View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repository of the Academy's Library

Evaluation of Dictionary Creating Methods for Finno-Ugric Minority Languages

Zsanett Ferenczi, Iván Mittelholcz, Eszter Simon, Tamás Váradi
Research Institute for Linguistics, Hungarian Academy of Sciences,
H-1068 Budapest, Benczúr u. 33.
{ferenczi.zsanett, mittelholcz.ivan, simon.eszter, varadi.tamas}@nytud.mta.hu

Abstract

In this paper, we present the evaluation of several bilingual dictionary building methods applied to {Komi-Permyak, Komi-Zyrian, Hill Mari, Meadow Mari, Northern Saami, Udmurt}–{English, Finnish, Hungarian, Russian} language pairs. Since these Finno-Ugric minority languages are under-resourced and standard dictionary building methods require a large amount of pre-processed data, we had to find alternative methods. In a thorough evaluation, we compare the results for each method, which proved our expectations that the precision of standard lexicon building methods is quite low for under-resourced languages. However, utilizing Wikipedia title pairs extracted via inter-language links and Wiktionary-based methods provided useful results. The newly created word pairs enriched with several linguistic information are to be deployed on the web in the framework of Wiktionary. With our dictionaries, the number of Wiktionary entries in the above mentioned Finno-Ugric minority languages can be multiplied.

Keywords: bilingual dictionaries, evaluation, under-resourced languages, dictionary building methods

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 revitalize the digital functions of some FU minority languages. The practical objective of the project is to create bilingual dictionaries for six small FU languages (Komi-Permyak, Komi-Zyrian, Meadow Mari, Hill Mari, Northern Saami, and Udmurt) paired with four major languages that are important for these small communities (English, Finnish, Hungarian, Russian) as well as to deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary.

The status of any particular language of the world is usually described using the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Lewis and Simons, 2010), which gives an estimation of the overall development versus endangerment of the language. In this scale, the highest level is 0, where languages are world-wide used *koiné*s, while languages on level 10 are already extinct. Northern Saami is on the highest level among the above mentioned FU languages: its level is 2 (provincial), thus it is used in education, work, mass media, and government within some officially bilingual region of Norway, Sweden, and Finland. In the case of the Meadow Mari language, the EGIDS level is 4 (educational), which means that it is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education. The EGIDS level of the other FU languages (Komi-Permyak, Komi-Zyrian, Hill Mari, Udmurt) is 5, i.e. they are developing, which means that there is literature which is available in a standardized form, though it is not yet widespread or sustainable.

The above mentioned FU languages are not endangered but under-resourced, hence we could not collect enough data for building parallel and comparable corpora, on which the standard dictionary building methods are based. The standard approach of bilingual lexicon extraction from parallel and comparable corpora is based on context similarity methods (e.g. Fung and Yee (1998; Rapp (1995)). Recently, source and target vectors are learned as word embeddings in neural networks based on gigaword corpora (e.g. Vulić and Moens (2015)). However, these methods need a large amount of (pre-processed) data and a seed lexicon which is then used to acquire additional translations of the context words. One of the shortcomings of this approach is that it is sensitive to the choice of parameters such as the size of the context, the size of the corpus, the size of the seed lexicon, and the choice of the association and similarity measures.

For these reasons, the above mentioned standard dictionary building methods cannot be used for our purposes. Therefore, it was necessary to conduct experiments with alternative methods. We made experiments with several lexicon building methods utilizing crowd-sourced language resources, such as Wikipedia and Wiktionary (Simon et al., 2015; Benyeda et al., 2016). 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. Depending on the method used, they either comprise more incorrect translation candidates and provide greater coverage, or provide precise word pairs at the expense of some decrease in recall; their right size depends on the specific needs.

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 methods used for creating the proto-dictionaries are shortly presented. We conducted thorough evaluation of the dictionaries produced for each language pair. In Section 3., the results of the evaluation is presented: in Section 3.1., we present the workflow of the manual validation of the automatically generated dictionaries, in Section 3.2., we detail the precision of each dictionary creating method applied here, and in Section 3.3., we figure out a kind of coverage for the newly created dictionaries in each language pair. The article ends with some conclusions and future directions in Section 4..

2. Creating the Proto-dictionaries

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 can also be used for bilingual lexicon extraction in several ways. For example, Erdmann et al. (2009) used pairs of article titles for creating bilingual dictionaries, which were later expanded with translation pairs extracted from the article texts. Mohammadi and Ghasem-Aghaee (2010) extracted parallel sentences from the English and Persian Wikipedia using a bilingual dictionary generated from Wikipedia titles as a seed lexicon. We followed the approach which is common in both articles, thus we created bilingual dictionaries from Wikipedia title pairs using the interwiki links.
Besides Wikipedia, *Wiktionary* is also considered as a crowd-sourced language resource that 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. Ács et al. (2013) extracted translations from the so-called translation tables. Since their tool `Wikt2dict` is freely available[1], we could apply it for our language pairs. We parsed the English, Finnish, Russian and Hungarian editions of Wiktionary looking for translations in the small FU languages we deal with.
Ács (2014) expanded the collection of translation pairs, discovering previously non-existent links between translations with a triangulation method. It 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.

3. Evaluation

The proto-dictionaries for each language pair were merged, and repeated lines were filtered out. Besides the above mentioned proto-dictionaries, the large merged file also contains a proto-dictionary which was not created by us but was downloaded from the Opus corpus (Tiedemann, 2009). For the Northern Saami–{English, Finnish, Hungarian} language pairs, there are available dictionaries which are lists of "reliable" alphabetic token links extracted from the automatic word alignment created with GIZA++ and the Moses toolkit. First, word pairs where the source and target words were character-level equivalents of each other were removed, since they are probably incorrect word pairs and remaining parts after (or in the lack of) boilerplate removal. The remaining part of the dictionary was also merged into the large dictionary, serving as an interesting example of applying standard lexicon extraction tools for an under-resourced language. The text material from which the Opus proto-dictionaries come is a parallel corpus of KDE4 localization files, where the Northern Saami–English parallel data contain 0.9M tokens, the Northern Saami–Finnish data contain 0.6M tokens, and the Northern Saami–Hungarian data contain 0.8M tokens. At the time of creating the proto-dictionaries, there were no dic files available for the other language pairs besides the three mentioned above in the Opus corpus.

3.1. Manual Validation

The large merged files were then manually validated by native speakers and linguist experts of the FU 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 word in the FU language, 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, the correct dictionary form should be added. If the target word is not a good translation, a new translation should be given.
The following categories come from these instructions:

- ok-ok: The source and the target word are valid words, they are dictionary forms, and they are translations of each other.
- ok-nd: The source and the target word are valid words, they are translations of each other, but the target word is not a dictionary form.
- nd-ok: The source and the target word are valid words, they are translations of each other, but the source word is not a dictionary form.
- nd-nd: The source and the target word are valid words, they are translations of each other, but none of them are dictionary forms.
- ok-wr: The source word is a valid word, it is a dictionary form, but the target word is not a valid word or it is not a correct translation of the source word.
- nd-wr: The source word is a valid word but not a dictionary form, and the target word is not a valid word or it is not a correct translation of the source word.
- wr-xx: The source word is not a valid word.

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 made sure that

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

the dictionary contains a source and a target word, checked 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 was 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.

As mentioned in Section 1., the manually validated word pairs are used as the source material of newly created Wiktionary entries, which contain several obligatory elements. These elements containing morphological and phonetical information are generated fully automatically. For example, in the case of the Northern Saami–English language pair, the Northern Saami word will be an entry in the English Wiktionary: the title of the entry will be the Northern Saami word, while its English definition will be its English translation equivalent.

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. Second, we get the number of word pairs which can be used for upload to the Wiktionary.

3.2. Precision

Category tags given to word pairs in the merged dictionaries were projected onto the corresponding word pairs in the proto-dictionaries. Results for each method were then summed up across all language pairs, as can be seen in Table 1. Besides category tags, the total number of dictionary entries of proto-dictionaries is presented in the first column. Abbreviations of the name of the methods are as follows: W2D ext: `Wikt2dict` extraction mode, W2D tri: `Wikt2dict` triangulation mode, WikiTit: Wikipedia title pairs, Opus: dic files downloaded from the Opus corpus.

In Table 1, methods are presented in a descending order based on their performance in the ok-ok category. This score is the precision of a method, i.e. the ratio of the number of the correct word pair to the total number of word pairs. Depending on the research purpose in question, word pairs containing non-dictionary forms can also be treated as correct translations, thus precision metrics may vary among approaches. Here we use it in a strict sense, thus a word pair is correct iff it is in the ok-ok category.

Some precision-like metrics are generally used for the evaluation of automatically generated bilingual dictionaries. For example, Vulić et al. (2011) use Precision@1 score, which is the percentage of words where the first word from the list of translations is the correct one, and mean reciprocal rank (MRR), where for a source word w, $rank_w$ denotes the rank of its correct translation within the retrieved list of potential translations. All these metrics are based on the assumption that the method used produces a list of translation candidates along with some confidence or probability measures. Even though it is not the case in our work, we can treat figures in the ok-ok column in Table 1 as Precision@1 scores calculated for a one-unit list of translation candidates.

The most precise method is using `Wikt2dict` in extraction mode thus extracting translation equivalents from Wiktionary translation tables. Word pairs coming from this method are quite reliable, since Wiktionary entries are manually created.

The second method is using `Wikt2dict` in triangulation mode, but there is a 15% decrease in the precision of this method compared to that of the first one. As this method does not directly use manually created links, its output may contain incorrect translations. The ok-wr figure for this method is the highest, mainly due to polysemy.

Wikipedia has very valuable translation texts since these translations were manually made by editors. Therefore, it is quite surprising, that using Wikipedia title pairs as a dictionary proved to be just the third most precise method. Consider the high nd-nd figure, which may be due to the fact that Wikipedia titles sometimes are not lemmas but plural forms, for example in the case of the names of families of plants and animals.

The worst result was produced by the method used in the Opus corpus, which is a standard dictionary building method based on parallel text material, using standard alignment and word pair extraction tools developed for well-resourced languages. Figures of this method are more flat, i.e. word pairs are distributed more uniformly across the categories compared to the other methods. This may be due to several reasons. First, the Opus dictionaries were generated from running text containing inflected and derived word forms and lemmas as well. Therefore, the number of non-dictionary forms and wrong translations is higher. (Inflected word forms were treated as valid words in non-dictionary form, while derived forms were categorized as wrong by the validators.) Second, the tools used within the Opus corpus project are not really feasible for under-resourced languages, therefore they produced more non-dictionary forms and wrong word pairs.

The large merged dictionary of each language pair was then evaluated for each category described in 3.1.; the results can be seen in Table 2. 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.

The first column of the table 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. The best language pair in this sense is sme–fin, which may be because Northern Saami is by far the best-resourced minority language of the ones we deal with, and it is an official language in several regions of Finland, where the Saami–Finnish bilingual population is quite large.

Since the validated dictionaries are the input of generating new Wiktionary entries, we need to extract all useful word pairs from the merged dictionary for each language pair. The second column of the table contains the ratio of the number of useful word pairs to the number of all word pairs.

method	all (#)	ok-ok (%)	ok-nd (%)	nd-ok (%)	nd-nd (%)	ok-wr (%)	nd-wr (%)	wr-xx (%)
W2D ext	1,965	71.76	1.22	5.75	15.17	4.63	0.36	0.76
W2D tri	23,066	56.61	1.79	2.98	3.06	30.38	1.1	3.82
WikiTit	16,854	54.11	2.97	5.57	32.5	2.92	0.49	0.75
Opus	8,401	27.57	3.99	10.4	18.64	13.99	14.57	10.69

Table 1: Results for the methods.

lang pair	all (#)	useful (%)	ok-ok (%)	ok-nd (%)	nd-ok (%)	nd-nd (%)	ok-wr (%)	nd-wr (%)	wr-xx (%)
koi–eng	1,251	96.64	74.82	0.16	7.83	0.00	13.67	0.16	3.36
koi–fin	592	98.82	65.20	3.04	9.97	0.84	19.59	0.17	1.18
koi–hun	540	93.15	70.19	3.33	4.63	1.30	13.52	0.19	6.85
koi–rus	611	98.85	65.47	2.95	16.69	1.47	11.62	0.65	1.15
kpv–eng	902	100.00	66.30	0.22	0.55	30.16	2.55	0.22	0.00
kpv–fin	577	100.00	57.89	3.29	0.69	37.09	0.87	0.17	0.00
kpv–hun	523	99.81	49.71	1.34	0.96	43.98	3.82	0.00	0.19
kpv–rus	544	100.00	63.60	8.64	9.93	14.52	3.31	0.00	0.00
mhr–eng	2,549	100.00	44.41	2.55	4.04	22.40	26.09	0.51	0.00
mhr–fin	2,565	100.00	50.80	1.05	3.31	20.74	23.63	0.47	0.00
mhr–hun	1,647	100.00	52.64	0.97	5.89	25.20	14.15	1.15	0.00
mhr–rus	1,707	100.00	40.01	2.11	4.28	17.28	35.56	0.76	0.00
mrj–eng	2,334	100.00	44.09	0.17	9.04	43.10	3.08	0.51	0.00
mrj–fin	1,013	100.00	20.24	7.70	9.77	52.32	8.59	1.38	0.00
mrj–hun	942	100.00	34.18	4.99	12.95	41.08	5.20	1.59	0.00
mrj–rus	835	100.00	27.07	11.26	9.58	31.38	16.89	3.83	0.00
sme–eng	6,041	91.97	47.57	3.77	7.33	6.56	21.65	5.08	8.03
sme–fin	7,100	91.03	42.03	3.42	5.42	12.56	19.92	7.66	8.97
sme–hun	4,969	90.78	48.48	1.67	6.72	6.62	17.05	10.24	9.22
sme–rus	4,373	95.40	71.35	0.50	2.56	0.18	20.05	0.75	4.60
udm–eng	2,087	99.14	77.19	3.07	0.91	0.29	17.59	0.10	0.86
udm–fin	1,700	99.65	49.12	2.06	1.06	18.82	28.06	0.53	0.35
udm–hun	1,204	99.50	57.14	1.74	1.50	23.17	15.45	0.50	0.50
udm–rus	1,226	98.78	8.56	2.04	0.98	65.25	20.64	1.31	1.22

Table 2: Results for the merged dictionaries.

In this case, useful word pairs comprise all word pairs minus the wr-xx category, since correct dictionary forms and translation equivalents were manually added by the human validators.

The remaining columns contain the results for each category coming from the instructions given to the validators (see Section 3.1.). A typical pattern can be recognized: if the ok-ok figure is low, the nd-nd figure will be high. It may be because of the high number of non-dictionary forms amongst Wikipedia titles, such as in the case of families of animals and plants, e.g. мерач-влак ∼ *nyúlfélék* ∼ *Leporidae*.

3.3. Coverage

If the number of the created dictionary entries can be treated as a kind of coverage, based on the figures of Table 1, it can be said that the `Wikt2dict` triangulation method has the best coverage, since it produced the largest number of translation candidates. As usual, the most precise method has the lowest coverage. We could gather much more word pairs from Wikipedia titles than from Wiktionary translation tables, which is likely due to the fact that Wikipedia contains more articles compared to the number of translations in Wiktionary's translation tables. Moreover, the number of entries highly depends on the activity of editors knowing these FU languages and willing to create new entries.

Coverage of a dictionary can also be measured by comparing the number of its entries to that of a 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 3 contains the figures for this kind of coverage evaluation. 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, 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 percentage of useful word pairs, while the third column ('useful #') contains the number of useful word pairs. In this case, useful word pairs comprise all word pairs minus the wr-xx category, since correct dictionary forms and translation equivalents were manually added by human validators.

As mentioned above, our Wiktionary articles are generated fully automatically. The part-of-speech (POS) tag of an entry is a compulsory element of an article, which is gathered from the output of morphological analyzers available for these languages through several disambiguating steps, as detailed in Ferenczi et al. (2018). 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 'remaining' contains the decreased number of the word pairs ready for upload. We have also got the number of the source language words already existing in the target language Wiktionary ('Wiktionary'), along with the number of the words being in both lists ('common'). These numbers come from the Wiktionary dumps[2] and are "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 'Wiktionary' and 'common', the number of brand new entries created by us ('new') can be easily counted, along with a kind of coverage ('coverage'), 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 ('improvement') 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.

4. Conclusions and Future Work

We presented several bilingual dictionary building methods applied to the {Komi-Permyak, Komi-Zyrian, Meadow Mari, Hill Mari, Northern Saami, Udmurt}–{English, Finnish, Hungarian, Russian} language pairs. Since these FU languages are under-resourced and standard dictionary building methods require a large amount of pre-processed data, we had to find alternative methods. In a thorough evaluation, we compared the results for each method, which proved our expectations that the precision of standard lexicon building methods is quite low.

The Wiktionary-based methods proved to be the most precise, but using Wikipedia title pairs extracted via interlanguage links also provided useful results.

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 the required pieces of linguistic information are to be uploaded as new entries into Wiktionary. Before uploading new entries, it must be checked whether an entry with the same word already exists in Wiktionary. From this, the number of brand new entries created by us could 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 could multiply the number of Wiktionary entries in the above mentioned Finno-Ugric minority languages. Since automatic uploading of entries is not supported by the Wiktionary community, we must obtain the permission to upload our newly created entries into Wiktionary. We have already permitted to upload the new entries into the Finnish and Hungarian versions of Wiktionary; in the time of writing, however, we still do not have the permission from the English and Russian Wiktionary editors.

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 work. 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[3] or via the repositories of dictionary families such as Giellatekno[4].

5. Acknowledgements

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

6. Bibliographical References

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

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

Benyeda, I., Koczka, P., and Váradi, T. (2016). Creating seed lexicons for under-resourced languages. In *GLOBALEX 2016 workshop*, Portorož. ELRA.

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

[3]`http://finnotka.nytud.hu`

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

lang pair	all (#)	useful (%)	useful (#)	remaining (#)	Wiktionary (#)	common (#)	new (#)	coverage (%)	improvement (%)
kom–eng	2,153	95.26	2,051	655	54	25	630	46.30	1166.67
kom–fin	1,169	95.54	1,117	687	42	27	660	64.29	1571.43
kom–hun	1,063	95.29	1,013	699	152	35	664	23.03	436.84
kom–rus	1,155	92.54	1,069	672	465	223	449	47.96	96.56
chm–eng	4,883	98.83	4,826	1,670	347	53	1,617	15.27	465.99
chm–fin	3,578	98.57	3,527	1,903	443	213	1,690	48.08	381.49
chm–hun	2,589	98.29	2,545	1,633	34	12	1,621	35.29	4767.65
chm–rus	2,542	98.11	2,494	1,493	848	201	1,292	23.70	152.36
sme–eng	6,041	91.97	5,556	2,531	4,072	882	1,649	21.66	40.50
sme–fin	7,100	91.03	6,463	2,862	817	422	2,440	51.65	298.65
sme–hun	4,969	90.78	4,510	2,392	206	146	2,246	70.87	1090.29
sme–rus	4,373	95.40	4,172	2,034	306	237	1,797	77.45	587.25
udm–eng	2,087	99.14	2,069	751	32	15	736	46.88	2300.00
udm–fin	1,700	99.65	1,694	828	55	45	783	81.82	1423.64
udm–hun	1,204	99.50	1,198	729	128	69	660	53.91	515.62
udm–rus	1,226	98.78	1,211	578	643	247	331	38.41	51.48

Table 3: Results for the language pairs.

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

Ferenczi, Zs., Mittelholcz, I., and Simon, E. (2018). Automatic Generation of Wiktionary Entries for Finno-Ugric Minority Languages. In *Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018)*, page 39–50, Helsinki, Finland. Association for Computational Linguistics.

Fung, P. and Yee, L. Y. (1998). An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *17th International Conference on Computational Linguistics*, pages 414–420, Stroudsburg. ACL.

Lewis, M. P. and Simons, G. F. (2010). Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique*, 55(2):103–120.

Mohammadi, M. and Ghasem-Aghaee, N. (2010). Building Bilingual Parallel Corpora Based on Wikipedia. In *2nd International Conference on Computer Engineering and Applications*, pages 264–268.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Stroudsburg. ACL.

Simon, E. and Mittelholcz, I. (2017). Evaluation of Dictionary Creating Methods for Under-Resourced Languages. In Kamil Ekštein et al., editors, *Text, Speech and Dialogue*, volume 10415 of *Lecture Notes in Artificial Intelligence*, pages 246–254, Prague, Czech Republic, August. Springer International Publishing.

Simon, E., Benyeda, I., Koczka, P., and Ludányi, Zs. (2015). Automatic creation of bilingual dictionaries for Finno-Ugric languages. In *1st International Workshop on Computational Linguistics for Uralic Languages*, Tromsø.

Tiedemann, J. (2009). News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. In Nicolas Nicolov, et al., editors, *Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007*, pages 237–248. John Benjamins, Borovets.

Vulić, I. and Moens, M.-F. (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *53rd Annual Meeting of the Association for Computational Linguistics*, pages 719–725, Stroudsburg. ACL.

Vulić, I., De Smet, W., and Moens, M.-F. (2011). Identifying word translations from comparable corpora using latent topic models. In *49th Annual Meeting of the Association for Computational Linguistics*, pages 479–484, Stroudsburg. ACL.