Automatic Generation of Wiktionary Entries for Finno-Ugric Minority Languages

Zsanett Ferenczi, Iván Mittelholcz, Eszter Simon

Helsinki, 8 January 2018

Research Institute for Linguistics, Hungarian Academy of Sciences

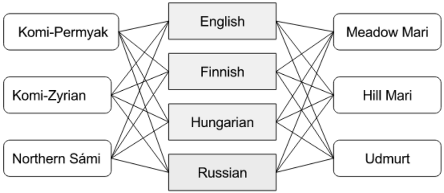
Objective of the project

General objective:

to provide linguistically based support for several small FU digital communities to generate online content and help to revitalize the digital functions of some FU minority languages

Practical objective:

to create bilingual dictionaries for six FU minority languages paired with four major languages and to deploy the enriched lexical material on the web in the framework of Wiktionary



Outline

- generating the translation units
    - creating proto-dictionaries → raw dictionaries
    - manual validation & preprocessing & consistency check → valid dictionaries = translation units
- generating the Wiktionary entries
    - providing POS tags
    - disambiguating the POS tags
    - adding IPA transcription
    - generating the final entries
- uploading the entries
- evaluation

The poster

Automatic Generation of Wiktionary Entries for Finno-Ugric Minority Languages

Zsanett Ferenczi & Iván Mittelholcz & Eszter Simon
Research Institute for Linguistics, Hungarian Academy of Sciences
{ferenczi.zsanett,mittelholcz.ivan,simon.eszter}@nytud.mta.hu

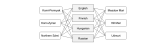
Abstract

The research presented here aims to generate online content and help to revitalize the digital functions of some Finno-Ugric (FU) minority languages. The practical objective of the research was to create bilingual dictionaries for six FU minority languages (Udmurt, Komi-Permyak, Komi-Zyrian, Meadow Mari, Hill Mari and Northern Saami) paired with four major languages which are important for these communities (English, Finnish, Hungarian, Russian) and to deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary. We give an overview of the workflow in which Wiktionary entries were fully automatically generated from automatically created and manually validated translation units. We also give a thorough evaluation, whose results show that we would multiply the number of Wiktionary entries in the aforementioned FU minority languages.

Introduction

General objective: to provide linguistically based support for several small FU digital communities to generate online content and help to revitalize the digital functions of some FU minority languages.

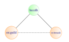
Practical objective: to create bilingual dictionaries for six FU minority languages paired with four major languages which are important for these communities and to deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary.



Generating the Translation Units

1. ***creating proto-dictionaries → raw dicts***
   (a) **Wikipedia title pairs:** we created bilingual dictionaries from Wikipedia title pairs using the interwiki links
   (b) **Wiktionary-based methods** using the Wikt2dict tool: `https://github.com/juditacs/wikt2dict`
      i. **Wikt2dict extraction:** we parsed the English, Finnish, Russian and Hungarian editions of Wiktionary and extracted translations from the translation tables

      ii. **Wikt2dict triangulation:** discovering previously non-existent links between translations with a triangulation method, which is based on the assumption that two expressions are likely to be translations, if they are translations of the same word in a third language



   (c) **downloaded dictionaries from the Opus corpus:** containing word pairs from the automatic word alignment created with GIZA++ and the Moses toolkit, only for Northern Saami–{English, Finnish, Hungarian}
2. ***manual validation & preprocessing & consistency check → valid dictionaries = translation units***

Generating the Wiktionary Entries

The manually validated word pairs were used as the source material of newly created potential Wiktionary entries, which contain several obligatory elements. These elements containing morphological and phonetical information were generated **fully automatically**. For example, in the case of the Northern Saami–Finnish language pair, the Northern Saami word would be an entry in the Finnish Wiktionary: the title of the entry would be the Northern Saami word, while its Finnish definition would be its Finnish translation equivalent.

Providing POS tags: from the output of morphological analyzers available for the languages we deal with:

- koi, kpv, mhr, mrj, sme, udm; fin, rus: Giellatekno
- hun: `emMorph`
- eng: `hunmorph`

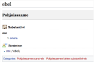
Disambiguating the POS tags: words without context → standard morphosyntactic disambiguation techniques based on contextual information cannot be used

1. only the morphological information of the given word is considered
2. horizontal comparison: the POS tags of a source word and the POS tags of the corresponding target word are compared, and we get the disambiguated POS tag from this comparison
3. vertical comparison: the sets of POS tags added to a word acting as a source word in more translation units are compared

Adding IPA transcription: gathering phonetic transcription to enrich the content of Wiktionary entries

- koi, kpv, mhr, mrj, udm: the Mari Web Project's automatic transcription tool
- sme: FST compiled from the `text2ipa` source files of the Giellatekno infrastructure

Putting the bits together: generating the final entries to be uploaded to Wiktionary



ebel

Pohjoissaame

Substantiivi

ebel

1. omena

Ääntäminen

- IPA: /ebel/

Categories: Pohjoissaamen sanat-eb | Pohjoissaamen kielen substantiivit-eb

Uploading the Entries

Uploading multiple entries to Wiktionary can be automated. MediaWiki has a bot called `Pywikibot`, that can automate work on MediaWiki sites such as Wiktionary or Wikipedia. Since automatic uploading of entries is not supported by the Wiktionary community, we have to ask for permission to upload our newly created entries into Wiktionary.

Evaluation

langs	all (#)	useful (#)	remain (#)	wikt (#)	comm (#)	new (#)	cover (%)	improv (%)
kom–eng:	2,153	2,111	656	54	25	631	46.30	1,168.52
kom–fin:	1,169	1,162	687	42	27	660	64.29	1,571.43
kom–hun:	1,063	1,025	699	152	35	664	23.03	436.84
kom–rus:	1,155	1,148	673	465	223	450	47.96	96.77
chm–eng:	4,883	4,883	1,671	347	53	1,618	15.27	466.28
chm–fin:	3,578	3,578	1,905	443	213	1,692	48.08	381.94
chm–hun:	2,589	2,589	1,634	34	12	1,622	35.29	4,770.59
chm–rus:	2,542	2,542	1,497	848	202	1,295	23.82	152.71
sme–eng:	6,041	5,556	2,531	4,073	882	1,649	21.65	40.49
sme–fin:	7,100	6,463	2,862	817	422	2,440	51.65	298.65
sme–hun:	4,969	4,509	2,392	206	146	2,246	70.87	1,090.29
sme–rus:	4,373	4,172	2,034	306	237	1,797	77.45	587.25
udm–eng:	2,087	2,069	754	32	15	739	46.88	2,309.38
udm–fin:	1,700	1,694	828	55	45	783	81.82	1,423.64
udm–hun:	1,204	1,198	739	128	69	670	53.91	523.44
udm–rus:	1,226	1,211	578	644	247	331	38.35	51.40

- **'all':** the total number of word pairs gathered with all methods for the language pair;
- **'useful':** the number of useful word pairs which comprise all word pairs except of the ones in which the source word is not a valid word, since correct dictionary forms and translation equivalents were manually added by human validators;
- **'remain':** the decreased number of the word pairs ready to upload (useful word pairs minus the number of word pairs for which we could not provide a POS tag minus the number of entries with the same word already existing in Wiktionary);
- **'wikt':** the number of the source language words already existing in the target language Wiktionary;
- **'comm':** the number of the common words being both in the Wiktionary and in our dictionaries;
- **'new':** the number of brand new entries created by us ('remain' minus 'comm');
- **'cover':** the ratio of the number of common words to the number of words already being in Wiktionary, thus it is the degree of overlap with Wiktionary;
- **'improv':** the ratio of the number of the new Wiktionary entries to that of the already existing ones which shows the improvement in the amount of Wiktionary entries of the given source language in the given target language edition of Wiktionary.

Conclusion and Future Work

- Wiktionary entries were created fully automatically after proto-dictionaries were validated;
- POS tags and IPA transcriptions were added to the translation pairs;
- automatically generated Wiktionary entries would multiply the number of Wiktionary entries in these FU minority languages;
- since automatic uploading of entries is not supported by the Wiktionary community, we have to ask for permission to upload our newly created entries into Wiktionary;
- we provide freely available professional online multilingual lexical data for digital communities of some FU minority languages with Wiktionary entries → also providing them in standard data formats (e.g. `tsv, XML, RDF`) via our web site or via the repositories of dictionary families such as Giellatekno and Apertium.

Acknowledgements

The research reported in the paper was conducted with the support of the Hungarian Scientific Research Fund (OTKA) grant #107885. We are grateful to the reviewers for their valuable comments.