Automatikusan generált online szótárak: az EFNILEX projekt eredményei

Héja Enikő, Takács Dávid

MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr u. 33.
{eheja,takdavid}@nytud.hu

Kivonat: Az előadás összefoglalja a 2008-ban kezdődő EFNILEX lexikográfiai projekt munkálatait, különös tekintettel a 2012-ben elért eredményekre. A projekt célja annak a vizsgálata volt, hogy a nyelvtechnológiai eszközök és eljárások mennyiben alkalmasak a kétnyelvű szótárak előállításának támogatására. Ennek elsősorban a kevéssé használt nyelvek esetében van jelentősége, hiszen ezeket a szótárakat a kereskedelmi kiadók nem tartják piacképesnek, így az elkészítésükbe sem invesztálnak jelentősen.

Mivel még nem léteznek olyan módszerek, amelyek a szótárak teljesen automatikus előállítását lehetővé teszik, eredeti célkitűzésünk az volt, hogy a lexikográfusokat olyan automatikusan generált erőforrásokkal lássuk el, amely a lehető legjobban csökkentik a szótárak elkészítéséhez szükséges munkát. Ezeket az erőforrásokat protoszótáraknak nevezzük. Természetesen annak a lehetőségét sem zártuk ki, hogy ezek az erőforrások valamilyen formában közvetlenül is számot tarthatnak a szótárhasználók érdeklődésére.

A protoszótárak előállítási folyamatának lényege, hogy a fordítási ekvivalenciákat szóillesztés útján nyerjük ki kétnyelvű párhuzamos korpuszokból. A módszer egyik előnye, hogy a lexikai megfeleltetések kiválasztása korpuszvezérelt módon történik, amely által a lexikográfusi intuíció csökkenthető. További előny, hogy az egyes fordításokhoz kétnyelvű konkordanciák állnak rendelkezésre, amelyek segítséget nyújtanak az egyes fordítások használati feltételeinek karakterizálásában. Ezen felül, nézetünk szerint ez a fajta megközelítés jobban illeszkedik a szótárhasználók igényeihez, hiszen szótárhasználatkor az elsődleges cél általában szövegek, és nem elszigetelt szavak megértése, illetve létrehozása (pl. [3]).

A protoszótárak újdonsága a hagyományos szótárakkal szemben, hogy a tartalmuk automatikusan testre szabható bizonyos mérőszámok mentén, amelyeket a statisztikai szóillesztés során határozunk meg. Az így testre szabott szótárak egy megfelelő lekérdező felülettel ellátva pedig a szótárhasználók számára már közvetlenül is hasznosak lehetnek. Az online szótárak lekérdezhetőek a http://efnilex.efnil.org oldalon (pl. [4]).

Az eredményül kapott szótárak a módszer lényegéből fakadóan számos újdonságot tartalmaznak a hagyományos szótárakkal szemben. Először is, a szótárak testreszabhatósága lehetővé teszi, hogy a szótárak különböző felhasználói szinteknek feleljenek meg. Így például a megfelelő paraméterek beállításával a felhasználó kiválaszthat egy

olyan szótárt, amely csak a leggyakoribb forrásnyelvi szavakat és ezek legvalószínűbb fordításait tartalmazza. Egy ilyen szótár tökéletes egy kezdő nyelvtanuló számára. Egy további lehetőség, hogy a felhasználó egy viszonylag nagy lefedettségű szótárt vág ki és kérdez le, amely már tartalmazhat hibás fordítási jelölteket is. Egy ilyen szótár a professzionális felhasználók, pl. fordítók számára lehet érdekes, akiket sokszor a speciális fordítási lehetőségek érdekelnek, ugyanakkor rendelkeznek kellő nyelvismerettel ahhoz, hogy az esetleges téves fordítási jelölteket kiszűrjék.

A felhasználói felület további előnyei közé tartozik, hogy a nyelvek közötti szemantikai relációkra is javaslatot tesz. Ez azért is nagyon fontos, mert a szigorú értelemben vett fordítási ekvivalencia – amikor a forrás- és célnyelvi kifejezés pontosan ugyanolyan kontextusokban jelennek meg – ritka jelenség (pl. [1]). Így fontos, hogy a szótár tartalmazzon arra vonatkozó javaslatokat, hogy a célnyelvi kifejezés használata megszorítottabb vagy általánosabb-e, mint a forrásnyelvi kifejezésé (pl. a magyar *tisztán* szó egyaránt fordítható *clearly*-nek és *distinctly*-nek az angolban, de hasznos, ha a szótár jelzi, hogy az utóbbi fordítás megszorítottabb környezetekben fordulhat csak elő).

Mindazonáltal ezek az újdonságok még nem teljes mértékben kidolgozottak, a paraméterbeállítások még további pontosításra szorulnak. Ezenfelül a felhasználói felületet is célunk felhasználóbarátabbá tenni. Részben ezen célokat szolgálja, hogy elkészítettük a Hunglish 1.0-n [5] alapuló angol-magyar, magyar-angol protoszótárainkat is, amelyek lekérdezhetővé tételük után reményeink szerint segítenek eredményeink disszeminálásában, valamint a felhasználói javaslatok alapján a további fejlesztések pontos meghatározásában is.

A módszer hátrányai közé tartozik, hogy a megfelelő méretű párhuzamos korpusz összegyűjtése főleg a kevéssé használt nyelvek esetében nehézkes. További hátrány, hogy a módszer önmagában nem kezeli a többszavas kifejezéseket.

A projekt 2012-es szakaszában a többszavas kifejezések kinyerésére is koncentráltunk, ezen belül is a többnyelvű kollokációk kinyerésére. Többnyelvű kollokációkat a következő nyelvpárokra vontunk ki: magyar-szlovén, magyar-litván, illetve magyar-angol. A munkafolyamat 3 lépésből áll. (1) Minden nyelvre külön-külön kinyerjük az egynyelvű kollokációkat. (2) A kinyert kollokációkat felismerjük a párhuzamos korpuszok releváns részében és egytokenes kifejezéssé alakítjuk, hogy ezek is a szóillesztő algoritmus bemenetéül szolgálhassanak. (3) A szóillesztő algoritmust futtatva nyerjük ki a kollokációkat és a hozzájuk tartozó fordítási jelölteket.

Az egynyelvű kollokációk kinyerése során ezúttal csak szomszédos tokeneket vettünk figyelembe, amelyekre szófaji megkötést is tettünk. A magyar-szlovén, a magyar-litván és a magyar-angol nyelvpárok esetében az AN, AdvV, NN formájú kollokációkat vettük figyelembe, ahol A jelöli a mellékneveket, N a főneveket, Adv a határozószókat és V az igéket. Az angol-magyar esetében további kollokációtípusokat is figyelembe vettünk: a magyar oldalon az NV formájú kollokációkat, melyek az igemódosítós igéket tartalmazzák, angol oldalon pedig a VN formájú kollokációkat, amely kategória, hipotézisünk szerint, elsősorban ige+névelőtlen tárgy szerkezeteket tartalmaz. Az előbbiekkel szemben ez a két szerkezet nem teljesen párhuzamos egymással szintaktikailag, hiszen a magyarban igemódosító pozícióban nemcsak tárgyesetű főnevek jelenhetnek meg (pl. *iskolába jár*). A névelőtlenség miatt

mégis azt gondoltuk, hogy ez a kategória felel meg legjobban az angol VN szerkezeteknek.

A kollokációk kinyerésére az UCS 0.6 szabadon elérhető kollokációkinyerő eszközt használtuk [2]. A kollokációjelöltek az ötnél nem ritkábban előforduló szópárok voltak, amelyek a fent említett formai kritériumoknak megfeleltek. A következő lépésben ezeket két különböző asszociációs mérték szerint szűrtünk: MI és Z-score szerint. A szóillesztő futtatása után azokat a fordítási párokat vettük figyelembe, amelyeknek vagy a célnyelvi vagy a forrásnyelvi oldalán kollokáció szerepelt. Az eredményül kapott fordítási jelölteket lekérdezhetővé tettük. Így az automatikusan generált szótárakból kiderül, hogy az *arc* lehet *beesett*, *eltorzult*, *kipirult*, *sápadt*, és a *beesett arc* egy lehetséges angol fordítása: *hollow cheek*. Az eredmények részletes kiértékelése a közeljövő feladata.

A párhuzamos kollokációk kinyerésének megkönnyítésére egy fontos fejlesztést vezettünk be: minden rendelkezésre álló párhuzamos korpuszt (Hunglish 1.0, litván-magyar, szlovén-magyar) egységes XML-annotációval láttunk el. Ez kettős célt szolgál: (1) a vizsgálni kívánt szerkezeti egységek kinyerését lényegesen megkönnyíti; (2) a kollokációk egységes kezeléséhez célszerűnek tűnt egy (kvázi) egységes morfoszintaktikai annotáció bevezetése. Az így újragenerált szótárak a kollokációkon túl szófaji információt is tartalmaznak, sőt bizonyos esetekben megadják azt is, hogy egy tipikus fordítás jellemzően milyen típusú szövegekben fordul elő.

Hivatkozások

1. Atkins, B.T. S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press (2008)
2. Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, URN urn:nbn:de:bsz:93-opus-23714 (2005)
3. Héja, E.: The Role of Parallel Corpora in Bilingual Lexicography. In: Proceedings of the LREC2010 Conference. La Valletta, Malta, May 2010 (2010) 2798–2805
4. Héja, E., Takács D.: Automatically Generated Customizable Online Dictionaries. In: Daelemans W. et al. (eds.): Proceedings of EACL2012. The Association for Computational Linguistics, Avignon, France (2012) 51–57
5. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy V.: Parallel corpora for medium density languages. In: Proceedings of the RANLP 2005 (2005) 590–596