Automatic creation of bilingual dictionaries for Finno-Ugric minority languages

Eszter Simon – Iván Mittelholcz

26 May 2017

Research Institute for Linguistics, Hungarian Academy of Sciences

Outline

Introduction

Digital revolution

- digital revolution $\rightarrow$ dramatic impact on nearly all aspects of society
- language communities are most sensitive to new paradigms of communication technologies
- information need is being covered from online, collaboratively edited material, e.g. Wikipedia
- in personal spheres of life: interaction is being conducted via social media applications
- the role of language in these novel situations is prominent, language is the vehicle $\rightarrow$ it needs to adapt itself to inevitable changes

Digital language death

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

- a language is digitally viable only to the extent it produces new, publicly available digital material
- loss of function & loss of prestige & loss of competence $\rightarrow$ language death
- small language communities: to what extent will these language communities be pervaded or corrupted by new media?
- our project aims to support Finno-Ugric (FU) language communities so that they are able to cope with some of the digitally performed functions of their native languages

Language technology

- language technology aspires to become a technology that helps people collaborate, share knowledge and participate in social interactions regardless of language barriers and computer skills
- but cutting-edge language processing applications are only available for widely-spoken languages
- based on partially existing language resources, we create freely accessible online lexical resources for small FU languages
- to provide linguistically-based support for several small FU digital communities in generating online content, and thereby promote multilingualism, and help revitalize the digital functions of these FU languages

The project

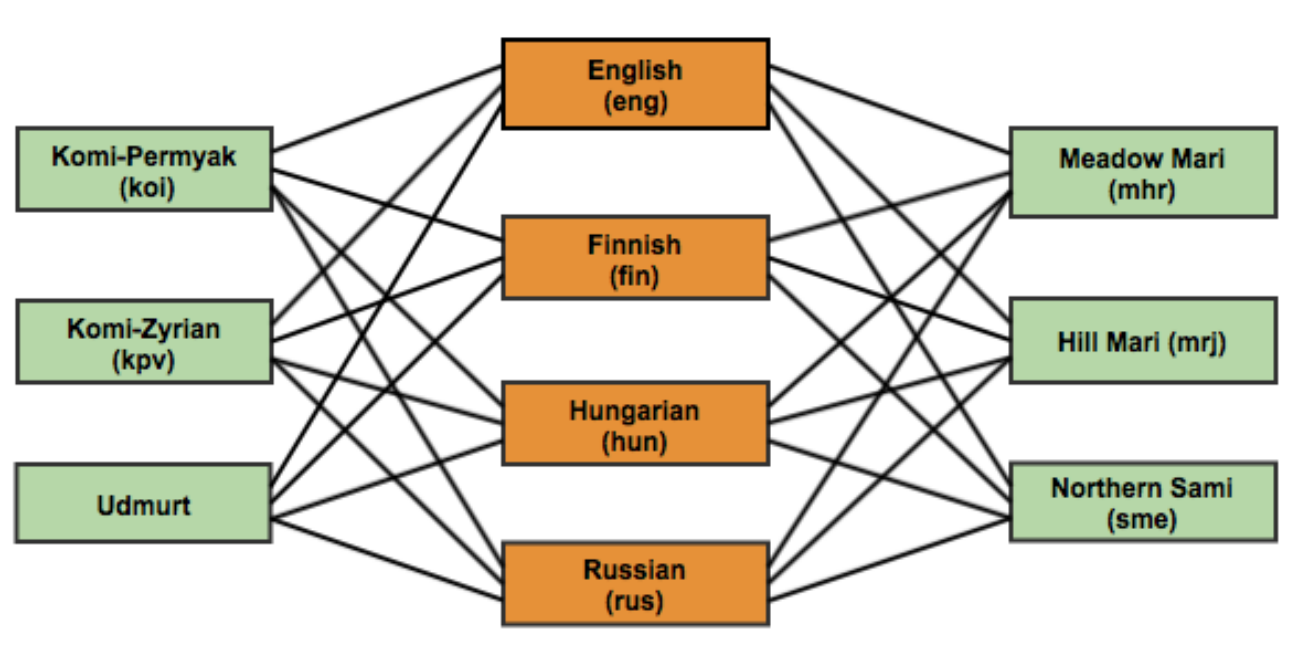
The project

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

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



The FU languages concerned

The FU languages concerned I.

Language	ISO	EGIDS	Population	Location	Writing
Saami, North	sme	2	26,000	Norway, Sweden, Finland	Latin
Mari, Meadow	mhr	4	470,000	Russia	Cyrillic
Mari, Hill	mrj	5	30,000	Russia	Cyrillic
Komi-Zyrian	kpv	5	156,000	Russia	Cyrillic
Komi-Permyak	koi	5	63,000	Russia	Cyrillic
Udmurt	udm	5	340,000	Russia	Cyrillic

The FU languages concerned II.

Northern Saami

- provincial (2): used in education, work, mass media, and government within officially bilingual regions of Finland, Norway, Sweden
- huge (and successful) revitalization efforts
- Northern Saami Wikipedia (7000+ articles), TV, radio, newspapers
  $\rightarrow$ newly produced electronic Northern Saami web material

Meadow Mari

- educational (4): in vigorous use, with standardization and literature, institutionally supported education
- co-official language of the Mariy El Autonomous Republic of the Russian Federation
- Meadow Mari Wikipedia (9000+ articles), newspapers

The FU languages concerned III.

Hill Mari, Komis, Udmurt

- developing (5): in vigorous use, with literature in a standardized form, though not yet widespread or sustainable
- co-official languages of the Mariy El Autonomous Republic, the Komi Republic and Udmurtia, respectively
- blogs and newspapers
- Hill Mari Wikipedia (10.000+ articles), Komi-Permyak Wikipedia (3000+ articles), Komi-Zyrian Wikipedia (5000+ articles), Udmurt Wikipedia (3000+ articles)

Proto-dictionaries

Bilingual dictionaries

- bilingual dictionaries play a critical role in foreign language teaching and in several NLP applications (e.g. machine translation, computational semantics)
- manual dictionary building takes a large amount of skilled work $\rightarrow$ not affordable in the case of lesser used languages
- fully automatic generation of clean bilingual resources is not possible $\rightarrow$ proto-dictionaries containing candidate translation pairs
- manual validation $\rightarrow$ input of further lexicographic work

Standard automatic dictionary building methods

Parallel corpora

1. Sentence alignment → aligned parallel sentences
2. Extraction of word pairs based on some similarity metrics → word pairs with their confidence measures

Comparable corpora

1. Extracting real parallel sentences
2. Applying context similarity methods

all methods require a large amount of (pre-processed) data and a seed lexicon

Under-resourced languages

- lacking pre-processing tools (tokenizer, sentence splitter, morphological analyser and disambiguator)
- lacking lexical resources
- even if there are monolingual data, comparable and parallel text material is far from enough
- no parallel texts for koi–{fin, rus}, mrj–ALL, udm–ALL: no Bible translation, no UN Declaration of Human Rights

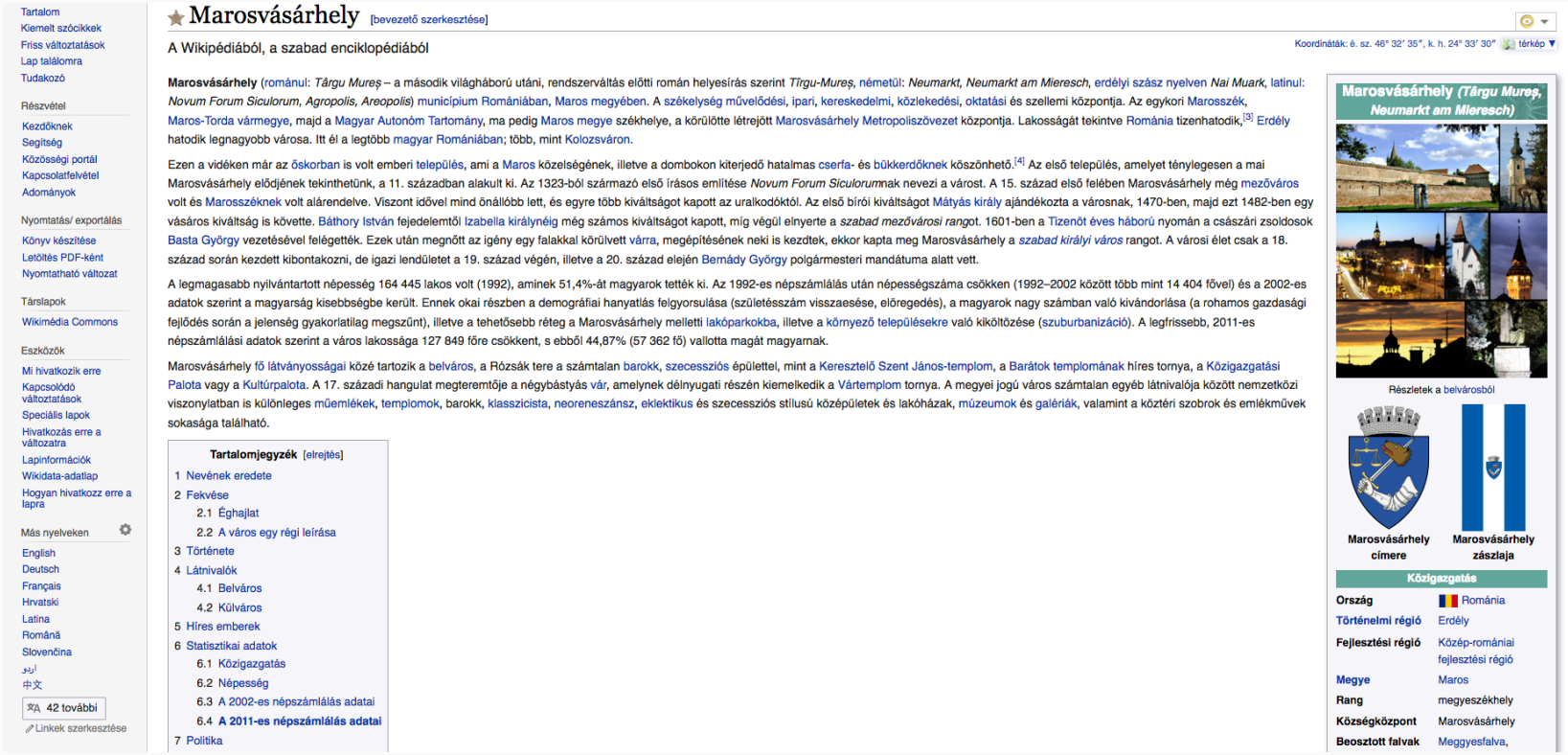
Alternative methods

utilizing crowd-sourced language resources, such as Wikipedia and Wiktionary

Wiktionary

- a collaborative multilingual dictionary project
- a sister project of Wikipedia
- free-content (CC-BY-SA 3.0 and GNU Free Documentation License)
- aims to describe all words of all languages
- editions in several languages using definitions and descriptions in the given language

Wikipedia title pairs



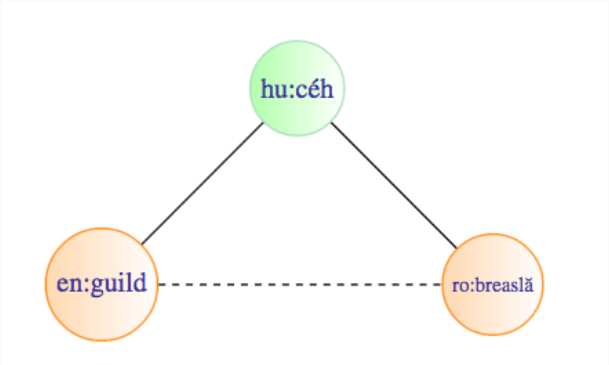
using the interwiki links, we created bilingual dictionaries from Wikipedia titles → thousands of word pairs for each language pair

Wiktionary-based methods

wikt2dict

`https://github.com/juditacs/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



Evaluation of the proto-dictionaries

only for Northern Saami

Steps

1. the proto-dictionaries for each language pair were merged
2. repeated items were filtered out
3. the merged files were manually validated by a linguist expert of Northern Saami

Instructions

- the source (S) and the target (T) word must be valid words, must be dictionary forms, and must be translations of each other
- if the source word is not a valid word, the word pair is treated as wrong
- if the source word is a valid word but not a dictionary form, the correct dictionary form should be added
- if the target word is a good translation of the source word but is not a dictionary form, the correct dictionary form should be added
- if the target word is not a good translation, a new translation should be given

Categories

- `ok-ok:` the S and the T word are valid words, are dictionary forms, and are translations of each other
- `ok-nd:` the S and the T word are valid words, are translations of each other, but the T word is not a dictionary form
- `nd-ok:` the S and the T word are valid words, are translations of each other, but the S word is not a dictionary form
- `nd-nd:` the S and the T word are valid words, are translations of each other, but none of them are dictionary forms
- `ok-wr:` the S word is a valid word and a dictionary form, but the T word is not a valid word or it is not a correct translation of the S word
- `nd-wr:` the S word is a valid word but not a dictionary form, and the T word is not a valid word or is not a correct translation of the S word
- `wr-xx:` the S word is not a valid word

Results

Results for the merged dictionaries

lang pair	sme–eng	sme–fin	sme–hun	sme–rus
all (#)	6,042	7,100	4,969	4,373
useful (%)	63.38	54.39	63.67	75.39

Results for the methods

method	WikiTitle	W2D ext	W2D tri	KDE4
all (#)	2,989	921	11,714	8,401
useful (%)	97.03	95.54	66.20	61.09

useful = ok-ok + ok-nd + nd-ok + nd-nd

Conclusions and future work

Conclusions

Conclusions

- under-resourced FU languages $\rightarrow$ alternative dictionary building methods
- the most precise method is using Wikipedia title pairs, but Wiktionary-based methods also provided useful results

Future work

- Wiktionary is not only the source, but the target as well: translation pairs enriched with obligatory pieces of linguistic information will be uploaded as new entries into Wiktionary

We want to give our results back to the community thus to support digital vitality of small FU language communities and thereby promote multilingualism.

Thank you for your attention!

simon.eszter@nytud.mta.hu
mittelholcz.ivan@nytud.mta.hu