Automatic Generation of Wiktionary Entries for Finno-Ugric Minority Languages

Zsanett Ferenczi
Research Institute for Linguistics
Hungarian Academy of Sciences
ferenczi.zsanett@nytud.mta.hu

Iván Mittelholcz
mittelholcz.ivan@nytud.mta.hu

Eszter Simon
simon.eszter@nytud.mta.hu

Abstract

The research presented in this paper 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.

Tiivistelmä

Tutkimuksen tavoitteena on tuottaa digitaalista sisältöä usealle suomalais-ugrilaiselle vähemmistökielelle, ja edistää niiden kielten elvytystä, eli pelastaa niiden uhanalaisia kieliä häviämiseltä. Tutkimuksen käytännöllisenä tavoitteena oli luoda kaksikielisiä sanakirjoja kuudelle suomalais-ugrilaiselle vähemmistökielelle (nimittäin udmurtille, komipermjakille, komisyrjäänille, niittymarille, vuorimarille ja pohjoissaamelle), yhdistettynä neljään, näille yhteisöille tärkeisiin kieliin (englanti, suomi, unkari ja venäjä). Automaattisesti luodut, sitten käsin tarkastetut, ja morfologisien ja ääntämistietojen kanssa vahvistetut käännökset ladattiin Wikisanakirjaan. Artikkelissa pyrittiin esittelemään koko prosessi tarkasti, minkä aikana Wiktionary-artikkelit luotiin kokonaan automaattisesti. Tutkimuksessa esittelemme myös, miten onnistuisimme moninkertaistamaan Wikisanakirjassa jo olemassa olevien edellä mainittujen suomalais-ugrilaisten vähemmistökielien sanojen lukumäärää.

Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018), pages 39–50,
Helsinki, Finland, January 8–9, 2018. ©2018 Association for Computational Linguistics

Kivonat

A cikkben bemutatott kutatás célja, hogy kisebbségi finnugor nyelvek számára nyelvi erőforrásokat állítson elő, melyek segíthetik ezeket a veszélyeztetett nyelvi közösségeket a revitalizálási folyamatokban. A kutatás során kétnyelvű szótárakat állítottunk elő olyan nyelvpárokra, melyeknél a forrásnyelv az udmurt, komi-permják, komi-zürjén, mezei mari, hegyi mari és északi számi nyelvek egyike, míg a célnyelv az angol, finn, magyar és orosz közül kerül ki. Az automatikusan előállított, majd kézzel ellenőrzött fordítási párok kiejtési és morfológiai információkkal kiegészítve kerülnek feltöltésre a Wiktionarybe. A cikk bemutatja a teljes munkafolyamatot, amelynek során a Wiktionary-bejegyzések teljesen automatikusan előállnak. Egy alapos kiértékelésben megmutatjuk azt is, hogy az általunk létrehozott bejegyzésekkel megsokszorozható a fent említett finnugor nyelvű szavak száma a célnyelvi Wiktionary-kiadásokban.

1 Introduction

The research presented in this paper is part of a project whose general objective is to provide linguistically based support for several small Finno-Ugric (FU) digital communities to generate online content and help to revitalize the digital functions of some FU minority languages. The practical objective of the project is 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.

Even for widely used languages, freely available professional online multilingual lexical data are scarce; exceptions being BabelNet (Navigli and Ponzetto, 2012) and open wordnets in a variety of languages, such as Multilingual Central Repository (Atserias et al., 2004) and MultiWordNet (Pianta et al., 2002). Smaller communities are often left to their own devices, which can manifest in their affinity towards mastering other languages to be able to translate or localize information that is unavailable in their native language.

In the current global economic and information space, we interact via new types of media, applications of which are e.g. Facebook, Twitter, YouTube, Wikipedia and other Wikimedia-related initiatives. Social media, powered by Web 2.0 technology – which actively involves language technology –, are becoming extremely popular, not only in the Western world where they typically originate from, but also among virtually any speech communities with internet connection. The new concepts that are brought to the smaller language communities – such as speakers of FU minority languages – are likely impact everyday lives to a bigger extent than in the case of larger speech communities, shifting new segments of native language use towards "globalized" language use. It is an empirical question to what extent, and which dimensions of the language of these speech communities – having been heavily affected by neighboring or dominating language contacts already – will be pervaded (or even corrupted) by the usage of new media.

Wiktionary[1] is a collaborative multilingual dictionary project, a sister project of Wikipedia, available under the same license (CC-BY-SA 3.0 and GNU Free Documentation License). It aims to describe all words of all languages. It has editions in sev-

[1] https://www.wiktionary.org/

eral languages using definitions and descriptions in the given language. Entries that are being maintained by a large active digital community are typically fully-fledged, whereas entries in the language domain of a small community can be very sparse, or missing. This situation can be improved by applying language technology methods and automatically creating Wiktionary entries. Using the Wiktionary infrastructure, lexical entries across FU and widely used language versions of Wiktionary can be interlinked. This will enable user communities to access rich, networked lexical material.

The aforementioned FU languages are under-resourced, hence we could not collect enough data for building parallel and comparable corpora, on which the standard dictionary building methods are based. Therefore, conducting experiments with alternative methods was needed. We made experiments with several lexicon building methods utilizing crowd-sourced language resources, such as Wikipedia and Wiktionary (see Section 2). Completely automatic generation of clean bilingual resources is not possible according to the state of the art, but it is possible to create certain lexical resources, termed proto-dictionaries, that can support lexicographic and NLP work. Proto-dictionaries contain candidate translation pairs produced by bilingual dictionary building methods.

Once the proto-dictionaries were prepared, they were merged for each language pair and repeated lines were filtered out. These files were then the object of manual validation by native speakers and linguist experts of the languages. These validated dictionaries containing translation units were the input of generating new Wiktionary entries which were created fully automatically. As the last step of the project, we upload the entries to Wiktionary.

The rest of the article is as follows. In Section 2, the workflow of generating the translation units is shortly presented. Section 3 gives an overview of the process how the Wiktionary entries are generated from the previously created translation units. In Section 4, the steps of uploading the newly created entries are described. We conducted a thorough evaluation of the coverage for proto-dictionaries created by us, which is described in Section 5. The article ends with some conclusions and plans for future work in Section 6.

2 Generating the Translation Units

For the creation of the proto-dictionaries, we applied several lexicon building methods utilizing Wikipedia and Wiktionary. For more details on the dictionary creating methods we used, see Benyeda et al. (2016) and Simon and Mittelholcz (2017) – here we only provide a short description.

Wikipedia is not only the largest publicly available database of comparable documents, but it also can be used for bilingual lexicon extraction in several ways. Following Erdmann et al. (2009), we created bilingual dictionaries from Wikipedia title pairs using the interwiki links.

Besides Wikipedia, Wiktionary is also considered as a crowd-sourced language resource which can serve as a source of bilingual dictionary extraction. Although Wiktionary is primarily for human audience, the extraction of underlying data can be automated to a certain degree. Following Ács et al. (2013) and Ács (2014), we applied the `Wikt2dict` tool[2] in two modes. First, we parsed the English, Finnish, Russian and Hungarian editions of Wiktionary and extracted translations from the so-called

[2] `https://github.com/juditacs/wikt2dict`

translation tables for the small FU languages we deal with. Second, the collection of translation pairs were expanded 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.

Besides the proto-dictionaries created by us, the large merged files for the Northern Saami–{English, Finnish, Hungarian} language pairs also contain proto-dictionaries which were not created by us but were downloaded from the Opus corpus (Tiedemann, 2009). These dictionaries contain word pairs from the automatic word alignment created with GIZA++ and the Moses toolkit.

Once the proto-dictionaries were prepared, they were merged for each language pair and repeated lines were filtered out. These raw dictionary files were then the object of manual validation by native speakers and linguist experts of the languages. The instructions for the validators were as follows. The source and the target word must be a valid word in the language concerned, they must be dictionary forms, and they must be translations of each other. If the source word is not a valid source language 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 manually added. If the target word is a good translation of the source word but is not a dictionary form, similarly to the former case, the correct dictionary form should be added. If the target word is not a good translation, a new translation should be given.

The validated dictionaries, however, were not fully clean and ready-to-use, thus several checking and correcting steps were required. As a sanity check, we checked whether the dictionary contains a source and a target word, whether any cells contain suspicious characters, etc. As a consistency check, cases when the target word was provided with a dictionary form as well as a new translation and cases when the source word was treated as wrong but a new translation were added for the target word were filtered out. A cross-language consistency check was also done, in which we checked whether source words were treated consistently in all languages. At the end of this workflow, we got the validated dictionaries containing the translation units, which served then as the input of the evaluation and the newly generated Wiktionary entries.

3 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–English language pair, the Northern Saami word would be an entry in the English Wiktionary: the title of the entry would be the Northern Saami word, while its English definition would be its English translation equivalent.

Each language edition of Wiktionary has its own rules that describe how to create new entries. These determine the structure of the entries and the pieces of information which must be included in each entry. From these descriptions of the four Wiktionary editions into which our entries were uploaded, a generalized description was created that contains the word itself, its language, its POS tag, and its translation equivalent. The only information missing from that list is the POS tag, which could be gathered from morphological analyzers available for these languages. Additional information can also be added to the entries, such as etymology or phonetic (IPA) transcription, however, these are not compulsory elements. IPA transcription is also included in our entries, since these FU languages have freely available tools that provide phonetic transcription and we wanted to enrich the Wiktionary entries with as many pieces of linguistic information as possible applying only automatic tools.

3.1 Providing POS Tags

New Wiktionary entries cannot be created without applying templates, which are provided for several word categories including POS classes. Therefore, providing the correct POS tag of a word is essential for generating a Wiktionary entry for that word. POS tags can be gathered from the output of morphological analyzers available for the languages we deal with. However, these are only words without context, thus the standard morphosyntactic disambiguation techniques based on contextual information cannot be used. Therefore, we had to find alternative ways for disambiguation, see Section 3.2.

There are available morphological analyzers for all languages we deal with that we could use to get POS tags for the words. We used the morphological analyzers of Giellatekno[3] for all of the source languages and for Finnish and Russian of the target languages. For Hungarian, we used the `emMorph` morphological analyzer (Novák et al., 2016), which is also based on the Helsinki Finite-State Technology[4] infrastructure just like the Giellatekno analyzers. For English, we used the `hunmorph` toolkit (Trón et al., 2005) with English-specific `aff` and `dic` files created from English lexicon and grammar files of `morphdb`, an open source morphological database (Trón et al., 2006). Since we work with different kinds of morphological analyzers providing different output formats, a kind of normalization of tags was needed. Having the normalized tagset, there is no difference in the format of analyses, so that the tags can be used in further steps without having different notations for the same POS tag.

Due to the fact that morphological analyzers only give analysis for single words, multi-word expressions (MWEs) had to be handled differently. In these cases, the last element of the MWE was split, and the MWE was temporarily substituted by its last word. The hypothesis behind this solution is that FU languages are typically head-final languages, thus the head follows its complements, i.e. the head is at the end of the phrase. Therefore, if we get the POS tag of the last element of the phrase, we will know the POS tag of the whole phrase. However, English and Russian are said to be strongly head-initial languages, moreover, even the FU phrase is not always head-final, thus the last element of a MWE in our dictionaries is not always the head. Handling of this phenomenon is described in Section 3.2.

Some validators inserted the particle ‘to’ before the English translation of verbs. This particle was removed from the input of the morphological analyzer but was kept as a background info and was used in the disambiguating step. The English analyzer gives back many possible analyses for a single word, since most of the English words can be a noun and a verb at the same time. There are cases when the disambiguation is difficult or almost impossible without this extra information. In the English Wiktionary, the ‘to’ particle must be included in the definition before each English verb, therefore they are later pasted back before the verb.

The output of this process contains five columns that consist of the source word, its possible POS tags, the target word, its possible POS tags and a column that contains information about the ‘to’ particle.

[3] `http://giellatekno.uit.no`

[4] `https://hfst.github.io/`

3.2 Disambiguating the POS Tags

Disambiguating the POS tags happens in circles. First, we only consider the morphological information of the given word. The second step of the POS disambiguation is a horizontal comparison, when 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. The third step is a vertical comparison, in which the sets of POS tags added to a word acting as a source word in more translation units are compared.

3.2.1 Considering the Morphological Information

Not only the POS tag of the output of the morphological analyzers are utilized, but we keep the lemma, and information on the case and the number. When a word has more than one analysis, a decision has to be made, and these pieces of morphological information can help in this process.

A kind of filtering is possible based on the assumption that the lemma and the input word are the same. Since only dictionary forms were sent to the analyzers, those tags that are the analysis of a non-dictionary form are rejected. For example, in Hungarian, the dictionary form is the nominative singular form in the case of nouns and adjectives and the present 3rd person singular indefinite form in the case of verbs. For example, if the input word is *várat*, the possible analyses are `vár[/N] + at[Acc]`, and `vár[/V] + at[_Caus/V] + [Prs.NDef.3Sg]`, thus the POS tag can only be `V` (verb), since the other one is an accusative word form, not a dictionary form.

However, there are cases when none of the lemmas is equal to the input word. In the case of MWEs, the lemma is only the last element of the input word, and therefore they must match at the end of the string. If none of the analyses matches the conditions, the set of possible POS tags is left empty.

A filtering happens at the end of this stage, because there are cases when a word gets more than one POS tag, and yet, it can contain redundant information. We keep the POS tags that are the most precise ones, e.g. if the set of POS tags contains both `N` and `Prop` (proper noun), then `N` is removed from the set.

3.2.2 Horizontal Comparison

In this step, the disambiguation of POS tags happen based on the comparison of the POS tag set of the source word and that of the target word. A translation pair has two sets of possible POS tags, and assuming that the words participating in a translation pair belong to the same POS, these tags can be reduced in number.

However, it is not correct in all cases. Not all source words has a one-word translation in the target language, and in such cases, the validator gave a MWE that seemed to be the most correct translation of the source word. Since MWEs may not have the head at the end of the phrase, they do not belong to the same POS as the original word.

Within the horizontal comparison, we investigate the intersection of the POS tag set of the source word and that of the target word. The following cases come from this comparison.

When the intersection of the two sets does not contain any POS tags, a decision has to be made in order to get some results for those translation units as well. Once, if the analyzer did not provide any output for the source word, it is the target word that determines the POS tag of the translation. Second, if the set of POS tags of the target word is empty, the source word is the one that determines. In those cases, when neither of the words has a possible tag, the translation candidate has to be removed, because no correct POS tag can be provided.

Another possible difficulty is that a single POS tag cannot be determined because of the fact that English and Hungarian phrases were split up by the last space, and although in Hungarian the head of a phrase is likely to be at the end, in English it is only so in some noun phrases. If the target word is a MWE, the target language is English, and the possible POS tags do not contain the `N` tag, these candidates are removed from the list of possible Wiktionary entries.

If the intersection of the two sets contains only one element, it is treated as the correct POS tag for that translation pair. If a correct POS tag is found, the result is saved into a list with the source word, because it may be used in the vertical comparison.

When the translation pair has more than one POS tag in common, the number of common tags is tried to be reduced by some rules. One of them is based on that the verbs of the FU languages have a particular ending (e.g. Northern Saami verbs end with '-t'), so if the source word has this verb ending, and the `V` tag is found among the common tags, then the source word is possibly a verb and is marked as that.

The number of the common POS tags can also be reduced for verbs, if the fifth column in the input contained the word 'TO'. It means that the validator inserted an extra 'to' before the English target word. Since it is a manually added information, it is assumed to be a reliable information about the POS of the word.

3.2.3 Vertical Comparison

When the sets of the POS tags of the source word and that of the target word have only one POS tag in common, the result is saved in a list with the source word, and it can be used for disambiguation. This is based on the observation that if a source word occurs in more than one translation unit, its corresponding target words are synonyms in most of the cases. Therefore, we assumed that the source word has the same POS tag in all of the translation units. When two sets have more than one POS tag in common, it is checked whether the source word has a former meaning with only one possible tag.

There are, however, cases when each translation unit with the same source word has multiple POS tags. In this case, the aforementioned method cannot be used, but those can still be disambiguated, if their sets are compared. For example, the Komi-Permyak word *ань* has three different equivalents in English: *female*, *mother*, and *woman*. These words have different sets of POS tags, namely *female* is marked as a `N` and as an `A` (adjective), while *mother* and *woman* have the tags `N` and `V`. The intersection of these three sets is undoubtedly `N`. A specific case of this is when the source word and the target word also have more POS tags, and all of them are correct. For example, the Meadow Mari word *нарынче* ('yellow') is an adjective and a noun, just like in English. In these cases, both tags are kept.

This process outputs three columns: the source word, the target word, and the correct POS tag. If a translation unit has more than one POS tag, the first two columns are repeated, thus it is treated as a new translation unit.

3.3 Adding IPA Transcription

The next step was to gather phonetic transcription to enrich the content of Wiktionary entries. We used the Mari Web Project's automatic transcription tool (Bradley, 2017) for generating IPA transcription for Hill Mari, Meadow Mari, Komi-Permyak, Komi-Zyrian, and Udmurt. For Northern Saami, we used an FST compiled from the `text2ipa` source files of the Giellatekno infrastructure[5].

All of the source FU languages has a transcription tool available, so every source word was sent to the tool and the result was saved so that it could be used when generating entries. The only problem occurred when the string contained digits and when proper nouns were sent to the transcription tool. Since the pronunciation of proper nouns might differ from the phonetics of the language, IPA transcription was not added to entries having only a proper noun as POS tag or entries having a digit in the source word.

3.4 Putting the Bits Together

Having all pieces of information, the next step is putting them together thus generating the final entries to be uploaded to Wiktionary in the last step. Although different editions of Wiktionary have different rules determining the structure of the articles, it was possible to create a template that covers all four editions to which the generated entries would be uploaded. (Consider that the languages called as target languages so far are now the languages of the Wiktionary editions to which the entries containing source words are to be uploaded.)

Before generating actual entries, it must be checked whether the word already exists in Wiktionary, and some further modifications concerning the existing data also had to be made. First, those words that already exist in the given edition of Wiktionary are filtered out: entries for those words which are in the last Wiktionary dump are not generated. Second, if a source word has more than one translation, the translation units can have the same POS tag, and in this case, they must be listed under the same POS header. If the translation unit has more than one POS tag, the translation must be repeated under each POS header in the entry.

After having extracted the words to be uploaded and having the list of translations for each POS tag, entries can be created. Each entry has a headword which is the source word. When uploading to Wiktionary, `Pywikibot` (see Section 4) will create a page that has the same name as this headword. Each entry contains one or more POS headers, and one or more translations under each header. If a source word is an existing word in more languages, then these two (or more) entries have to be merged and listed under the same headword. At the end of this step, an output file is created which meets the requirements of the input file of `Pywikibot`.

4 Uploading the Entries

Uploading multiple entries to Wiktionary can be automated. MediaWiki has a bot called `Pywikibot`[6], that can automate work on MediaWiki sites such as Wiktionary or Wikipedia. This library has a script called `pagefromfile`[7], which allows to create pages on Wiktionary (or other MediaWiki sites) from text files. That script reads the file and recognizes the template that can be configured, and it will create Wiktionary entries according to these. Each page must be separated by some characters, and each headword is used to define the name of the page. We run it with the option `--safe`,

[5] https://victorio.uit.no/langtech/trunk/langs/sme/src/phonetics/

[6] https://www.mediawiki.org/wiki/Manual:Pywikibot

[7] https://www.mediawiki.org/wiki/Manual:Pywikibot/pagefromfile.py

which means that if a certain page already exists, the bot will not upload or refresh the existing page but skips it.

Fully automated uploading of large amounts of newly created Wiktionary entries is however not supported in the Wiktionary community. We have to ask the administrators of each Wiktionary edition to allow us to upload our entries. Unfortunately, we did not get the permission from all Wiktionary editions, therefore, now we can only provide numbers based on the last downloaded Wiktionary dumps, see Table 1.

5 Evaluation

The manual validation and correction of the automatically generated proto-dictionaries has a twofold aim. First, the performance of dictionary creating methods can be compared. For more details on the results, see Simon and Mittelholcz (2017). Second, we get the number of word pairs which can be used for upload to the Wiktionary. In this section, we provide a thorough evaluation of generating Wiktionary entries.

Measuring of the coverage of a dictionary is far from trivial. It can be measured by comparing it to a word list of a corpus, or to a frequency list generated from a corpus. Or, it can be measured by comparing the number of its entries to that of another – ideally hand-crafted – dictionary. Since our newly created word pairs are to be transformed into Wiktionary articles, for this purpose, here we used Wiktionary, which is not an expert-built lexicon but manually edited by thousands of contributors.

Table 1 contains the figures for this evaluation. We use ISO 639-3 language codes for the individual languages: koi: Komi-Permyak, kpv: Komi-Zyrian, mhr: Meadow Mari, mrj: Hill Mari, sme: Northern Saami, udm: Udmurt; eng: English, fin: Finnish, hun: Hungarian, rus: Russian. However, several Wiktionary editors do not differentiate between individual languages but use macrolanguage codes (chm for Mari languages, kom for Komi languages), therefore we had to merge the dictionaries for the two Mari and for the two Komi languages.

The first column of the table (‘all’) shows the total number of word pairs gathered with all methods for the language pair. As can be seen, hundreds or thousands of translation candidates were generated for each language pair. However, not all of these word pairs are correct translation candidates, therefore we needed to extract the useful word pairs from the merged dictionary for each language pair. The second column (‘useful’) shows 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.

As mentioned above, our Wiktionary articles are generated fully automatically. The POS tag of an entry is a compulsory element of an article, which is gathered from the output of morphological analyzers through several disambiguating steps, as detailed in Section 3.1 and 3.2. The number of the useful word pairs drops in line with the increase of source language words for which we could not provide a POS tag. Before uploading new entries, it must be checked whether an entry with the same word already exists in Wiktionary. If yes, it also decreases the number of uploadable word pairs. Column ‘remain’ contains the decreased number of the word pairs ready to upload. We have also got the number of the source language words already existing in the target language Wiktionary (‘wikt’), along with the number of the words being in both lists (‘comm’). These numbers come from the Wiktionary dumps[8] and are

[8] Wiktionary dumps used in the evaluation: eng: 06-Nov-2017, fin: 05-Nov-2017, rus: 07-Nov-2017, hun:

"theoretical" numbers in the sense that they are not the numbers of actually uploaded entries, which can only be known after uploading.

From the columns 'wikt' and 'comm', the number of brand new entries ('new') created by us can be easily counted, along with a kind of coverage ('cover'), which is a 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. Consider that the coverage for each language pair drops as the size of the relevant Wiktionary grows. The last column ('improv') contains the ratio of the number of the new Wiktionary entries to one 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.

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

Table 1: Results for the language pairs.

6 Conclusion and Future Work

Wiktionary is not only used for extracting data from it, but we want to give our results back to the community, thus translation pairs enriched with obligatory pieces of linguistic information are to uploaded as new entries into Wiktionary. Before uploading new entries, it is needed to be checked whether an entry with the same word already exists in Wiktionary. From this, the number of brand new entries created by us can be easily counted, along with a kind of coverage and improvement in the number of Wiktionary entries. As can be seen from the results, the latter is very impressive, thus, with our dictionaries, we would multiply the number of Wiktionary entries in the aforementioned FU minority languages. Since automatic uploading of entries is

06-Nov-2017.

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. However, lexical data can be provided in several other ways. We plan to make them available in standard data formats (e.g. `tsv, XML`) which are easy to apply in further lexicographic or NLP subtasks. We also want to convert our data into the data format following the conventions of linguistic linked open data and provide them via our web site or via the repositories of dictionary families such as Giellatekno and Apertium.

Acknowledgments

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.

References

J. Ács. 2014. Pivot-based multilingual dictionary building using Wiktionary. In *9th Language Resources and Evaluation Conference*. ELRA, Reykjavik.

J. Ács, K. Pajkossy, and A. Kornai. 2013. Building basic vocabulary across 40 languages. In *6th Workshop on Building and Using Comparable Corpora*. ACL, Sofia, pages 52–58.

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. 2004. The MEANING Multilingual Central Repositoryse. In *Proceedings of the Second International Global WordNet Conference (GWC'04)*. Brno, Czech Republic.

Ivett Benyeda, Péter Koczka, and Tamás Váradi. 2016. Creating seed lexicons for under-resourced languages. In *GLOBALEX 2016 workshop*. ELRA, Portorož.

Jeremy Bradley. 2017. Transcribe.mari-language.com. *Acta Linguistica Academica* 64(3):369–382. https://doi.org/10.1556/2062.2017.64.3.3.

M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. 2009. An Approach for Extracting Bilingual Terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications* 5(4):1–17.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193:217–250.

Attila Novák, Borbála Siklósi, and Csaba Oravecz. 2016. A new integrated open-source morphological analyzer for Hungarian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing and Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*. Mysore, India, pages 293–302.

Eszter Simon and Iván Mittelholcz. 2017. Evaluation of Dictionary Creating Methods for Under-Resourced Languages. In Kamil Ekštein and Václav Matoušek, editors, *Text, Speech and Dialogue*. Springer International Publishing, Prague, Czech Republic, volume 10415 of *Lecture Notes in Artificial Intelligence*, pages 246–254.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolas Nicolov, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007*, John Benjamins, Borovets, pages 237–248.

Viktor Trón, Gyögy Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. 2005. Hunmorph: Open Source Word Analysis. In *Proceedings of the ACL Workshop on Software*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 77–85.

Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. pages 1670–1673.