Az EFNILEX első szakasza

Héja Enikő[1]

[1] Nyelvtudományi Intézet
heja.eniko@nytud.hu

1. Bevezetés

A cikkben ismertetett munka az EFNIL által finanszírozott EFNILEX projekt első szakasza, amely 2008-tól 2012-ig tartott. A projekt azt vizsgálta, hogy a nyelvtechnológiai módszerek és eszközök – különös tekintettel a párhuzamos korpuszokon végzett szóillesztésre – mennyiben képesek támogatni a szótárkészítési folyamatot. A szótárkészítés automatikus támogatása elsősorban a kevéssé használt nyelvek esetében bír jelentőséggel, hiszen az ilyen nyelvpárokra íródott szótárakra alacsony a kereslet, így az ilyen munkálatok finanszírozása is korlátozott. A bemutatandó munka eredeti célja egy közepes méretű (kb. 15 000 szócikk) litván–magyar szótár létrehozása volt. A munkafolyamat részeként tesztelési célokra a magyar–szlovén nyelvpárt is vizsgáltuk.

A projektben részt vettek: František Čermak (Institute of the Czech National Corpus, Charles University), John Simpson (Oxford English Dictionary) és Jolanta Zabartskaite (Institute of the Lithuanian Language). A projektet Váradi Tamás koordinálta az MTA Nyelvtudományi Intézet részéről.

Az elvégzett munkából egy disszertáció is született Váradi Tamás témavezetésével (Héja, 2016), melynek főbb eredményeit az *International Journal of Lexicography* is közölte (Héja, 2017).

Az EFNILEX projekt célja egy megfelelő lefedettségű és pontosságú lexikai erőforrás előállítása volt, amely emberi utószerkesztési munkálatokat is igényel. Vagyis célunk az volt, hogy a lexikográfusok számára olyan erőforrásokat biztosítsunk, amelyek a lehető legjobban csökkentik a teljes értékű, emberi felhasználásra alkalmas szótárak elkészítéséhez szükséges munkát. A fenti elvárásoknak megfelelő automatikusan generált erőforrásokat protoszótáraknak fogjuk nevezni a cikk hátralevő részében. Az automatikusan létrehozott protoszótárak az efnilex.efnil.hu weboldalon kérdezhetőek le.

DOI: https://doi.org/10.18135/VT70.7

Az általunk javasolt módszer alapját párhuzamos korpuszokon végzett automatikus szóillesztés képezi. Bár az automatikus szóillesztést széles körben használták szótárfejlesztésre elsősorban a gépi fordítás területén, amennyire tudjuk, a kutatás előtt ezt a megközelítést nem használták emberi felhasználásra szánt szótárak készítésének támogatására lexikográfiai projektekben.

A megfelelő módszer kiválasztásakor egyik fő szempontunk az volt, hogy a lehető legnagyobb mértékben csökkentsük a lexikográfusi nyelvi intuíció szerepét a szótárkészítési folyamat során, ezért a protoszótárakat felügyelet nélküli tanulási algoritmussal akartuk létrehozni. Elsősorban ezért választottuk ezt a módszert például a *hub-and-spoke* (Martin, 2007) modellel szemben, amelynek lényege, hogy már létező egynyelvű adatbázisokat (pl. egynyelvű értelmező szótárakat vagy wordneteket) köt össze egy olyan sok erőforrással rendelkező közbülső nyelv felhasználásával, amely rendelkezik kétnyelvű szótárakkal mind a forrásnyelv mind a célnyelv tekintetében.

A cikk hátralévő részében a munkafolyamatot, valamint az elért eredményeket mutatom be.

2. A munkafolyamat leírása

A munkafolyamtot számos cikkben ismertettem már (pl. Héja, 2010, Héja és Takács, 2012), így ezt most csak vázlatosan fejtem ki.

A protoszótárak készítése minden vizsgált nyelvpár esetében három lépésből állt. Az első szakaszban elkészítettük a párhuzamos korpuszok egységes morfológiai annotációval ellátott XML-verzióját a vizsgált nyelvpárokra. A második szakaszban a párhuzamos korpuszok lemmatizált változatából szóillesztéssel létrehoztuk a protoszótárakat. A harmadik szakaszban kiértékeltük a protoszótárakat.

2.1. A párhuzamos korpuszok elkészítése

A projekt során az eredetileg célul kitűzött nyelvpárok mellett (litván–magyar, szlovén–magyar) további nyelvpárokat, a francia–holland, illetve az angol–magyar nyelvpárokat is vizsgáltunk. Mivel ez utóbbiak esetében volt elérhető párhuzamos korpusz (DPC [Macken és mtsai., 2007], Hunglish [Varga és mtsai., 2005]), csak a litván–magyar, illetve szlovén–magyar nyelvpárokra építettünk párhuzamos korpuszt. Az első szakasz során a szövegeket összegyűjtöttük, normalizáltuk és morfológiailag elemeztük. A párhuzamosítást a hunalign

(Varga és mtsai., 2005) mondatillesztővel végeztük el. A párhuzamosítást egynyelvű szövegek lemmatizált változatain végeztük el. Következő lépésben elkészítettük a morfológiailag elemzett párhuzamos korpuszok egységes XML-reprezentációját. A munkaszakaszhoz kapcsolódó legfontosabb tapasztalat az volt, hogy a javasolt módszer legnagyobb nehézségét a kevéssé használt nyelvpárok esetében a digitálisan elérhető párhuzamosítható szövegek mennyisége jelenti. Amennyiben a megfelelő méretű párhuzamos korpusz rendelkezésre áll, a protoszótár már könnyen előállítható.

2.2. A szóillesztés

A második szakaszban a protoszótárakat állítottuk elő a párhuzamos korpuszok alapján. A szótárkinyerésre a GIZA++-t (Och és Ney, 2003) választottuk a számos versengő módszer közül (pl. Ribeiro és mtsai., 2000). A módszer lényege, hogy a szóillesztés elvégzése során a forrásnyelv és a célnyelv lemmái között feltételes valószínűséget becsül ($P(lemma_{cél}|lemma_{forrás})$). Az így becsült szópárok közötti feltételes valószínűségek képezik a protoszótárak alapját. A lehetséges szótárkinyerő algoritmusok közül azért éppen ezt választottuk, mert azt gondoltuk, hogy aszimmetrikus jellegénél fogva a feltételes valószínűségek becslése különösen alkalmas megfordítható, kódoló szótárak előállítására. Erről bővebben lesz még szó jelen írás 3.2 fejezetében.

2.3. A kiértékelés

A munkafolyamat harmadik szakasza a kiértékelés volt. A kiértékelést több lépésben végeztük. Az első lépésben az annotátorok közötti egyeztetés után megállapítottuk, hogy milyen típusú hibák fordulnak elő a fordítási jelöltek között. Ezen tapasztalatok alapján készítettünk egy kiértékelési útmutatót is, amely során törekedtünk arra, hogy disztribúciós alapokon határozzuk meg az egyes hibatípusokat. Ezzel az volt a célunk, hogy a lehető legjobban lecsökkentsük a szubjektív egyéni nyelvi értékítélet szerepét a kiértékelés során.

A kiértékelés második lépése már az útmutató alapján zajlott. Ennél a lépésnél már nemcsak a fordítási valószínűséget vettük figyelembe, de a forrásnyelvi és célnyelvi lemmák előfordulási gyakoriságát is felvettük paraméternek.

A kiértékelés második szakasza alapján a következő főbb következtetéseket tettük: (1) Szükséges a lemmáknak egy minimális előfordulási

gyakorisága ahhoz, hogy becsülhető legyen a valószínűség. (2) Ha van megfelelő mennyiségű adat, akkor általában igaz az, hogy minél nagyobb a fordítási valószínűség, annál jobb a fordítás. (3) De magas fordítási valószínűség esetén is lehet nagy a hibás fordítások aránya: gyakran előforduló forrásnyelvi lemma és ritka célnyelvi fordításjelölt esetén, ha a forrásnyelvi és célnyelvi lemmák sokszor fordulnak elő együtt párhuzamos mondatokban. Azért, hogy az ilyen eseteket kiszűrjük, az eddigi paraméterek mellett figyelembe vettük még a forrásnyelvi és célnyelvi lemmák gyakoriságának hányadosát is: ennek egy előre meghatározott küszöbérték alatt kellett maradnia. (4) A következő megfigyelésünk az volt, hogy a forrásnyelvi lemmák gyakoriságai és a fordítási valószínűségek „fordítottan arányosak": azaz minél gyakrabban fordul elő a forrásnyelvi lemma, annál kisebb fordítási valószínűségek is még jó fordításokat eredményeznek.

Így harmadik lépésben egy sávos kiértékelést is elvégeztünk, amely során a forrásnyelvi lemma növekvő gyakorisági intervallumaihoz csökkenő valószínűségi küszöbértékeket rendeltünk. Azt találtuk, hogy a fordítási párok ilyen szűrése alkalmas a fedés növelésére is.

3. A projekt eredményei

3.1. Gyakorlati eredmények

A projekt során lekérdezhető protoszótárakat hoztunk létre négy nyelvpárra: magyar–litván (v.v.), magyar–szlovén (v.v.), francia–holland (v.v.), angol–magyar (v.v.). A lekérdezhető protoszótárak megalkotása során az adatbázisok mellett kialakítottunk egy – a hagyományostól némileg eltérő – lekérdezőfelületet is, amely lehetővé teszi a választott módszer adatvezérelt jellegéből fakadó új információk megjelenítését, illetve lekérdezését.

A protoszótárak számos sajátossággal bírnak. Először is: a választott módszer miatt megfordíthatók, így nem négy, hanem nyolc protoszótárat hoztunk létre. A protoszótárak az efnilex.efnil.org weboldalon kérdezhetők le.

A protoszótárak kódoló szótárak, így különösen alkalmasak arra, hogy szövegek írásában segítsék a felhasználót azáltal, hogy hasznos információkat nyújtanak a fordítás helyes használatára vonatkozóan. Egyfelől megjelenítik azokat a párhuzamos kontextusokat, amelyben a forrásnyelvi és a célnyelvi szavak előfordulnak. Ezen túl a protoszótárak segítik a fordítás helyes használatát azzal is, hogy az előfordulási gyakorisá-

gok alapján megbecsülik, hogy a fordítási jelölt használati köre szűkebb-e vagy tágabb, mint a forrásnyelvi szóé. Előbbi esetben a szöveg megalkotásakor a célszó kontextusaira kiemelt figyelmet kell fordítani. A lekérdezhető protoszótárak további érdekessége, hogy testre szabhatók annak függvényében, hogy milyen felhasználói csoportot céloznak meg. Ha csak a leggyakoribb szavak fordításait kérdezzük le magas feltételes valószínűséggel, akkor megkapjuk egy nyelv alapszókincsét kevés, ám biztosan jó fordítási jelölttel. Ez a beállítás kezdő nyelvtanulók számára ajánlott. Ezzel szemben a protoszótárakat úgy is testre szabhatjuk, hogy a ritkább szavak nem tipikus fordításait is megjelenítsék. Ebben az esetben több lesz a hibás fordítási jelölt, de mivel ezekre a fordításokra már elsősorban a biztos nyelvismerettel rendelkezők kíváncsiak, ők kézzel kiszűrhetik a helytelen fordítási jelölteket.

A projekt gyakorlati eredményei közé soroljuk az egységes morfológiai annotációval ellátott litván–magyar, szlovén–magyar és angol–magyar párhuzamos XML-korpuszokat is. A párhuzamos korpuszok méretét az 1. táblázatban adjuk meg:

1. táblázat. A morfológiailag annotált párhuzamos XML-korpuszok mérete.

	Magyar	**Litván**	**Magyar**	**Szlovén**	**Magyar**	**Angol**
Token	4.813.956	4.141.521	723.857	809.448	6.921.127	8.312.795
Mondat	319.489	320.678	40.926	42.659	494.044	494.044
Fordítási egység	304.419		38.791		494.044	

3.2. Elméleti eredmények

A projekt legfontosabb elméleti eredménye, hogy a javasolt módszer, vagyis a fordítási párok automatikus tanulása párhuzamos korpuszokból feltételes valószínűségek becslésével, számos előnnyel rendelkezik a hagyományos és korpuszalapú kétnyelvű lexikográfiai módszerekkel szemben is. Ezek közül a legfontosabb, hogy a javasolt módszer a forrásnyelvi oldalon kiküszöböli a lemmákhoz tartozó egyes jelentések elkülönítésének problémáját. Továbbá, a módszer lehetővé teszi a fordítási reláció korpusz adatokon való kvantifikálható újraértelmezését. A szakirodalom (pl. Atkins és Rundall, 2010, Adamska-Sałaciak, 2010) alapján azt találtuk, hogy a fordítási reláció általában valamilyen értelemben aszimmetrikus és fokozatos. Azt állítjuk, hogy a hagyományos relációs

felfogás helyett a fordítási relációra érdemes feltételes valószínűségként gondolni. Hiszen a feltételes valószínűség megragadja a fordítási reláció aszimmetrikus és fokozatos jellegét. Sőt ezen túlmenően ez a matematikai konstrukció számot ad arról a speciális esetről is, amikor a fordítási reláció szimmetrikus. Ez a tökéletes fordítási ekvivalencia esetében áll fenn.

3. Összefoglalás

A cikkben az EFNILEX projekt első szakaszának (2008–2012) főbb eredményeit ismertettem, melyet az EFNIL tagszervezeteként végeztünk Váradi Tamás koordinálásával. Számos magyar és nemzetközi publikáció mellett a disszertációm is ebből a munkából született, melynek témavezetője szintén Váradi Tamás volt. A disszertáció főbb elméleti eredményei az *International Journal of Lexicography*-ban is megjelentek.

Végezetül néhány személyes gondolatot szeretnék leírni. Az egyetemről frissen kikerülve sokunknak volt a Korpusznyelvészeti, majd később a Nyelvtechnológiai Osztály az első munkahelye. Vezetési stílusából fakadóan Tamás gyakran előlegezett bizalmat nekünk a feladatok kiosztása során. Bár ennek kapcsán olykor előfordult velem, hogy azt éreztem, túl nagy ez a kabát, egyúttal ez nagyon motiváló is volt. Remélem, hogy Tamás is úgy gondolja, hogy ehhez a megelőlegezett bizalomhoz a legtöbb esetben sikerült felnőnünk.

A projekteket és a kapcsolódó kutatásokat gyakran mutathattuk be neves külföldi konferenciákon, amelyet az ünnepelt mindig támogatott anyagilag is, ennek köszönhetően már pályánk elején bekapcsolódhattunk a nemzetközi vérkeringésbe. Így sokunkat Tamás indított el a nyelvtechnológiai pályán. Ezért nagyon hálás vagyok, és ezzel a rövid írással szeretnék boldog 70. születésnapot kívánni Neki. Kedves Tamás, Isten éltessen!

Bibliográfia

Atkins, B. T. S., Rundell, M.: *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford (2008)

Adamska-Sałaciak, A.: Examining Equivalence. *International Journal of Lexicography* 23/4, 387–409 (2010)

Héja E.: Dictionary Building based on Parallel Corpora and Word Alignment. In: Dykstra, A. and Schoonheim, T. (eds) Proceedings of the XIV. EURALEX International Congress. pp. 341–352. Fryske Akademy, Afûk, Ljouwert (2010)

Héja, E.: The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography. Quantifying Translational Equivalence. PhD-értekezés (2016)
Héja, E.: Revisiting Translational Equivalence: Contributions from Data-Driven Bilingual Lexicography International Journal of Lexicography 30/4, 483–503 (2017)
Héja, E., Takács, D.: Automatically Generated Customizable Online Dictionaries. In: Daelemans W. et al. (eds.) Proceedings of EACL2012. pp. 51–57. The Association for Computer Linguistics (2012)
Macken, L., Trushkina, J., Paulussen, H., Rura, L., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus. A multilingual annotated corpus. In: Davies, M., Rayson, P., Hunston, S., Danielsson, P. (eds.) Proceedings of Corpus Linguistics 2007. University of Birmingham, Birmingham, United Kingdom (2007)
Martin, W.: Government Policy and the Planning and Production of Bilingual Dictionaies: The 'Dutch' Approach as a Case in Point, International Journal of Lexicography 20/3, 221–237 (2007)
Ribeiro, A., Pereira Lopes, G., Mexia, J.: Extracting Equivalents from Aligned Parallel Texts: Comparison of Measures of Similarity. In: Monard M.C., Sichman J. S. (eds.) Advances in Artificial Intelligence. IBERAMIA 2000, SBIA 2000. Lecture Notes in Computer Science, vol 1952. pp. 339–349. Springer, Berlin, Heidelberg (2000)
Och, F. J.; Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29/1, 19–51 (2003)
Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. In: Angelova, G., Bontcheva, K., Mitkov, R. Nicolov, N., Nikolov, N. (eds.) Proceedings of the RANLP 2005. pp. 590–596. Borovets, Bulgaria (2005)