Proceedings of 1st International Workshop in Computational Linguistics for Uralic Languages (IWCLUL 2015); ‹http://dx.doi.org/10.7557/scs.2015.2›

Automatic creation of bilingual dictionaries for Finno-Ugric languages

Eszter Simon
simon.eszter@nytud.mta.hu

Ivett Zs. Benyeda
benyeda.ivett@nytud.mta.hu

Péter Koczka
koczka.peter@nytud.mta.hu

Zsófia Ludányi
ludanyi.zsofia@nytud.mta.hu

Research Institute for Linguistics, Hungarian Academy of Sciences

December 15, 2014

Abstract

We introduce an ongoing project whose objective is to provide linguistically based support for several small Finno-Ugric digital communities in generating online content. To achieve our goals, we collect parallel, comparable and monolingual text material for the following Finno-Ugric (FU) languages: Komi-Zyrian and Permyak, Udmurt, Meadow and Hill Mari and Northern Sami, as well as for major languages that are of interest to the FU community: English, Russian, Finnish and Hungarian. Our goal is to generate proto-dictionaries for the mentioned language pairs and deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary. In addition, we will make all of the project's products (corpora, models, dictionaries) freely available supporting further research.

1 Introduction

In his survey on language death, Kornai [1] states that language has became a function that is performed digitally, and that a language is digitally viable only to the extent it produces new, publicly available digital material. Language death implies loss of function, entailing the loss of prestige, and ultimately the loss of competence. To avoid such deterioration, our project aims to support Finno-Ugric language communities so

that they would be able to cope with some of the digitally performed functions of their native languages.

In this context, language technology aspires to become an enabler technology that helps people collaborate, conduct business, share knowledge and participate in social debate regardless of language barriers and computer skills [2]. However, cutting-edge technologies are typically available only for widely-spoken languages which are in the class of digitally thriving languages according to Kornai's classification [1].

In this paper, we introduce an ongoing project whose objective is to provide linguistically based support for several small Finno-Ugric digital communities in generating online content and help revitalize the digital functions of some endangered Finno-Ugric languages. The project is based on comparable corpora collected from the web. We generate proto-dictionaries for several endangered Finno-Ugric and major language pairs and deploy the enriched lexical material on the web in the framework of the collaborative dictionary project Wiktionary.

The first major component of the research project is the compilation and development of parallel and comparable corpora. We collect text material for the following Finno-Ugric (FU) languages: Komi-Zyrian and Permyak, Meadow and Hill Mari, Udmurt and Northern Sami, as well as for major languages that are of interest to the FU community: English, Russian, Finnish and Hungarian, see Section 3.1.

The parallel and comparable corpora will be automatically pre-processed. Since these small FU languages have weak language technology support, we experiment with language-independent tools applying machine learning methods, see Section 3.2.

Having the data pre-processed, we conduct experiments with automatic dictionary generation both from parallel and comparable texts. As a result, we have bilingual proto-dictionaries containing more than one thousand translation candidates for each language pair, see Section 4.

Dictionary entries will be automatically enriched with linguistic information and manually corrected by native speakers then uploaded to Wiktionary, see Section 5.

2 Related work

Bilingual dictionaries play a critical role not only in machine translation [3] and cross-language information retrieval [4], but also in other NLP applications, like language learning [5], computational semantics and several tasks requiring reliable lexical semantic information [6]. Since manual dictionary building is time-consuming and takes a significant amount of skilled work, it is not affordable in the case of lesser used languages. However, completely automatic generation of clean bilingual resources is not possible according to the state of the art. As a middle course, rough equivalence

at the conceptual level is already a useful notion, and filtering out candidate translation pairs produced by standard bilingual dictionary building methods can support lexicographic work.

The standard dictionary building methods are based on parallel corpora. However, as foreseen by Rapp [7], "the availability of a large enough parallel corpus in a specific field and for a given pair of languages will always be the exception, not the rule", such corpora are still available only for the best-resourced language pairs. This is the reason of the increased interest in compiling comparable (non-parallel) corpora.

The standard approach of bilingual lexicon extraction from comparable corpora is based on context similarity methods (e.g. [7, 8]), which consist of the following steps: building context vectors, translation of context vectors, and comparison of source and target vectors. These methods need 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. Since there are no sufficiently large corpora and lexicons for these FU languages, conducting experiments with alternative methods is needed. There are several newer approaches to extracting translation pairs from non-parallel corpora, e.g. independent component analysis [9], label propagation [10], and topic model based methods [11]. One of the hot topics in NLP is using deep learning algorithms for obtaining vector representations for words, which are applied for a wide range of NLP tasks, as well as for extracting translation candidates from large amounts of unstructured text data (e.g. [12]). Yet another method for lexicon building is extracting the real parallel sentences from comparable corpora (e.g. [13]), which are then used as standard parallel texts for generating proto-dictionaries.

3 Creating parallel and comparable corpora

As a first step, we collected parallel and comparable texts for the language pairs in question. Linguistic processing of the collected data is inevitable before the next steps of dictionary building. Since all lexicon building methods require sentence-level aligned text, the running text must be split into sentences. Dictionary entries usually are words, thus word-level pre-processing, i.e. tokenization is also needed. Providing part-of-speech tag and lemma for each token is a very important step, since Wiktionary entries cannot be a finite word form without a lemma. In this section, we describe the subtasks of the corpus building workflow, including text collection and text processing steps.

3.1 Text collection

To build parallel corpora, we collected source texts and translations in parallel. In a strict sense, only Bible translations, novel translations, software documentation, and official documents, such as the Universal Declaration of Human Rights, can be treated as real parallel texts. For building comparable corpora, multilingual text collections have been created by applying several approaches.

Parallel corpora. Using the Bible as a parallel text in dictionary building has a long tradition [14]. To the extent feasible we tried to use modern Bible translations to avoid extracting archaic or extinct words. We downloaded the New Testament in the investigated FU and major languages from the Parallel Bible Corpus [15], Bible.is and The Unbound Bible. The translations are provided in a verse-aligned plain text format, thus they can be easily used for further processing.

Additionally, we found Northern Sami software documentation aligned with all major languages in question in the OPUS corpus [16] and some parallel texts on the websites of officially bilingual regions of Norway, Finland, and Russia. We did not find any parallel texts, not even Bible translations, for the language pairs where L1 is Komi-Permyak, Hill Mari or Udmurt.

Comparable corpora. For creating comparable corpora, the most commonly used source is Wikipedia. First of all, we downloaded the Wikipedia dumps for the languages we are dealing with and extracted each interlanguage-linked article pair. We used a slightly modified version of Wikipedia Extractor[1] for extracting the plain text, some metadata and in-text inter-wiki links with Wikidata IDs. Wikidata is a sister project of Wikipedia: it is a free collaborative multilingual knowledge base where inter-linked Wikipedia titles are instances of one and the same entity with one Wikidata ID. Using these IDs will help us to find and anchor named entities in the text of articles – regardless of language.

The length of the corresponding Wikipedia articles can be highly different: articles that are being maintained by a large, active digital community are typically fully-fledged, whereas articles from the language domain of a small community can be very brief. To improve the comparability measure, we consider only the first x sentences of each article in the major languages, where x is the number of sentences in the corresponding FU article, supposing that they are roughly each other's translation (corresponding to the first, defining paragraph).

Another approach to building a comparable corpus is downloading domain-specific monolingual texts by specifying a keyword [8]. We collected documents about the Sami culture, education and society in English and Northern Sami from several websites.

[1] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

Yet another way of collecting comparable text material is downloading multilingual daily newspaper materials from the same time interval and, if feasible, from the same country or region. Since these articles are about the most important local and global events, even if they are not translations of each other, they can be treated as comparable corpora [8]. Based on this hyphothesis, we collected articles from online newspapers in Finland for Northern Sami–{Finnish, English, Russian} language pairs.

Monolingual texts. We also collected monolingual texts for all FU languages. While parallel and comparable corpora are used to create dictionaries, monolingual texts serve as training data for the tokenizer and sentence splitter. We collected monolingual data from several websites in various domains, e.g. literature, news, personal blogs, official texts.

Table 1 shows the number of tokens for parallel, comparable and monolingual corpora. Numbers for the latter comprise all data, even the texts included in parallel and comparable corpora. Token numbers for comparable corpora contain only the restricted size Wikipedia articles, i.e. only the first x sentences of each article pair (see above). The time interval-based texts have been compared on a yearly basis, i.e. news files which do not have a corresponding pair from the same time period were not considered. The numbers in the table show the current state of the text processing task; they will indeed change at later phases of the project.

3.2 Text processing

The pre-processing and segmentation is a particularly important step, largely because any error made at this stage is likely to cause complications at later phases of text processing. However, these small FU languages have weak language technology support. An outstanding contributor is Giellatekno[2], which provides NLP tools for several FU languages. Northern Sami by far has the largest available online resources, but, as far as we know, tools for the other small FU languages are still under development.

Pre-processing steps. As mentioned in Section 3.1, we gathered a relatively large amount of text, but there are a number of issues with the collected material, since an easily observable portion can be classified as “dirty text”.

First, all texts must undergo character normalization before any further processing steps. All sources have been converted to plain text files using standard Unicode characters in UTF-8 encoding.

Second, closely related languages (Komi-Zyrian and Permyak, Meadow and Hill Mari) are often mixed even within a single document, thus they have to be separated. For language discrimination, we used the Blacklist Classifier[3], which resulted in a

[2]http://giellatekno.uit.no

[3]https://bitbucket.org/tiedemann/blacklist-classifier/wiki/Home

lang	mono	lang pairs	parallel		comparable	
			L1	L2	L1	L2
sme	1,364,254	sme–eng	691,260	724,750	253,930	1,754,968
		sme–fin	245,440	273,973	239,651	5,259,591
		sme–rus	173,179	220,790	212,332	233,748
		sme–hun	171,668	224,014	86,244	106,391
kpv	480,609	kpv–eng	121,108	174,742	89,580	183,602
		kpv–fin	121,120	133,715	88,507	80,797
		kpv–rus	117,903	125,085	108,013	141,369
		kpv–hun	121,319	134,344	68,179	74,274
koi	719,325	koi–eng	0	0	257,871	194,784
		koi–fin	0	0	137,578	77,696
		koi–rus	0	0	188,334	139,976
		koi–hun	0	0	95,120	64,794
mhr	1,335,457	mhr–eng	128,316	175,075	121,588	250,583
		mhr–fin	128,328	133,965	118,120	115,028
		mhr–rus	109,449	109,818	158,977	215,724
		mhr–hun	128,565	134,618	106,813	121,453
mrj	366,964	mrj–eng	0	0	137,088	306,465
		mrj–fin	0	0	85,134	93,622
		mrj–rus	0	0	124,289	187,687
		mrj–hun	0	0	77,855	90,168
udm	584,113	udm–eng	0	0	67,306	135,450
		udm–fin	0	0	56,222	49,961
		udm–rus	0	0	80,800	129,293
		udm–hun	0	0	41,883	48,736

Table 1: Number of tokens for monolingual, parallel and comparable corpora. We use the ISO 639-3 language codes: sme – Northern Sami, kpv – Komi-Zyrian, koi – Komi-Permyak, mhr – Meadow Mari, mrj – Hill Mari, udm – Udmurt.

97.47% accuracy for Komi-Zyrian and Permyak and a 96.77% accuracy for Meadow and Hill Mari languages.

Third, majority of FU language bloggers use blog publishing services that only support Russian or English, therefore these texts are mixed, thus dates and some elements of websites are not in the desired FU language but in one of the languages supported by the service provider. To filter out foreign parts, we use Langid[4], a language identifier using trigram statistics with Katz's back-off smoothing. Models were created using manually selected text samples. Since dates are valuable information in order to build time frame-based comparable corpora, we preserved them.

Sentence segmentation and tokenization. For sentence segmentation and tokenization, we use the sentence detection and tokenizer tools of Apache OpenNLP[5]. Since Northern Sami is quite well-supported with NLP tools, we built models only for the FU languages using Cyrillic script. We created gold standard data for training and testing the tokenizer and sentence splitter moduls of Apache OpenNLP. Over ten thousand sentences were randomly selected for each language, and, after manual correction, the data was divided into training and test sets (90%-10%). Both moduls performed over 98% F-measure, which is partly due to the abbreviation dictionary support of Apache OpenNLP, blocking the false sentence segmentation at abbreviations. Using such lists is a common practice in sentence segmentation, however, building an exhaustive list is beyond the bounds of possibility, especially for the FU languages in the scope of our research. For this reason, our abbreviation dictionary is mainly based on the Russian abbreviation list of Wiktionary, but we plan to extend it with more abbreviations and acronyms found at later stages of our work. Nevertheless, using only Russian abbreviations would be sufficient for our needs as the FU languages written in Cyrillic script tend to use a number of Russian abbreviations (consider the abbreviations for units of measurement, place names and internationalisms).

Morphological analysis and disambiguation. Morphological analysers are available as online applications for Udmurt and Komi-Zyrian[6] and Hill Mari[7]. The website of Giellatekno contains source files for a finite state transducer-based morphological analyser for almost all FU languages we deal with. However, as far as we know, morphological analyser for Komi-Permyak do not exist.

For languages that lack any morphological analysers, there are two possibilities we can choose from. Once, we can use semi-supervised or unsupervised morphological segmentation tools such as Morfessor[8] and develop additional tools to extend

[4]https://github.com/juditacs/langid

[5]https://opennlp.apache.org/

[6]http://www.morphologic.hu/urali/index.php?lang=english

[7]http://www.univie.ac.at/maridict/site-2014/morph.php

[8]http://morfessor.readthedocs.org/en/latest/general.html#techrep

its functionality to meet our needs. We made some experiments with Morfessor: we trained it on an Udmurt word list and compared its segmented output to the output of an Udmurt morphological analyser [17]. The results are convincing, but it would still take great effort to develop additional utilities that could produce lemmas and POS tags from Morfessor's output.

The other option is to use existing tools developed for closely related languages. The most simple solution is the direct application of tools developed for the related language, thus the models built for Komi-Zyrian could be applied on the Komi-Permyak data directly. Moreover, we expect that morphological tags in Komi-Zyrian can be transferred to the Komi-Permyak version of the same text. Since large amounts of data for training do not exist for the majority of languages, experimenting with several methods of the annotation transfer between closely related languages is a hot topic in NLP (e.g. [18, 19]). We plan to investigate the approach for transferring POS annotations from a resourced language towards a closely related non-resourced language by Scherrer and Sagot [18].

4 Creating proto-dictionaries

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 work. Proto-dictionaries are expected to provide greater coverage but comprise more incorrect translation candidates; their right size depends on the specific needs.

We made experiments with several lexicon building methods, which are detailed below. Applying each method resulted in bilingual resources containing translation candidates for almost all language pairs. These dictionary files will then be used as the starting point to create the final dictionaries, where only the most likely translation candidates are kept on the basis of some heuristics, developed in a later phase of the project by manually evaluating the results. At this stage of the project, we have raw, i.e. still un-cleaned proto-dictionaries for all language pairs each containing more than one thousand translation candidates. The most under-resourced language pair is Komi-Permyak–Hungarian with ca. 1300 word pairs, while the Northern Sami–Finnish proto-dictionaries contain more than 20,000 word pairs.

Wikipedia titles. 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. Erdmann et al. [20] 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 [13] extracted parallel sentences from the

English and Persian Wikipedia using a bilingual dictionary generated from Wikipedia titles as a seed lexicon. Following this approach, we also created bilingual dictionaries from Wikipedia title pairs using the interwiki links.

Wiktionary-based methods. Besides Wikipedia, Wiktionary is also considered as a crowdsourced 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. Ács et al. [21] extracted translations from the so-called translation tables. Since their tool Wikt2dict is freely available[9], 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 [22] 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. With the triangulation mode of Wikt2dict, we could further expand our dictionaries.

Hundict. Hundict[10] is an experimental project for bilingual lexicon extraction from parallel corpora. It extracts word pairs based on high co-occurence in corresponding text segments, using the Sørensen-Dice co-efficient. The tool's performance can be improved by adding a gold standard dictionary and a list of stopwords. We made some experiments with Bible translations for Northern Sami–Finnish and Komi-Zyrian–English language pairs, which resulted in word pairs along with their confidence measures. The system has more parameters which can be fine-tuned, and we plan to test it with more options and on more language pairs. However, the tool needs lemmatized text as an input, thus we have to lemmatize our parallel corpora before further experiments.

5 Conclusion and future work

We introduced an ongoing project which is based on parallel and comparable corpora collected from the web. The project's main objective is to generate dictionaries for language pairs where the source language is one of the following small FU languages: Komi-Zyrian and Permyak, Meadow and Hill Mari, Udmurt and Northern Sami, while the target language is one of the following major languages: English, Finnish, Russian and Hungarian. However, collecting text for these under-resourced FU languages and processing them poses several problems. Most of these FU languages are digitally not

[9] https://github.com/juditacs/wikt2dict

[10] https://github.com/zseder/hundict

really viable, since they produce very few digital text material. For this reason, finding large amounts of text in these languages is challenging. Since there are language pairs for which we did not find any parallel text material, standard dictionary building methods cannot be used. Moreover, these small FU languages have weak or no language technology support, thus language-independent supervised tools are needed to be used. On the level of morphology we faced other kind of problems: some analysers are available only as an online application, some are still under development.

In spite of the difficulties, we collected some text material for these languages and built proto-dictionaries by applying several methods. The final dictionaries will be uploaded to Wiktionary, where lexical entries contain morphological, etymological and lexico-semantic information, and translation equivalents across languages. We will generate the dictionaries and the linguistic information as automatically as possible. The manual validation and correction of Wiktionary input files will be conducted only in the last phase of the project by native speakers.

Using the Wiktionary infrastructure, lexical entries across the language versions of Wiktionary can be interlinked. This will enable user communities to access rich, networked lexical material that can be used for translation purposes. Content in Wiktionary is formatted in a lightweight markup system, but we will recast the data into an XML format suitable for further processing. After cleaning the copyright issues, we will make all of the generated resources (corpora, dictionaries, models) freely available.

Acknowledgments

The research reported in the paper was conducted with the support of the Hungarian Scientific Research Fund (OTKA) grant #107885. We thank the anonymous reviewers for their constructive comments.

References

[1] A. Kornai. Digital Language Death. *PLoS ONE*, 8(10), 2013.

[2] E. Simon, P. Lendvai, G. Németh, G. Olaszy, and K. Vicsi. *A magyar nyelv a digitális korban – The Hungarian Language in the Digital Age*. Georg Rehm and Hans Uszkoreit (Series Editors): META-NET White Paper Series. Springer, 2012.

[3] R. D. Brown. Automated dictionary extraction for "knowledge-free" example-based translation. In *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 111–118, 1997.

[4] G. Grefenstette. The Problem of Cross-Language Information Retrieval. In G. Grefenstette, editor, *Cross-Language Information Retrieval*, page 1–9. Kluwer Academic Publishers, 1998.

[5] A. Kilgarriff, F. Charalabopoulou, M. Gavrilidou, J. B. Johannessen, S. Khalil, S. Johansson Kokkinakis, R. Lew, S. Sharoff, R. Vadlapudi, and E. Volodina. Corpus-Based Vocabulary lists for Language Learners for Nine Languages. *Language Resources and Evaluation*, 48(1):121–163, 2013.

[6] T. Zesch, C. Müller, and I. Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[7] R. Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting of Association for Computational Linguistics*, ACL '95, page 320–322, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

[8] P. Fung and L. Y. Yee. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 17th International Conference on Computational Linguistics – Volume 1*, COLING '98, page 414–420, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[9] A. Hazem and E. Morin. ICA for Bilingual Lexicon Extraction from Comparable Corpora. In *The 5th Workshop on Building and Using Comparable Corpora*, pages 126–133, Istanbul, Turkey, May 2012.

[10] A. Tamura, T. Watanabe, and E. Sumita. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, page 24–36, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[11] I. Vulić and M. F. Moens. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th*

Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, page 449–459, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[12] R. Al-Rfou, B. Perozzi, and S. Skiena. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[13] M. Mohammadi and N. Ghasem-Aghaee. Building Bilingual Parallel Corpora Based on Wikipedia. In *2010 Second International Conference on Computer Engineering and Applications (ICCEA)*, volume 2, pages 264–268, March 2010.

[14] P. Resnik, M. B. Olsen, and M. Diab. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33(1–2):129–153, 1999.

[15] T. Mayer and M. Cysouw. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[16] J. Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[17] A. Novák. Morphological Tools for Six Small Uralic Languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC '06)*. European Language Resources Association (ELRA), 2006.

[18] Y. Scherrer and B. Sagot. A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 502–508, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[19] A. K. Ingason, H. Loftsson, E. Rögnvaldsson, E. F. Sigurdsson, and J. C. Wallenberg. Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 91–95, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[20] M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. An Approach for Extracting Bilingual Terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 5(4):31:1–31:17, November 2009.

[21] J. Ács, K. Pajkossy, and A. Kornai. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[22] J. Ács. Pivot-based multilingual dictionary building using Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

Freeware dictionary program to download (Windows or Android) with many wordlists to browse offline or look up online. We also provide free human translation and professional translation.Â Freelang dictionary is a free program to download for Windows. It is very easy to install : pick up a language in the list below, then download and install the program and the wordlist. Once the program has been installed, you can download and install as many wordlists as you want. The program enables you to browse the lists, look up a word, add or modify an entry, and learn words at your own rhythm from a personal learning list. All wordlists are bilingual (from/to English).Â Finno-Ugric group. Finnish branch. ESTONIAN - Estonia. For the creation of the proto-dictionaries, we applied sev-eral lexicon building methods utilizing Wikipedia and Wik-tionary. For more details on the dictionary creating methods we used, see Benyeda et al. (2016) and Simon and Mittel-holcz (2017) â€“ 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 bilin-gual lexicon extraction in several ways. For example, Erd-mann et al.Â (2015). Automatic creation of bilingual dictionaries for. Finno-Ugric languages. In 1st International Workshop on Computational Linguistics for Uralic Languages, TromsÃ¸. Tiedemann, J. (2009). Simon, Eszter and Benyeda, Ivett Zsuzsanna and Koczka, PÃ©ter and LudÃ¡nyi, ZsÃ³fia (2015) Automatic creation of bilingual dictionaries for Finno-Ugric languages. In: Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages. University of Tromso, Tromso, pp. 119-131. Preview. Text 3474_13193_1_PB_u.pdf Download (134kB) | Preview. The Finno-Ugric languages form a subfamily of the Uralic languages. The majority of linguists believe that Finnish, Hungarian and Estonian, among other languages, should be included in the group. Unlike most of the other languages spoken in Europe, the Finno-Ugric languages are not part of the Indo-European family of languages. The Uralic languages also include the Samoyedic languages, and some linguists use the terms Finno-Ugric and Uralic as synonyms.Â The "Urheimat" of Proto-Finno-Ugric, the proto-language of the modern Finno-Ugric languages, cannot be located with any certainty. The area west of the Ural mountains is generally assumed as a likely candidate, at a time of maybe the 3rd millennium BC.