon, Ivett Benyeda, Péter Koczka

Research Institute for Linguistics,
Hungarian Academy of Sciences

19th August 2015
XII International Congress for Finno-Ugric Studies

The project

Finno-Ugric Digital Natives: Linguistic support for Finno-Ugric digital communities in generating online content

- supported by the Hungarian Scientific Research Fund (OTKA No. FNN 107885)
- project investigator: Tamás Váradi
- September 2013 – August 2017
- partners:
  - Research Institute for Linguistics, Hungarian Academy of Sciences
  - Institute of Behavioural Sciences, University of Helsinki

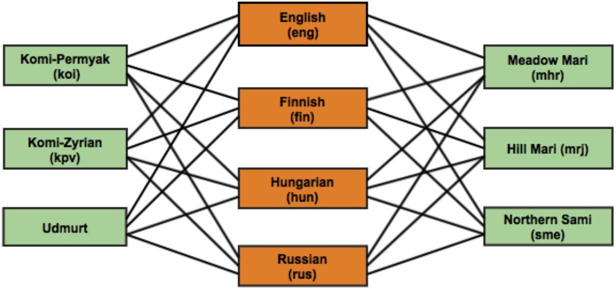
The objective of the project

Kornai (2013): a language is digitally viable only to the extent it produces new, publicly available digital material

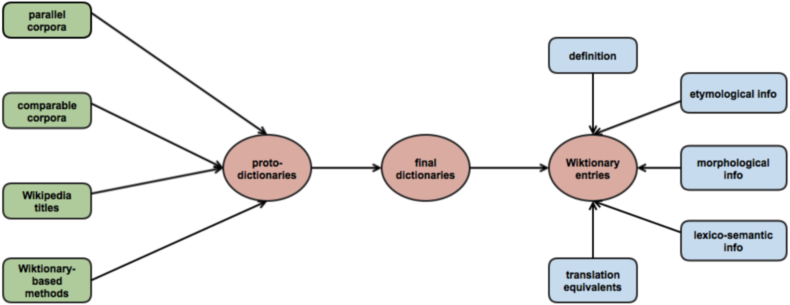
- to provide language technology support for several small Finno-Ugric digital communities in generating online content
- to support small Finno-Ugric language communities to be able to cope with some of the digitally performed functions of their native languages
- to provide language resources for further research

The objective of the project

generating dictionaries for several language pairs, and deploying the enriched lexical material on the web in the framework of Wiktionary



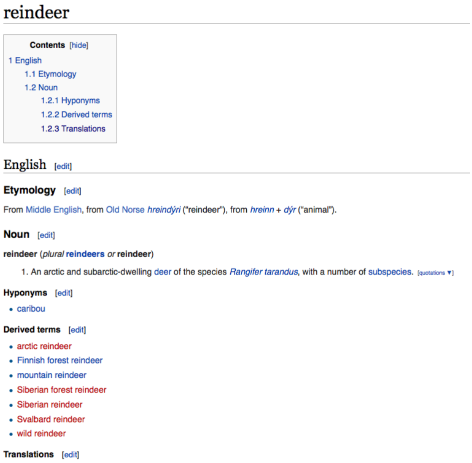
The workflow of the project



Wiktionary

- a collaborative multilingual dictionary project
- a sister project of Wikipedia
- licensed under CC-BY-SA 3.0 and GNU Free Documentation License
- Wiktionary editions in the major languages: eng, fin, hun, rus

A Wiktionary entry



reindeer

Contents [hide]

English [edit]

Etymology [edit]

From Middle English, from Old Norse *hreindýri* ("reindeer"), from *hreinn* + *dýr* ("animal").

Noun [edit]

reindeer (*plural* **reindeers** *or* **reindeer**)

1. An arctic and subarctic-dwelling deer of the species *Rangifer tarandus*, with a number of subspecies. [quotations ▼]

Hyponyms [edit]

- caribou

Derived terms [edit]

- arctic reindeer
- Finnish forest reindeer
- mountain reindeer
- Siberian forest reindeer
- Siberian reindeer
- Svalbard reindeer
- wild reindeer

Translations [edit]

Enriching the entries

generating the entries as automatically as possible

- definition: the {English, Finnish, Hungarian, Russian} equivalent of the entry
- etymological info: from UraloNet etymological database
- morphological info: from morphological analysers, from downloaded dictionaries
- lexico-semantic info: idioms, phrases from downloaded dictionaries
- translation equivalents: from the proto-dictionaries

Uploading the data

- manual validation and correction
- uploading the entries into Wiktionary
- conversion from the Wiktionary lightweight markup system to XML/RDF format
- clearing the copyright issues
- making all the generated resources publicly available

Creating proto-dictionaries

- parallel corpora
- comparable corpora
- Wikipedia title pairs
- Wiktionary-based methods

Text collection – parallel corpora

- Bible translations (Parallel Bible Corpus, Bible.is, The Unbound Bible)
- software documentation (OPUS)
- websites of officially bilingual regions of Norway, Finland and Sweden

Text collection – comparable corpora

- Wikipedia
  - downloading the dumps for the languages we are dealing with
  - extracting each interlanguage-linked article pair
  - extracting the plain text
  - considering only the first x sentences of each article in the major languages, where x is the number of sentences in the corresponding FU article
- domain-specific monolingual texts by specifying a keyword (Sami culture, education, society, etc.)
- multilingual daily newspaper materials from the same time interval and from the same region (YLE, Lapin Kansa)

Text processing

- plain text extraction from several formats (HTML, PDF)
- character normalization (UTF-8)
- language discrimination → removing text parts in other languages
- tokenization, sentence segmentation
- *morphological analysis and disambiguation*

Methods of dictionary creation – parallel corpora

1. Sentence alignment
    - Hunalign
    - result: aligned parallel sentences
    - side-product: automatically bootstrapped dictionary in the realignment phase
2. Extraction of word pairs based on some similarity metrics
    - Hundict (Dice co-efficient)
    - result: word pairs with their confidence measures

Methods of dictionary creation – comparable corpora

1. Extracting real parallel sentences
    - Hunalign/Yalign → standard methods for parallel texts
2. Applying context similarity methods
    - Hundict

Number of tokens

lang pair	**parallel**		**comparable**	
	L1	**L2**	**L1**	**L2**
sme–fin	771,749	763,234	230,011	5,312,884
sme–rus	328,019	342,984	174,352	231,914
kpv–rus	135,228	145,749	78,763	139,281
kpv–fin	121,142	129,826	65,087	97,903
mhr–eng	13,202	13,276	96,183	241,377
koi–eng	4,970	5,782	64,783	138,122
koi–hun	1,650	1,573	37,137	47,658
mrj–fin	0	0	121,023	128,469
udm–hun	0	0	39,994	56,139

Wikipedia title pairs

using the interwiki links, we created bilingual dictionaries from Wikipedia titles

lang pair	entries (#)	lang pair	entries (#)
mrj–eng	9,313	kpv–rus	3,168
mrj–rus	6,676	koi–eng	2,137
sme–fin	5,564	koi–fin	1,215
sme–hun	4,149	koi–hun	882

Wiktionary-based methods

Wikt2dict

- Parsing
  - parsing the English, Finnish, Russian and Hungarian editions of Wiktionary
  - extraction of translations from the translation tables
- Triangulating
  - discovering previously non-existent links between translations
  - further expansion of our dictionaries

Results – Wiktionary parsing

lang pair	entries (#)	lang pair	entries (#)
sme–eng	573	koi–eng	37
udm–rus	276	mrj–rus	27
kpv–rus	245	mhr–rus	25
sme–rus	227	mrj–eng	20
koi–rus	201	sme–fin	15
sme–hun	106	udm–hun	11
udm–eng	102	kpv–eng	4
mhr–eng	96		

Results – Wiktionary triangulation

lang pair	entries (#)	lang pair	entries (#)
sme–eng	3,428	koi–eng	625
sme–rus	2,972	koi–fin	379
mhr–fin	1,849	koi–rus	190
mhr–hun	941	mrj–rus	96
udm–rus	811	kpv–fin	14
udm–hun	723	kpv–eng	4

Results – All proto-dictionaries

koi–eng	2,799	mrj–eng	9,409
koi–fin	1,594	mrj–fin	5,495
koi–hun	1,194	mrj–hun	5,109
koi–rus	2,144	mrj–rus	6,799
kpv–eng	3,368	sme–eng	10,076
kpv–fin	2,776	sme–fin	8,795
kpv–hun	2,234	sme–hun	6,355
kpv–rus	3,418	sme–rus	8,365
mhr–eng	5,474	udm–eng	4,005
mhr–fin	5,143	udm–fin	2,876
mhr–hun	3,910	udm–hun	2,162
mhr–rus	5,040	udm–rus	3,606

Conclusion and future works

- these small FU languages are under-resourced → the standard dictionary building methods output dirty and small dictionaries → more manual work
- systematic test of the dictionary building methods
- collecting more texts → generating larger dictionaries
- creating and uploading the Wiktionary entries

Thank you for your attention!

simon.eszter@nytud.mta.hu
benyeda.ivett@nytud.mta.hu
koczka.peter@nytud.mta.hu

UraloNet: http://www.uralonet.nytud.hu/
Hunalign: http://mokk.bme.hu/en/resources/hunalign/
Hundict: https://github.com/zseder/hundict
Wikt2dict: https://github.com/juditacs/wikt2dict