ZSANETT FERENCZI

AUTOMATIC DICTIONARY BUILDING METHODS FOR FINNO-UGRIC LANGUAGES

Doctoral (PhD) dissertation

THESIS BOOKLET

Pázmány Péter Catholic University
Faculty of Humanities and Social Sciences
Doctoral School of Linguistics
Language Technology Workshop

Supervisor:
Eszter Simon PhD

Budapest
2023

1 Aims

The primary aim of the thesis is to create and evaluate automatic dictionary building methods for two Finno-Ugric languages (Finnish and Hungarian), and to establish which one of these methods performs best in terms of their precision. A further goal is to improve interoperability between many existing resources for these languages and to provide a language-independent database which can be utilized as the basis of an automatically reversible online dictionary. Another objective is the creation of a language learning application, which would help learners of Finnish and learners of Hungarian to practice different aspects of these languages.

2 Data and methods

The methods presented in this thesis use three kinds of resources: Wiktionary, WordNet and OPUS. Each of the proposed scripts parses the contents of one of these resources and obtains bilingual translation candidates and other lexical information, such as definitions and example sentences for Finnish and Hungarian.

Resources: WordNet (Miller 1995) is a lexical database which organizes nouns, verbs, adjectives and adverbs into synonym sets (sets of words, synonyms describing a specific concept). The English WordNet was translated into several languages, including Finnish (Lindén and Carlson 2010) and Hungarian (Miháltz et al. 2008). Wiktionary (https://www.wiktionary.org/) is an online multilingual dictionary. It is a crowd-sourced project, and has many language editions. The language of the edition defines the target language of the dictionary, and any language can serve as source language in this resource. Three editions of Wiktionary (the Finnish, the Hungarian and the English language editions) were used in order to obtain bilingual translation pairs, definitions and example sentences. The OPUS corpus (Tiedemann and Nygaard 2004) is a collection of publicly available data. It contains several types of resources, such as parallel texts and bilingual word alignments for several language pairs, including the Finnish–Hungarian pair.

Natural Language Processing Tools: During the research work, it was observed that the bilingual translation candidates obtained from OPUS reach low quality at the preliminary evaluation, because they were extracted from running texts, and the list contained different word forms instead of lemmata. To investigate the impact of lemmatization in this case, it was necessary to utilize language processing tools. In another phase of the project, in order to generate language learning exercises, example sentences needed to be tok-

enized, lemmatized, morphologically analyzed, and dependency parsed. For these tasks, I used the emtsv (Indig et al. 2019) text processing pipeline for Hungarian, and the omorfi (Pirinen 2015) and uralicNLP (Hämäläinen 2019) tools for Finnish.

3 The structure and the main theses of the dissertation

The dissertation is comprised of six chapters. The first chapter gives a short introduction to the main topics of the research work, including lexicography, vocabulary acquisition and Computer-Assisted Language Learning (CALL). The second chapter explores the different dictionary building methods and proposes three new alternative solutions to extract Finnish–Hungarian bilingual translation pairs from three different resources. Chapter 3 describes the language-independent database that has been created to contain the lexical data obtained with the help of the proposed methods. In the next chapter, the importance and benefits of CALL are emphasized. In Chapter 5, the components of the proposed framework (Finno-Ugric Lexical Resources (FULR)) are introduced. The last chapter concludes the work, summarizes the results, and contains directions for further research. The main theses of the research work and a more detailed description of the main chapters are provided below.

Chapter 2 describes and compares the main resources and approaches which can be used to build bilingual dictionaries with automatic methods. Finnish and Hungarian have several existing lexical resources that had not been exploited before to automatically generate translation pairs. These resources, as well as the utilized language processing tools for these languages are described in this chapter. The proposed dictionary building methods that parse the above-mentioned resources are presented in detail. The results of the intermediate evaluation are reported. The main theses of this chapter are the following:

1. I created Wiktionary Parser, a script that parses the Finnish and Hungarian Wiktionary editions and collects bilingual translation candidates with their part-of-speech information, as well as monolingual lemma–definition and lemma–example sentence pairs. The script is freely available and the resulting data set is uploaded to the FULR database, where every data point will be manually validated.

2. I developed WordNet Connector, a script that links the Finnish and Hungarian WordNets. It collects bilingual translation candidates by connecting the synsets of these two databases. The algorithm can also

extract monolingual synonym lists, as well as Hungarian definitions and example sentences from the Hungarian database. The Finnish WordNet does not contain these kinds of data. The script is freely available and the resulting data set is uploaded to the FULR database, where every data point will be manually validated.

3. I created a script called OPUS Extractor that extracts bilingual word pairs using the Finnish and Hungarian word alignments. It is possible to sort the word pairs by their co-occurence number or in alphabetical order. The script is freely available and the resulting data set is uploaded to the FULR database, where every data point will be manually validated.

4. I lemmatized the data set extracted from the OPUS corpus, since the word alignments were generated from running texts, resulting in only 6.25% precision. As it was also observed by Simon and Mittelholcz (2017), it was found that the extracted translation candidates contained many suffixed word forms, which resulted in a low quality proto-dictionary. Lemmatization led to a 73.13% decrease in the number of word pairs, and resulted in 93.137% precision. This result clearly shows that the precision of bilingual translation pair extraction from running texts in morphologically rich languages can greatly benefit from lemmatization, when the word pairs are intended to be included in a dictionary as headwords.

5. By validating hundreds of translation candidates, synonym pairs, definitions and example sentences, it was possible to compare the precision of the applied methods. I showed that the highest precision could be achieved by the Wiktionary Parser method proposed in this dissertation, followed by one of the modes (extract) of the wikt2dict tool developed by Ács et al. (2013). This proves that the crowd-sourced resource, Wiktionary, contains mostly high-quality, reliable data.

Chapter 3 presents the details of a language-independent lexicographical database that has been created to store the data extracted by the dictionary building methods. A relational database was chosen over the XML data structure to ensure that multiple users can manipulate the data simultaneously. The tables of the proposed database, as well as its views and triggers are described in detail. The main result of this chapter can be summarized as follows:

6. I designed and developed a lexicographical, MariaDB-based relational database that stores data in a language-independent way. This reposi-

tory of data stores all kinds of information (translation candidates, example sentences, definitions, and synonyms) in a universal way. The main units of the proposed structure are entities, as opposed to traditional lexicography approaches which consider entries as the basic units of the dictionary. This database schema makes it possible to create an automatically reversible bilingual dictionary from the obtained data set.

Chapter 4 first introduces how computers can help language learners master foreign languages. This chapter presents the evolution of CALL and the typical activities it can offer to learners, as well as its place in the syllabus of language classrooms. It describes how natural language processing can further improve the quality of CALL applications. At the end of this chapter, some of the limitations of this field are mentioned.
Chapter 5 elaborates on the proposed framework and its components. Following the introductory section, the steps of manual entity validation are presented in the proposed dictionary writing system, that is intended to facilitate the dictionary editing process. The macrostructure and the microstructure of the online bilingual dictionary are described. The following section provides a detailed description of the CALL application and its two modules: a flashcard module that learners can use to learn new vocabulary items, and a cloze tasks module which aims to help learners practice different grammar aspects of both Finnish and Hungarian. The theses related to this chapter are the following:

7. I created the FULR (Finno-Ugric Lexical Resources) platform, which consists of three components: an online bilingual dictionary, a dictionary writing system and a language learning application.

8. I designed the online bilingual dictionary interface for learners of Finnish and learners of Hungarian. This dictionary builds on the data set stored in the language-independent database described earlier. The generation of the dictionary entries happens automatically from the entities and relations found in the database defined by certain rules.

9. I developed the dictionary writing system that allows editors to easily manipulate the data, without advanced IT knowledge.

10. I created the language learning application that incorporates two modules: a flashcard module and a module with cloze exercises. The flashcard module facilitates vocabulary acquisition and offers two types of cards: monolingual cards with the help of lemma–definition pairs, and bilingual flashcards with translation pairs. The cloze exercises can help learners practice different grammar aspects. I determined some rules to

describe three Finnish and three Hungarian grammar aspects in order to automatically generate examples for these exercise types.

11. A subset of the examples for each cloze exercise has been validated and the predetermined rules proved to generate the tasks very precisely (with above 90% precision in all exercise types except one). This method can be applied to alleviate the tedious manual creation of such exercises by language instructors.

One of the prominent theoretical results of the thesis is the evaluation and comparison of different dictionary building methods for morphologically rich languages. New methods have been developed and described, and the positive impact of lemmatization on data gathered from non-lemmatized corpora has been proved.

The practical outcome of the dissertation is a language-independent framework that incorporates an online dictionary, a dictionary writing system, and a language learning application. This framework can be extended with data in other Finno-Ugric languages in the future. The language learning application and the database is structured in a way that learners' responses are stored anonymously. This collected learner data set is expected to provide valuable information for future research that aims to investigate the difficulties that learners face when learning these languages.

4 Relevant publications

Publications:

Ferenczi, Zsanett 2022. Automatically Generated Language Learning Exercises for Finno-Ugric Languages. In: *Linguistics Beyond and Within (LingBaW)*. In press.

Ferenczi, Zsanett. 2022. Nyelvtanulást elősegítő feladatok automatikus előállítása finn és magyar nyelvekre [Automatic Generation of Finnish and Hungarian Language Learning Exercises]. In: Berend, Gábor – Gosztolya, Gábor – Vincze, Veronika (eds.): *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem Informatikai Intézet. 213–226.

Ferenczi, Zsanett 2021. Finn–magyar fordítási párok kinyerése automatikus módszerekkel [Automatic Extraction of Finnish and Hungarian Translation Pairs]. In: Gráczi, Tekla Etelka and Ludányi, Zsófia (eds.): *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2021: XV. Alkalmazott Nyelvészeti Doktorandusz Konferencia.* Budapest: Nyelvtudományi Kutatóközpont. 131–150. https://doi.org/10.18135/Alknyelvdok.2021.15

Ferenczi, Zsanett. 2019. Kenesei István (szerk.): Nyelv, biológia, szabadság. A 90 éves Chomsky jelentősége a tudományban és azon túl [Language, Biology, Freedom. The Impact of the 90-year-old Noam Chomsky in Science and Beyond]. In: Magyar Pszichológiai Szemle. 74(4). Akadémiai Kiadó. 611–614.

Ferenczi, Zsanett. 2019. Jogi szövegek automatikus fordítása [Automatic Translation of Legal Texts]. In: *Édes Anyanyelvünk*. 41(3). 16.

Simon, Eszter – Mittelholcz, Iván – Ferenczi, Zsanett. 2018. Automatikus szótárépítés kisebbségi finnugor nyelvekre [Automatic Dictionary Building for Finno-Ugric Minority Languages]. In: Pletl, Rita and Kovács, Gabriella (eds.): *Soknyelvűség és többnyelvűség Európában*. Cluj-Napoca: EME-Scientia Publishing House. 53–64.

Ferenczi, Zsanett – Mittelholcz, Iván – Simon, Eszter – Váradi, Tamás. 2018. Evaluation of Dictionary Creating Methods for Finno-Ugric Minority Languages. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). 1989–1994.

Simon, Eszter – Mittelholcz, Iván – Ferenczi, Zsanett. 2018. Lexikai erőforrások automatikus előállítása kisebbségi finnugor nyelvekre [Automatic Creation of Lexical Resources for Finno-Ugric Minority Languages]. In: Veronika Vincze (ed.): *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport. 260–271.

Ferenczi, Zsanett – Mittelholcz, Iván – Simon, Eszter. 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)*. Helsinki, Finland: Association for Computational Linguistics. 39–50.

Conference presentations:

Ferenczi, Zsanett. 2022. Nyelvtanulást elősegítő feladatok automatikus előállítása finn és magyar nyelvekre [Automatic Generation of Finnish and Hungarian Language Learning Exercises]. *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Hosted virtually at the University of Szeged, 27–28 January 2022. (talk)

Ferenczi, Zsanett. 2021. Automatically Generated Language Learning Exercises for Finno-Ugric Languages. *Linguistics Beyond And Within*. Hosted virtually at John Paul II Catholic University of Lublin, 14–15 October 2021. (talk)

Ferenczi, Zsanett. 2021. Automatic Generation of Vocabulary and Grammar Exercises for Finnish and Hungarian. *Tenth Workshop on NLP4CALL*. Hosted virtually, 31 May 2021. (talk)

Ferenczi, Zsanett. 2021. Finn–magyar fordítási párok kinyerése automatikus módszerekkel [Automatic Extraction of Finnish and Hungarian Translation Pairs]. *Alkalmazott Nyelvészeti Doktorandusz Konferencia*. Hosted virtually, 5 February 2021. (talk)

Ferenczi, Zsanett – Mittelholcz, Iván – Simon, Eszter – Váradi, Tamás. 2018. Evaluation of Dictionary Creating Methods for Finno-Ugric Minority Languages. *Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, 7–12 May 2018. (poster)

Ferenczi, Zsanett. 2018. Szócikképítés a Wiktionaryben lépésről lépésre [Creating Dictionary Entries in Wiktionary Step by Step]. *Alkalmazott nyelvészeti kutatások a kisebbségi finnugor nyelvek szolgálatában*. Budapest, 13 February 2018. (talk)

Simon, Eszter – Mittelholcz, Iván – Ferenczi, Zsanett. 2018. Lexikai erőforrások automatikus előállítása kisebbségi finnugor nyelvekre [Automatic Creation of Lexical Resources for Finno-Ugric Minority Languages]. *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szeged, 18–19 January, 2018. (talk)

Ferenczi, Zsanett – Mittelholcz, Iván – Simon, Eszter. 2018. Automatic Generation of Wiktionary Entries for Finno-Ugric Minority Languages. *Fourth International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018)*. Helsinki, 8–9 January 2018. (poster)

References

Ács, J., Pajkossy, K., and Kornai, A. (2013). Building Basic Vocabulary Across 40 Languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, (pp. 52–58). Sofia, Bulgaria, Association for Computational Linguistics.

Hämäläinen, M. (2019). UralicNLP: An NLP Library for Uralic Languages. *Journal of Open Source Software*, 4(37):1345.

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., and Makrai, M. (2019). One Format to Rule Them All – The emtsv Pipeline for Hungarian. In *Proceedings of the 13th Linguistic Annotation Workshop*, (pp. 155–165). Florence, Italy, Association for Computational Linguistics.

Lindén, K. and Carlson, L. (2010). FinnWordNet–Finnish WordNet by Translation. *LexicoNordica–Nordic Journal of Lexicography*, 17:119–140.

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., and Váradi, T. (2008). Methods and Results of the Hungarian WordNet Project. In *Proceedings of The Fourth Global WordNet Conference*, (pp. 311–321). Szeged, Hungary.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Pirinen, T. A. (2015). Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. *SKY Journal of Linguistics*, 28:381–393.

Simon, E. and Mittelholcz, I. (2017). Evaluation of Dictionary Creating Methods for Under-Resourced Languages. In *International Conference on Text, Speech, and Dialogue*, (pp. 246–254). Prague, Czech Republic, Springer.

Tiedemann, J. and Nygaard, L. (2004). The OPUS Corpus - Parallel and Free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal.