FERENCZI ZSANETT

# AUTOMATIKUS SZÓTÁRÉPÍTÉSI MÓDSZEREK FINNUGOR NYERLVEKRE

Doktori (PhD) értekezés

# TÉZISFÜZET

Pázmány Péter Katolikus Egyetem
Bölcsészet- és Társadalomtudományi Kar
Nyelvtudományi Doktori Iskola
Nyelvtechnológia Műhely

**Témavezető:**
**Dr. Simon Eszter**

Budapest
2023

## 1. Célkitűzés

A disszertáció elsődleges célja automatikus szótárépítési módszerek létrehozása és kiértékelése, valamint ezen metódusok minőségének összevetése a finn és magyar nyelvek számára. A doktori kutatás további célja az ezen nyelvekre már létező erőforrások közötti átjárhatóság megteremtése, és egy olyan nyelvfüggetlen adatbázis biztosítása, amely egy kétnyelvű reverzibilis online szótár alapjául szolgálhat. Célja továbbá egy nyelvtanuló alkalmazás létrehozása finnül, valamint magyarul tanulók számára.

## 2. A kutatás módszere

A disszertációban bemutatott módszerek háromféle forrást használnak: a Wiktionaryt, a WordNetet és az OPUS korpuszt. A kidolgozott szkriptek e források tartalmát dolgozzák fel, és kétnyelvű fordításjelölteket, valamint egyéb lexikai információkat (pl. definíciókat és példamondatokat) nyernek ki a finn és a magyar nyelvre.

**Erőforrások:** A WordNet (Miller 1995) egy olyan lexikális szemantikai adatbázis, amely a főneveket, igéket, mellékneveket és határozószókat szinonima-halmazokba, azaz synsetekbe rendezi, és ezek között különböző szemantikai relációkat ír le. Az angol WordNetet több nyelvre is lefordították, többek között finnre (Lindén and Carlson 2010) és magyarra (Miháltz et al. 2008) is. A Wiktionary (https://www.wiktionary.org/) egy többnyelvű online szótár. Ezen erőforrást az interneten bárki szerkesztheti, és számos különböző nyelvű kiadása létezik. A kiadás nyelve határozza meg a szótár célnyelvét, míg forrásnyelvként a világ bármely nyelve szolgálhat bármely kiadásban. Három Wiktionary kiadást (a finn, a magyar és az angol nyelvű változatot) használtunk a kétnyelvű fordítási párok, definíciók és példamondatok kinyeréséhez. Az OPUS korpusz (Tiedemann and Nygaard 2004) olyan adatgyűjtemény, melyet szabadon elérhető szövegek alkotnak. Az adatbázis többféle erőforrást tartalmaz, pl. párhuzamos korpuszokat és kétnyelvű szópárlistákat számos nyelvpárra, köztük a finn–magyar nyelvpárra.

**Nyelvtechnológiai eszközök:** Az OPUS-ból kinyert kétnyelvű fordításjelöltek az előzetes kiértékelés során feltűnően alacsony pontosságot értek el. Mivel ezek a fordítási párok elemzetlen szövegekből lettek kinyerve, a kapott lista lemmák helyett különböző (ragozott) szóalakokat tartalmazott. Ahhoz, hogy a lemmatizálás hatását megvizsgálhassam ezen esetben, különböző nyelvfeldolgozó eszközök felhasználására volt szükség. A kutatás egy későbbi szakaszában a nyelvtanulást elősegítő gyakorlatok létrehozásához példamondatokat kellett tokenizálni, lemmatizálni és morfológiailag elemezni,

valamint függőségi elemzést végezni ezen mondatokon. A magyar nyelvű adatokon az emtsv (Indig et al. 2019) szövegfeldolgozó eszközláncot, míg a finn adatok elemzéséhez az omorfi (Pirinen 2015) és uralicNLP (Hämäläinen 2019) eszközöket használtam.

## 3. A disszertáció felépítése és főbb tézisei

A disszertáció hat fejezetből áll. Az első fejezet bemutatja a kutatást érintő főbb témákat, köztük a lexikográfiát, a szókincsfejlesztést és a számítógéppel támogatott nyelvtanulást (Computer-Assisted Language Learning (CALL)). A második fejezet ismerteti a különböző szótárépítési eljárásokat, majd bemutatja azt a három új alternatív megoldást, amelyek segítségével finn–magyar kétnyelvű fordítási párokat nyerhetünk ki különböző forrásokból. A harmadik fejezet ismerteti egy olyan nyelvfüggetlen adatbázis felépítését, amely alkalmas a kinyert lexikai adatok tárolására. A negyedik fejezet a CALL fontosságát és előnyeit hangsúlyozza. Az ötödik fejezetben a kutatás során létrehozott, szabadon hozzáférhető lexikai erőforrás (Finno-Ugric Lexical Resources (FULR)) komponensei kerülnek bemutatásra. A disszertáció utolsó fejezete összefoglalja a kutatás eredményeit, és további kutatási irányokat vázol fel. A kutatómunka téziseit és a főbb fejezetek részletes leírását az alábbiakban ismertetem.

**A második fejezet** ismerteti és összehasonlítja azokat a főbb erőforrásokat és megközelítéseket, amelyeket gyakran alkalmaznak kétnyelvű szótárak automatikus létrehozása során. A finn és a magyar nyelvek számos olyan lexikai erőforrással rendelkeznek, amelyeket korábbi kutatások nem használtak fel kétnyelvű fordítási párok automatikus generálására. Ebben a fejezetben ezeket az erőforrásokat, valamint a felhasznált nyelvtechnológiai eszközöket ismertetem. Részletesen bemutatok három új szótárépítési módszert, amelyek a fent említett erőforrásokat (Wiktionary, WordNet, OPUS korpusz) dolgozzák fel. A fejezet ismerteti a módszerek előzetes kiértékelésének eredményeit. A fejezetben található fő téziseket a következőképpen lehet összefoglalni:

1. Létrehoztam a Wiktionary Parser szkriptet, amely elemzi a finn és a magyar Wiktionary kiadásokat, és kétnyelvű fordításjelölteket (szófaji információval ellátva), valamint egynyelvű lemma–definíció és lemma–példamondat párokat nyer ki ezen erőforrásokból. A szkript szabadon hozzáférhető, és a kapott adathalmaz a FULR adatbázisában is elérhető, ahol minden adat kézi validáláson esik át.

2. Létrehoztam a WordNet Connector szkriptet, amely összekapcsolja a finn és a magyar WordNetet. Ez a módszer kétnyelvű fordításjelölteket nyer ki a két adatbázis synsetjeinek összekapcsolásával. Az algoritmus alkalmas egynyelvű szinonima-listák létrehozására is, valamint a magyar adatbázisból definíciókat és példamondatokat is kinyerhetünk a segítségével. A finn WordNet nem tartalmaz ilyen adatokat. A szkript szabadon hozzáférhető, és a kapott adathalmaz a FULR adatbázisában is elérhető, ahol minden adat kézi validáláson esik át.

3. Létrehoztam az OPUS Extractor szkriptet, amely kétnyelvű szópárokat nyer ki a finn és a magyar szópárhuzamosítások felhasználásával. A szkript lehetőséget ad a szópárok együttes előfordulásának száma szerinti kilistázására, valamint ábécé sorrendben történő kiírására. A szkript szabadon hozzáférhető, és a kapott adathalmaz a FULR adatbázisában is elérhető, ahol minden adat kézi validáláson esik át.

4. Az OPUS korpuszból kinyert adathalmazt lemmatizáltam, mivel a szópárhuzamosítások elemzetlen szövegekből lettek generálva, és az így kapott szópárok csak 6,25%-os pontosságot értek el. Ahogyan azt Simon and Mittelholcz (2017) is megfigyelte az OPUS korpusszal kapcsolatban, a kinyert fordításjelöltek között számos ragozott szóalak is megjelent, amely kevésbé jó minőségű proto-szótárhoz vezetett. A lemmatizálás eredményeként a szópárok száma 73,13%-kal csökkent, viszont a pontosság 93,137%-ra javult. Ez az eredmény azt bizonyítja, hogy a lemmatizálás javíthat a szópárkinyerés pontosságán a morfológiailag gazdag nyelvek esetében, amennyiben a szópárokat elemzetlen korpuszokból nyerjük ki, és ezeket címszóként kívánjuk felvenni egy kétnyelvű szótárba.

5. Több száz fordításjelölt, szinonimapár, definíció és példamondat kézi ellenőrzésével sikerült meghatározni az alkalmazott módszerek pontosságát és összevetni azokat. Kimutattam, hogy a legjobb pontosságot a kutatás során létrehozott Wiktionary Parser módszerrel lehet elérni, amelyet az Ács et al. (2013) által fejlesztett wikt2dict eszköz egyik módja (extract) követ. Ez azt bizonyítja, hogy a Wiktionary, amely egy közösség által szerkesztett szótár, többnyire jó minőségű, megbízható adatokat tartalmaz.

**A harmadik fejezet** részletesen bemutatja azt a nyelvfüggetlen lexikográfiai adatbázist, amelyet a szótárépítési módszerekkel kinyert adatok tárolására hoztam létre. Az XML adatstruktúra helyett egy relációs adatbázist választottam erre a célra, mely biztosítja, hogy egy időben több felhasználó is módo-

síthassa az adatokat. A létrehozott adatbázis tábláit, valamint az adatbázisban definiált nézettáblákat és triggereket is ebben a fejezetben ismertetem. A fejezet fő eredménye a következőképpen foglalható össze:

**6.** Megterveztem és megvalósítottam egy MariaDB-alapú lexikográfiai adatbázist, amely nyelvfüggetlen módon képes adatokat tárolni. Ez az adattár mindenféle információt (pl. fordításjelölteket, példamondatokat, szinonimákat) univerzális módon ragad meg. A javasolt struktúra alapegysége az entitás, szemben a hagyományos lexikográfiai megközelítésekkel, amelyek a szócikket tekintik a szótár alapegységének. Az itt bemutatott adatbázis-séma lehetővé teszi, hogy a kapott adathalmazból egy kétnyelvű reverzibilis online szótárt hozzunk létre.

**A negyedik fejezetben** azt mutatom be, hogy a számítógépek hogyan lehetnek a nyelvtanulók segítségére idegen nyelvek elsajátításakor. Ez a fejezet egyrészt ismerteti a CALL evolúcióját és az egyes időszakokra jellemző típusfeladatokat, másrészt bemutatja a CALL nyelvoktatásban elfoglalt helyét. A fejezetben megemlítem, hogy nyelvtechnológiai eszközök segítségével tovább javítható bizonyos CALL-alkalmazások minősége, valamint felhívom a figyelmet azokra a korlátokra, amelyek a számítógéppel támogatott nyelvtanulás területén leggyakrabban megfigyelhetők.
**Az ötödik fejezet** részletesen ismerteti a létrehozott keretrendszert és annak komponenseit. A bevezetést követően az entitások kézi kiértékelésének lépéseit mutatom be az általam javasolt szótáríró rendszerben, amely a szótárszerkesztés folyamatát hivatott megkönnyíteni. Ezután az online kétnyelvű szótár makro- és mikroszerkezetét ismertetem. A következő részben részletesen bemutatom a CALL alkalmazást és annak két modulját: egy digitális szókártya modult, amellyel a tanulók fejleszthetik szókincsüket, valamint egy behelyettesítéses feladatokat tartalmazó modult, amelynek célja, hogy a finnül, illetve magyarul tanulók különböző nyelvtani aspektusokat gyakorolhassanak. A fejezethez kapcsolódó tézisek a következők:

**7.** Létrehoztam a FULR (Finno-Ugric Lexical Resources, finnugor lexikai erőforrások) platformot, amely három komponensből áll: egy online kétnyelvű szótárból, egy szótáríró rendszerből és egy nyelvtanuló alkalmazásból.

**8.** Megterveztem és felépítettem egy online kétnyelvű szótár felületét finnül, valamint magyarul tanulók számára. Ez a szótár a fent említett nyelvfüggetlen adatbázisban tárolt adathalmazra épül. A szócikkek generálása bizonyos szabályok segítségével automatikusan történik az adatbázisban található entitások és relációk felhasználásával.

9. Létrehoztam egy online szótáríró rendszert, amely lehetővé teszi, hogy a szótár szerkesztői az adatokat egyszerűen, különösebb informatikai ismeretek nélkül módosíthassák.

10. Létrehoztam egy nyelvtanuló alkalmazást, amely két modult tartalmaz: egy digitális szókártya modult és egy behelyettesítéses feladatokat tartalmazó modult. A szókártya modul lehetővé teszi a finn vagy magyar nyelv szókincsének gyakorlását. Ez kétféle formában érhető el: a modul egynyelvű kártyákat kínál a lemma–definíció párok felhasználásával, valamint kétnyelvű szókártyákat a fordítási párok segítségével. A behelyettesítéses feladatok célja a különböző nyelvtani szerkezetek elsajátításának elősegítése. Különböző szabályokat határoztam meg 3 finn és 3 magyar nyelvtani szerkezet leírására, melyek segítségével példák ezrei generálhatók automatikusan a különböző feladattípusokra.

11. Minden feladattípus esetén kiértékeltem a példák egy részét, és megfigyelhető volt, hogy az előre meghatározott szabályok segítségével nagy pontossággal lehetett behelyettesítéses feladatokat generálni (egyet kivéve minden feladattípus esetén 90% feletti pontosságot kaptam). Az itt bemutatott CALL alkalmazás alkalmas lehet arra, hogy felgyorsítsa a feladatok manuális létrehozásának folyamatát, megkönnyítve ezzel a nyelvtanárok munkáját.

A dolgozat egyik legfontosabb elméleti eredménye a különböző szótárépítési módszerek kiértékelése és összehasonlítása morfológiailag gazdag nyelvek számára. Három új szótárépítési módszert is kidolgoztam, és bemutattam, hogy a lemmatizálásnak pozitív hatása van az elemzetlen korpuszokból gyűjtött adatok pontosságára.

A disszertáció gyakorlati eredménye egy olyan keretrendszer, amely egy online szótárt, egy szótáríró rendszert és egy nyelvtanuló alkalmazást foglal magába. Ez a nyelvfüggetlen keretrendszer más finnugor nyelvek adataival is bővíthető a jövőben. A nyelvtanuló alkalmazás és az adatbázis úgy van felépítve, hogy a tanulók válaszait anonim módon rögzítse. Ez az adathalmaz várhatóan értékes információval fog szolgálni újabb nyelvészeti kutatások számára, amelyek azt vizsgálják és mérik tanulói adatokon, hogy a finnül, illetve magyarul tanulók számára mely nyelvtani szerkezetek okoznak nehézséget.

## 4. A témában végzett publikációs tevékenység

**Publikációk:**

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

Ferenczi Zsanett. 2022. Nyelvtanulást elősegítő feladatok automatikus előállítása finn és magyar nyelvekre. 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. 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. In: Magyar Pszichológiai Szemle. 74(4). Akadémiai Kiadó. 611–614.

Ferenczi Zsanett. 2019. Jogi szövegek automatikus fordítása. In: *Édes Anyanyelvünk*. 41(3). 16.

Simon Eszter – Mittelholcz Iván – Ferenczi Zsanett. 2018. Automatikus szótárépítés kisebbségi finnugor nyelvekre. 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. 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.

**Konferencia-szereplések:**

Ferenczi Zsanett. 2022. Nyelvtanulást elősegítő feladatok automatikus előállítása finn és magyar nyelvekre. *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Online, 2022. január 27–28. (előadás)

Ferenczi Zsanett. 2021. Automatically Generated Language Learning Exercises for Finno-Ugric Languages. *Linguistics Beyond And Within*. Online, 2021. október 14–15. (előadás)

Ferenczi Zsanett. 2021. Automatic Generation of Vocabulary and Grammar Exercises for Finnish and Hungarian. *Tenth Workshop on NLP4CALL*. Online, 2021. május 31. (előadás)

Ferenczi Zsanett. 2021. Finn–magyar fordítási párok kinyerése automatikus módszerekkel. *Alkalmazott Nyelvészeti Doktorandusz Konferencia*. Online, 2021. február 5. (előadás)

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, Japán, 2018. május 7–12. (poszter)

Ferenczi Zsanett. 2018. Szócikképítés a Wiktionaryben lépésről lépésre. *Alkalmazott nyelvészeti kutatások a kisebbségi finnugor nyelvek szolgálatában*. Budapest, 2018. február 13. (előadás)

Simon Eszter – Mittelholcz Iván – Ferenczi Zsanett. 2018. Lexikai erőforrások automatikus előállítása kisebbségi finnugor nyelvekre. *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018)*. Szeged, 2018. január 18–19. (előadás)

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, 2018. január 8–9. (poszter)

## Hivatkozások

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*, pages 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*, pages 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*, pages 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*, pages 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.