Compiling an Estonian-Slovak Dictionary with English as a Binder

Michaela Denisová[1]

[1] Masaryk University, Žerotínovo nám. 617/9, 601 77 Brno
E-mail: michaeladenisova@gmail.com

Abstract

For such a rare language combination as Estonian-Slovak, it is complicated to find study materials designated for Slovaks learning Estonian, especially a bilingual dictionary, an essential language study resource. However, building a bilingual dictionary from scratch requires a lot of work and effort. The half-automatic computational methods and available open-source language resources offer a possible solution for this complicated task. One approach is to merge two already existing dictionaries that share a common language to derive a new language pair dictionary. However, as words are polysemous, many mistakes could occur while attempting so. Therefore, it is required to edit the aligned translations afterwards.
This article describes the process of compiling the Estonian-Slovak dictionary created from English-Estonian and English-Slovak dictionaries. English was chosen as an intermediate language, as it is a well-resourced language, and all materials are easy to find. Various automatic techniques were applied in the editing step to decrease the number of incorrectly aligned translations. Finally, the techniques used and quality of the dictionary were manually evaluated on a random sample of 1,000 translations.
The final version of the dictionary consists of 138,779 translations, and the Estonian headword list covers about 85% of basic Estonian vocabulary, which contains around 5,000 lemmas. The correct translations form approximately 40% of the dictionary. Additionally, a web application is being developed for this dictionary.[1]

Keywords: bilingual dictionaries; (semi)automatic compilation; intermediate language; Estonian; Slovak

1. Introduction

This project was created as a master's thesis to provide more learning materials for Slovak students who learn Estonian at the department of Baltic Studies at Masaryk University. Students struggle with a lack of study resources in their mother tongue, especially at the beginning, and it is challenging to find an accurate translation. Therefore, this project assumed that a dictionary is one of the most crucial study materials when acquiring a new language, as students look up a foreign word in the dictionary to understand its meaning and use it correctly. Still, it is difficult to translate directly into the learners' mother tongue for a low-resource language pair, such as Estonian-Slovak. Many students thus use another major language as an intermediary, which provides more materials. However, this could lead them to incorrect translations

[1] https://estonian-slovak-dictionary.herokuapp.com (23 March 2021).

and cause mistakes in language usage. Unfortunately, creating a new bilingual dictionary is a non-trivial task, especially for rare language combinations. Using automatic methods and available open-source language resources could solve this problem. One option is to derive a new bilingual dictionary from existing dictionaries with well-resourced language as their common language. In this project, English-Estonian and English-Slovak open-source dictionaries were merged to create a new Estonian-Slovak dictionary. The direction from Estonian to Slovak was preferred as it may be more critical for the learners to grasp the meaning of the foreign words at first.

As words are polysemic, the incorrect translations are also aligned by this method. For example, *piť* (verb) ~ *drink* (verb, noun) ~ *jook* (noun). For this problem, several solutions were introduced, e.g. inverse consultation (Tanaka & Umemura, 1994) or inverse consultation combined with distributional similarity (Saralegi et al., 2011). The inverse consultation method was based on extracting translations via intermediate language and repeating the same process with the obtained words in the reverse direction. Meanings included in the acquired intersection were considered correct. In the latter method, the distributional similarity was computed from the custom-built parallel corpora to retrieve the distances between the translations.

The above mentioned approaches have been used and improved over the years in various projects, for instance, in a Japanese-French dictionary (Tanaka & Umemura, 1994), Korea-Japanese dictionary (Shirai & Yamamoto, 2001), Basque-Chinese dictionary (Saralegi et al., 2011), or in a project dealing with automatic generation of several bilingual dictionaries (Ordan et al., 2017).

The difference is that those approaches focused more on the process of combining the dictionaries. In this project, the merging is not as crucial as the automatic correction of the aligned translations after merging. Besides, the techniques applied and the quality of the resulting dictionary were manually evaluated to provide precise and accurate results for each technique separately and the whole dictionary.

This article is structured as follows. The second section explains the nature of the chosen dictionaries, and the following one focuses on the compilation process. The fourth section deals with techniques applied for automatic correction of the incorrectly aligned translations after the process of merging. Techniques are divided into separate subsections: alignment based on the part of speech, comparison with WordNet and Google Translation datasets, comparison of Estonian headword list with the EKI Combined Dictionary, and comparison of named entities. After that, the results of the evaluation are given. Finally, the conclusion is drawn, and new ideas are outlined.

2. Dictionaries

For this project, three types of open-source dictionaries were used, one English-Estonian

dictionary[2] obtained from the Estonian Language Institute, and two English-Slovak dictionaries, one from online dictionary platform dict.cc[3] where authors can contribute and share their dictionaries and the other one from DictionaryForMIDs[4], which is a free multi-purpose dictionary designed for cell phones, PDAs, or PCs.

The English-Estonian dictionary was a large dictionary with an extensive vocabulary, which contained 83,089 headwords. On the other hand, the English-Slovak dictionary obtained from DictionaryForMIDs had only 26,070 English headwords. Therefore, it was necessary to include the second English-Slovak dictionary to create a greater intersection with English words from the English-Estonian dictionary, so the resulting compiled Estonian-Slovak dictionary would contain more entries. The English-Slovak dictionary from dict.cc had 25,025 English headwords that belonged to the general vocabulary and words from specific fields, such as anatomy or biology. It contained explanatory notes and abbreviations as well, but those were eliminated in the text pre-processing phase.

As a result, these two English-Slovak dictionaries together had 41,516 English headwords. They overlapped in less than 10,000 headwords.

3. Merging Dictionaries

The first step required to merge English-Estonian and English-Slovak dictionaries was to extract their common English word list. According to this list, every Estonian translation was aligned with every Slovak translation, which caused an exponential increase of the translations, and it aligned together words with different meanings (see Figure 1). This merging created two dictionary directions: Estonian to Slovak and Slovak to Estonian. As mentioned above, in the next steps, only the Estonian-Slovak direction was processed.

The first version of the Estonian-Slovak dictionary contained 34,674 Estonian headwords.

It was necessary to perform manual control on a random sample of 1,000 translations to analyse the main mistakes after merging, so the solutions could be adjusted accordingly. It was also essential to find out the proportion of the correct and incorrect translations. According to the control performed, only around 25% were correctly aligned translations. The remaining translations consisted of mistakes.

[2] https://www.eki.ee/litsents/ (21 March 2021).

[3] https://www.dict.cc/ (21 March 2021).

[4] https://sourceforge.net/projects/dictionarymid/ (21 March 2021).

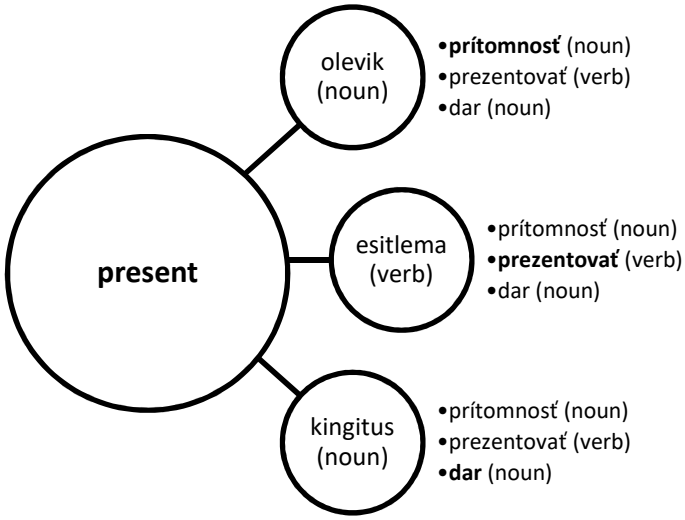

Figure 1: Aligning the Estonian words *olevik*, *esitlema, and kingitus* with the Slovak words *prítomnosť*, *prezentovať* and *dar* according to the English word 'present'

As it was assumed, most of the mistakes were caused by the ambiguity of the words, for instance, the diversity of the parts of speech typical for the English words (present (noun) vs to present (verb)) (see Figure 1). Moreover, it included the words describing nationalities, language groups and countries as in the word 'Italian', which could be in Estonian *itaallane* (nationality) or *itaalia keel* (Italian language). These types of mistakes occurred in 75% of all incorrectly aligned translations in the control group.

Other translations, around 7%, consisted of misspelled words, rarely used words, non-lemmas, proper nouns, or foreign words from other languages.

The analysis of Estonian headword lists revealed that there were incorrect headword candidates. 14% of them were multiword expressions. Multiword expressions were considered as word sequences with some unpredictable properties (Parmentier & Waszczszuk, 2019). They were manually detected during the control, and this group included mainly expressions untypical for learners' dictionaries, e.g. in Estonian *graafiliselt esitama* 'to present graphically' or *tuhast puhastama* 'to clean from the ash'.

2.5% of the Estonian headwords contained a hyphen, usually prefixes such as *eba-* 'un' or *silbi-* 'syllabic'. These headwords were easily removable, but the question was whether and which of these headwords are relevant for the learners, and thus worth keeping in the dictionary.

There are several possible explanations for why those incorrect headword candidates appeared in the headword list. The first is that English headwords' descriptive translations became a headword in Estonian and Slovak while merging the dictionaries. Other possibilities are that mistakes were made during the text pre-processing, or potential mistakes were already in the original dictionaries.

In the next section, the techniques applied for solving the errors mentioned above are stated.

4. Applied Techniques

This section describes the techniques which were applied after merging the dictionaries. All those techniques were adjusted to the mistakes revealed during the manual control described in the third section. They were chosen to efficiently eliminate as many incorrectly aligned translations or incorrect Estonian headword candidates as possible while maintaining the correct translations.

4.1 Alignment based on the part of speech

The first technique assumed that translations have the same part of speech in both languages. This means that the Slovak translation of an Estonian noun is a noun as well, and all word pairs with a different part of speech are incorrectly aligned together. This solution addressed the problem with the word classes' diversity of the English words (see Figure 1).

EstNLTK library version 1.4.1.[5] was used for annotating Estonian headwords, and the web application Slovak POS Tagger[6] developed by the Slovak University of Technology in Bratislava was chosen for Slovak translations.

The main problem while using this technique was with the accuracy of the libraries. Tagging libraries give more accurate results when the context of the word is available. However, there were no contextual words in this case, and the morphological analysis proposed only one part of the speech tag that could but did not have to be the correct one. For instance, Estonian verbs in a past passive participle form (e.g., *teatud* 'done') are translated into Slovak as adjectives, but the EstNLTK tagger marked them in different cases, either as verbs or adjectives, e.g. *tagatud* 'guaranteed' as a verb, *maetud* 'buried' as an adjective.

Another problem was that the Slovak tagger did not recognise around 13% of all Slovak translations and marked them as unknown, which reduced the number of aligned translations to compare. Between those words were rarely used words, inflected word forms or multiword expressions containing spelling errors. However, any multiword expressions could not be included in the part of speech comparison because they received a POS tag according to the first word in the expression. Although usually, the last word determines the part of speech of the whole expression.

---

[5] https://github.com/estnltk/estnltk (21 March 2021).

[6] http://morpholyzer.fiit.stuba.sk:8080/PosTagger/. The accuracy when choosing a single tag is 65%. (21 March 2021).

While comparing tags between aligned translations, only nouns, verbs, adjectives, numerals, and pronouns were considered, since the other word classes groups were more likely to contain mistakes and incorrect differences between the tags given by the libraries. Moreover, only Estonian headwords with more than one Slovak translation were included. This measure was taken because if the dictionary contained Estonian headwords aligned with a single Slovak translation with a different part of speech, the automatic comparison would remove it from the dictionary. This would result in the loss of the headwords while the objective was not only to remove aligned translations, but also to maintain the vocabulary.

As a result of this technique, around 25% of all aligned translations were removed from the dictionary (in contrast to the number of aligned translations occurring in the dataset before the part of speech comparison), which was a satisfactory result.

4.2 Comparison with WordNet and Google Translation datasets

The second technique that was applied was the extraction of new bilingual datasets and comparison of the results across them. One of the available language resources for both languages was WordNet.[7] WordNet is a network that connects words according to their semantic relationships, while every word carries its own index. According to this index, words can be looked up in wordnets in different languages. Although there are limitations of WordNet (Pedersen & Braasch, 2009), in this project it was considered a trustworthy language resource since it was made manually by lexicographers (compared to half-automatically derived resources).

Estonian WordNet[8] and Slovak WordNet[9] were used for these purposes. Words with the same index, which indicated an equivalent synonym, were matched together and thus created a new Estonian-Slovak dictionary with 6,829 translations. In comparison to the original Estonian-Slovak dictionary, only 1,254 translations occurred in both dictionaries. It was a very small number, as in the first extracted dictionary were 156,180 translations.[10]

WordNet can be used in several different ways. One option is to measure the distances between the words computed via the Open Multilingual WordNet module in the NLTK library [11] or EstNLTK. However, the similarities measurement works within one language, not across the languages. Additionally, the intersection between Estonian and

[7] https://wordnet.princeton.edu/ (7 April 2021).

[8] https://www.cl.ut.ee/ressursid/teksaurus/index.php?lang=et (21 March 2021).

[9] https://korpus.sk/WordNet.html (21 March 2021).

[10] After splitting the translation into the format – one Estonian headword with one Slovak translation per row.

[11] https://www.nltk.org/ (28 March 2021).

Slovak WordNets was trifling in terms of receiving reasonable results.

Another option for this technique to work was to extract another dictionary so the results could be more accurate. Using Google Translate API[12] appeared as a good option. All Estonian headwords from the original dictionary and the WordNet dictionary were extracted and translated via the Google Translate API library into the Slovak language. The result was a third Estonian-Slovak dictionary with 45,178 translations.

The Google API derived dataset was manually checked on a control group consisted of around 700 randomly chosen translations. The reason was to assess the quality for the next steps. Different types of mistakes were revealed; for example, headwords translated using the same word (*tõtlikult – tõtlikult*)[13] or headwords translated into languages other than Slovak (*ebaloomulikkus – unnaturalness*). Additionally, errors caused by polysemy appeared. For instance, the Estonian word *sepikoda* 'forge shop' was translated into Slovak as *falšovať* 'to fake', where both words came from the English word 'forge' containing both meanings. The percentage of correct translations in the control group was around 55%, slightly more accurate than the original Estonian-Slovak dictionary.

These three datasets were sufficient to make comparisons. The idea was to give a score to every translation according to its occurrences in the datasets. If the translation occurred only in the Google dataset or only in the original one, it received a score of 0.25. The score for the WordNet dataset was the highest - 0.5. Thus, if the translation was in all three datasets, it got a score of 1. If found in the WordNet and Google datasets it obtained a score of 0.75, etc. The logic behind this was that WordNet is the most trustworthy resource since it was compiled manually, whereas the other datasets were automatically derived.

The success rate of this technique depends on how many resources are available to compare. Naturally, most of the translations received a score of 0.25, and the smallest group consisted of translations with a score of 1 (see Table 1). On the other hand, this technique could serve as an indicator for users as to what extent they can rely on a current translation. Moreover, the score indicates which group of translations should be corrected when manually post-editing. Additionally, each score group was manually checked on a random sample of 500 translations, and the results are described in more detail in Section 5.

[12] https://cloud.google.com/translate/docs/ (21 March 2021).

[13] Those were easily removed.

The number of word pairs	Score
178,678	0.25
15,194	0.5
1314	0.75
502	1

Table 1: Comparison with WordNet and Google Translation datasets.

4.3 Comparison of headword list in EKI Combined Dictionary

This technique aimed to eliminate words that are not usually given as a headword in general-purpose dictionaries, e.g. proper names, inflected word forms, misspelled words, abbreviations, or foreign words from other languages. The EKI Combined Dictionary[14] and its user interface Sõnaveeb (Tavast et al., 2019) were used for these purposes. The EKI Combined Dictionary contains rich linguistic information about Estonian words. This technique's objective was to look up every Estonian headword automatically. Suppose the headword could not be found in the EKI Combined Dictionary; in that case, it could be eliminated from the dictionary since this technique assumed that the EKI Combined Dictionary contains all relevant words for users or learners.

As a result, 10,014 Estonian headwords were not found in the EKI Combined Dictionary. Manual control was performed on a random sample of 1,000 translations. The biggest group consisted of multiword expressions, around 72%. The rest of them made up foreign words from other languages, e.g. 'capriccio' or 'curling', proper names or rarely used words.

The problem with the inflected word forms persisted, as automatic searching through the EKI Combined Dictionary allows to look up a lemma of an inflected word form. The words found in the EKI Combined Dictionary also contained words with a comparison score of 0.75 or even 1 (see Section 4.2.). This was exactly 123 words (e.g., *varastaja* 'thief', *aadlinaine* 'noblewoman'), so the decision was to keep such words in the dictionary as headwords.

4.4 Comparison of named entities

This last technique focused on the polysemy problem between words referring to

[14] https://metashare.ut.ee/repository/browse/the-eki-combined-dictionary-2021/af363d08857111eba6e4fa163e9d4547c858d4634fcb44eea7b56db3e452675c/ (21 March 2021).

nationalities, countries, or language groups. For instance, in English, the word 'Italian' refers to the nationality (Italian men or Italian woman) or language. This problem could be solved by using a named entity recognition library which decides about every name if it is either person (PER), location (LOC) or organisation (ORG), e.g. the Estonian word *Itaallane* – PER (Italian men). The Estonian headword tag could then be compared to its Slovak translations' tags; when the tags differ, it is an incorrectly aligned translation.

Estonian headwords were tagged by the libraries EstNLTK and Polyglot.[15] Polyglot was also used for classifying Slovak translations.

This technique was the most unsuccessful because libraries gave different results and marked the same words with different tags. For example, some countries were classified as a location, while others as an organisation, even in some cases when the translation was correct (see Table 2). An interesting choice was for the word *European Union*, which received in Slovak language organisation tag and a location tag in Estonian. Differences occurred between the libraries used for the same language. For example, Polyglot tagged the Estonian word *Mars* as a person while EstNLTK as a location.

Another problem was that the EstNLTK library did not tag nationalities, and Polyglot tagged only a few exceptions, which significantly limited the group of translations compared. Polyglot classified 1,477 Estonian headwords and 928 Slovak translations, and the EstNLTK library marked 935 Estonian headwords. Due to the small number of translations that could be compared and significant differences between the given tags, the results were not considered.

Estonian headword	**EstNLTK library**	**Slovak translation**	**Polyglot library**
• *Filipiinid* (‘Philippines’)	• ORG	• *Filipíny* (‘Philippines’)	• LOC
• *Gruusia* ('Georgia')	• LOC	• *Georgia* ('Georgia')	• ORG
• *Somaali* ('Somalia')	• PER	• *Somálsky* ('Somalian’)	• LOC

Table 2: Results of the comparison of named entities

[15] https://polyglot.readthedocs.io/en/latest/ (21 March 2021).

5. Evaluation

The resulting automatically derived dictionary consisted of 138,779 translations (28,873 Estonian headwords), and it was evaluated from two points of view. Firstly, what kind of vocabulary it contained and, secondly, which of the applied techniques helped improve the dictionary's quality and what types of mistakes persisted.

The Estonian headword lists were compared to the lemma list of the Balanced Corpus of Estonian and then to the lemma lists in each sub-corpus individually from the same corpus.[16] This corpus comprises texts from newspapers, literature, and academic texts. 92% of the words with a frequency over 5,000 were included in the Estonian headword list. When looking at the various genres, including journalism, fiction, and scientific texts, the percentages of Estonian headwords included in the dictionary with occurrences over 1,000 and 5,000 were in the range from 88% up to 94%. The results are stated below in Table 3.

	Headwords with frequency over 5,000	**Headwords with frequency over 1,000**
• Whole corpus	• 92%	• 90%
• Journalism sub-corpus	• 94%	• 91%
• Fiction sub-corpus	• 90%	• 88%
• Scientific sub-corpus	• 93%	• 92%

Table 3: The percentage of Estonian headwords from the dictionary contained in different sub-corpora frequency lists

Since written language varies from the spoken language, the comparison with the wordlist extracted from the Basic Estonian Dictionary[17] provided a more accurate picture. This dictionary was compiled for learners at A2 to B1 CEFR levels and covers the basic Estonian vocabulary. Around 85% of headwords from the Basic Estonian Dictionary occurred in the Estonian-Slovak dictionary. Missing words were, for instance, zodiac signs or the word 'me'.

When assessing the headword list with regard to the part of speech representation, the biggest group was made up nouns and adjectives, around 67% and 12%, respectively. Those were followed by verbs with approximately 10%. The remaining 11% consisted

[16] https://www.cl.ut.ee/ressursid/sagedused1/index.php?lang=en (21 March 2021).

[17] http://www.eki.ee/dict/psv/ (21 March 2021).

of adverbs, pronouns, interjections, numerals, etc.

For the second evaluation, 1,000 translations were randomly chosen and manually controlled. This control revealed that around 40% of translations are correctly aligned, which is 15% more than during the first control before post-processing. The most frequent mistakes were still related to polysemy. Specifically, incorrectly aligned translations with the same or unknown part of speech and the persisting problem with nationalities, countries, and languages. The percentage of incorrect translations caused by polysemy was approximately 92% in this control group. Compared to the initial control, it is 17% more, which means that the percentage of other mistakes decreased.

Other errors that methods could not eliminate were related to multiword expressions, misspelled words and inflected word forms in Slovak translations, Estonian non-lemmas, and translations in other languages than Slovak (e.g., English, Czech etc.).

As a result, the most successful technique appeared to be alignment based on the part of speech, where the number of wrongly connected translations fell by around 25%. This percentage could be increased by using a more accurate tagger.

A comparison with WordNet and Google Translation datasets also gave good results. Each group with a different score (1, 0.75, 0.5, 0.25) was manually evaluated on a random sample of 500 translations.[18] The manual evaluation confirmed that the given score corresponds with the error percentage in the group. The results are stated in Table 4 below. Overall, the given score can be valuable for dictionary users as an appropriateness indicator or for further dictionary development.

Score	**The Percentage of Errors**
• 1	• 0.59%
• 0.75	• 4.6%
• 0.5	• 15%
• 0.25	• 66.4%

Table 4: The percentage of errors in each score group

On the other hand, the technique using named entity recognition failed because of immense differences between the results for the exact words given by different taggers.

[18] All translations from group with score 1 were checked.

An overview of all applied methods and results is provided in Table 5.

Method	**Impact**
• Alignment based on the part of speech	• 25% incorrectly aligned word pairs removed
• Comparison with WordNet and Google Translation datasets	• See Table 4
• Comparison with the EKI Combined Dictionary	• 24% of Estonian headwords were removed
• Comparison of named entities	• The method failed due to the immense differences between tags given by tagging libraries

Table 5: Applied techniques and their results summarisation.

6. Conclusion and Future Works

This article introduced the Estonian-Slovak dictionary, automatically derived from two already existing dictionaries that shared English as their common language. Merging of the dictionaries produced many incorrectly aligned translations, where most of the errors were caused by polysemy.

Several techniques were applied to reduce the number of incorrectly aligned translations: alignment based on the part of speech, comparison with WordNet and Google Translation datasets, comparison of Estonian headword list with the EKI Combined Dictionary, and comparison of named entities. The best approach turned out to be comparing the part of speech tags between the aligned translations. In contrast, tagging word pairs with named entity recognition feature failed due to the different tags.

In the end, the quality of the dictionary was evaluated. The evaluation revealed that the dictionary consists of 138,779 translations and the Estonian headword list covers 85% of the basic Estonian vocabulary. Regarding the quality of the translations, around 40% of the translations are correct, while in the remaining roughly 60% some errors persisted, mostly caused by polysemy.

There are several options to increase accuracy. Firstly, the scoring technique could be extended by another dictionary extracted from Estonian-Slovak parallel corpora. Estonian corpora could be used to control if translations contain all relevant meanings. On top of that, the web application for this Estonian-Slovak dictionary was built, and its further development is planned.

7. Acknowledgements

I express my gratitude to my supervisor, Dr. Kadri Muischnek, MA, Ph.D., and Jelena Kallas, MA, Ph.D. for their guidance and valuable advice.

8. References

Ordan, N., Gracia J. & Kernerman I. (2017). Auto-generating Bilingual Dictionaries. In I. Kosem et al. (eds.) e*Lex 2017 Proceedings*, pp. 474-484.

Pedersen, B. S. & Braasch, A. (2009). What do we need to know about humans? A view into the DanNet database. *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. Editors: Kristiina Jokinen and Eckhard Bick, pp. 158–166.

Saralegi, X., Manterola I. & San Vicente I. (2012). Building a Basque-Chinese Dictionary by Using English as Pivot. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* L12-1006, pp. 1443–1447.

Shirai, S. & Yamamoto K. (2001). Linking English Words in Two Bilingual Dictionaries to Generate Another Language Pair Dictionary. *Proc. of ICCPOL*, pp. 174-179.

Tanaka, K. & Umemura, K. (1994). Construction of a Bilingual Dictionary Intermediated by a Third Language. *COLING '94: Proceedings of the 15th conference on Computational linguistics*, pp. 297-303.

Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified data modelling for presenting lexical data: The Case of EKILEX. In J. Čibej V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress. EURALEX: Lexicography in Global Contexts, Ljubljana, 17–21 July 2018. Ljubljana: Ljubljana University Press, Faculty of Arts*, pp. 749−761.

Parmentier, Y. & Waszczuk, J. 2019. Preface. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, iii–ix. Berlin: Language Science Press.

Websites:

Basic Estonian Dictionary. Accessed at: http://www.eki.ee/dict/psv/. (21 March 2021)

dict.cc. Accessed at: http://www.dict.cc. (21 March 2021)

DictionaryForMIDs. Accessed at: https://sourceforge.net/projects/dictionarymid/. (21 March 2021)

EKI Combined Dictionary. DOI: 10.15155/3-00-0000-0000-0000-08979L. Accessed at: https://metashare.ut.ee/repository/browse/the-eki-combined-dictionary-2021/af363d08857111eba6e4fa163e9d4547c858d4634fcb44eea7b56db3e452675c/. (21 March 2021)

English-Estonian dictionary. Accessed at: https://www.eki.ee/litsents/. (21 March 2021)

EstNLTK. Accessed at: https://github.com/estnltk/estnltk. (21 March 2021)

Estonian-Slovak dictionary web application. Accessed at: https://estonian-slovak-dictionary.herokuapp.com/index/ee. (23 March 2021)
Estonian WordNet. Accessed at: https://www.cl.ut.ee/ressursid/teksaurus/index.php?lang=et. (21 March 2021)
Frequency lists of The Balanced Corpus of Estonian. Accessed at: https://www.cl.ut.ee/ressursid/sagedused1/index.php?lang=en. (21 March 2021)
Google translation API. Accessed at: https://cloud.google.com/translate/docs/ (21 March 2021)
Natural Language Toolkit. Accessed at: https://www.nltk.org/ (28 March 2021)
Polyglot. Accessed at: https://polyglot.readthedocs.io/en/latest/. (21 March 2021)
Slovak POS Tagger. Accessed at: http://morpholyzer.fiit.stuba.sk:8080/PosTagger/. (21 March 2021)
Slovak Wordnet. Accessed at: https://korpus.sk/WordNet.html. (21 March 2021)
Sõnaveeb. Accessed at: https://sonaveeb.ee/. (21 March 2021)
WordNet. Accessed at: https://wordnet.princeton.edu/. (7 April 2021)