MANYE XIX.

A tudomány nyelve – A nyelv tudománya

Alkalmazott nyelvészeti kutatások a magyar nyelv évében

XIX. Magyar Alkalmazott Nyelvészeti Kongresszus

Eger, 2009. április 16–18.

Szerkesztő:

Zimányi Árpád

MANYE Vol. 6

MANYE – Eszterházy Károly Főiskola
Székesfehérvár – Eger
2010

ISSN 1786-545X
ISBN 978-963-9894-54-9

Kiadó:

Magyar Alkalmazott Nyelvészek és Nyelvtanárok Egyesülete
8000 Székesfehérvár, Fürdő utca 1.
Tel./fax: (22) 543-311

Eszterházy Károly Főiskola
3300 Eger, Eszterházy tér 1.
Tel.: (36) 520-400

Nyomda: B. V. B. Nyomda és Kiadó Kft.
Felelős vezető: Budavári Sándor

Korpusznyelvészeti módszer a fordításkutatásban

Péch Olívia

0. Bevezetés

Napjaink kommunikációs társadalma erőn felüli kihívásként éli meg az információáramlás növekedésének hatására egyre nagyobb súllyal bíró kultúraközi kommunikáció térnyerését. A hétköznapi élet számos területén exponenciális mértékben sokasodik a fordított szövegek mennyisége (vásárlói tájékoztatók, hírek, uniós dokumentumok stb.), amelyeknél azonban az autentikus célnyelvi szövegek szövegépítési sajátosságaitól eltérő szövegépítési jellemzőket figyelhetünk meg. A fordítástudomány feladatai közé tartozik a fordított szövegekre jellemző szövegalkotási sajátosságok feltárása, és a tapasztalatok felhasználása a fordítóképzésben, az anyanyelvhasználat tudatosításában, illetve a számítógépes nyelvészetben.

Az utóbbi évtizedek robbanásszerű számítástechnikai fejlődésének köszönhetően számos új eszköz, számítógépes alkalmazás segíti a fordítástudomány területére tartozó alkalmazott nyelvészeti kutatásokat. Ezek az eszközök alkalmasak több szöveg együttes vizsgálatára, a forrásnyelvi és a célnyelvi szövegek összehasonlíthatóságát megkönnyítő szövegpárhuzamosításra, gyors, pontos és hatékony gépi lekérdezések elvégzésére.

1. Korpusznyelvészet a fordításkutatásban

A korpuszban összegyűjtött szövegek fordítástudományi vizsgálatának (corpus-based translation studies) előnyére Mona Baker hívta fel először a figyelmet (Baker 1993). Baker a korpuszalapú vizsgálatok mellett szóló legfőbb érvként azt emelte ki, hogy a korpuszban összegyűjtött szövegek a ténylegesen használt nyelv megtestesülései, illetve a nyelvi viselkedés természetes megvalósulásai, melyek a nyelvész beleavatkozása nélkül jöttek létre. A korpuszalapú fordításkutatás tehát azon túl, hogy empirikusan közelíti meg a nyelv leírását, ragaszkodik a hiteles használati előfordulások elsődlegességéhez. A valós nyelvi előfordulások vizsgálatának növekvő népszerűségéről árulkodnak az 1990-es évektől egyre szaporodó korpuszalapú fordításkutatások, melyek közül terjedelmi keretek miatt itt csak a legjelentősebbek említésére szorítkozunk: Baker (1993, 1995, 1996), Laviosa (1997,

1998, 2002), Kennedy (1998), Tymoczko (1998) Stubbs (2001), Kenny (2001), Olohan (2004), Teubert (2007).

A korpuszalapú fordítástudomány népszerűségének jövőbeni növekedését vetíti előre Tymoczko az e kutatási irány által megvalósítható távlati lehetőségek számbavételével (Tymoczko 1998). A fordítástudomány igényeinek megfelelő technológiák alkalmazása és adaptálása mellett Tymoczko kiemeli a korpuszok felhasználásának köszönhetően kutathatóvá váló témák közül a nyelvészeti és kulturális megközelítés integrálását a fordítástudományban és az ideológiák fordításokra gyakorolt befolyásának vizsgálatát. A cikk zárásaként Tymoczko már annak a reményének is hangot ad, hogy a korpuszalapú fordítástudománynak köszönhetően nyitva áll a lehetőség a fordítás elméleti és gyakorlati területének egymáshoz közelítése, sőt akár újraegyesítése előtt.

2. A korpuszok legalapvetőbb ismérvei

Az 1960-as évektől megjelenő korpusznyelvészet (corpus linguistics) az a „nyelvészeti irányzat, mely a nyelv és nyelvhasználat vizsgálatát speciális módszerek és számítógépes programok segítségével korpuszra alapozva végzi" (Szirmai 2005: 170). A korpuszok (corpora) maguk olyan a „nyelvészeti vizsgálatok céljából, bizonyos szempontok alapján összeválogatott írott vagy beszélt nyelvi szövegek" (Szirmai 2005: 170), amelyek elektronikus formában állnak rendelkezésünkre.

A nyelvész korpuszokkal kapcsolatos tevékenységét Kennedy négy nagyobb csoportba sorolja (Kennedy 1998), annak megfelelően, hogy a nyelvészeti tevékenység középpontjában (i) a korpuszépítés (számítógépes tárolásra és feldolgozásra alkalmas szövegek gyűjtése, és ezekből adatbázis építése), (ii) a nyelvleírás (lexikon és/vagy nyelvtan leírása adott nyelv esetében), (iii) a statisztikán alapuló elemzés (az adathalmaznak a számítógépes nyelvészet statisztikai módszereivel történő elemzése), vagy valamely (iv) a hétköznapi életben is elterjedt alkalmazás (lexikográfia, nyelv- és/vagy fordításoktatás és -tanulás stb.) áll.

2.1. Az elektronikusan elérhető korpuszok

Mivel a korpuszépítéssel és a nyelvészeti elemzés céljából történő saját korpusz összeállításával bővebben a 3. pontban foglalkozunk, most a nyelvleírás és a fordítástudományi vizsgálatok számára jelenleg is elérhető online és offline korpuszokat tekintjük át. Vizsgálhatunk írott- vagy beszélt nyelvi

korpuszokat, mely utóbbinál mindenképpen szükséges a beszélt nyelvi anyag átírása az élőbeszéd jellemzőinek (szünet, hangerősség, hanglejtés, hangsúly stb.) rögzítésével. A természetesnyelv-feldolgozás (Natural Language Processing vagy NLP) utóbbi évtizedben tapasztalható robbanásszerű fejlődésének köszönhetően az elemzetlen korpuszok mellett egyre több nyelven válnak elérhetővé az automatikus vagy félautomatikus eszközök segítségével elemzettek. (i) A szegmentálás, tokenizálás során azonosítják a bekezdés-, mondat-, és szóhatárokat, (ii) a lemmatizáló program az azonos szótőből származó alakokat szűri ki, (iii) a lexikai elemek meghatározásához nélkülözhetetlen a morfológiai elemzés, jelentésegyértelműsítés (word sense disambiguation) és tulajdonnév-felismerés, (iv) míg a mondat szerkezeti egységeinek azonosítását és a közöttük lévő összefüggések feltárását a szintaktikai elemzés (parsing) valósítja meg.

Az online és offline elérhető korpuszokat a bennük található szövegek nyelve alapján szokás csoportosítani. A világ első elektronikus korpusza az 1964-ben W. N. Francis és H. Kucera által *Brown University Standard Sample of Present-Day American English* néven publikált *Brown Corpus* volt (http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM). Ezt számos **egynyelv**ű **korpusz** (monolingual corpora) követte, melyek közül a teljesség igénye nélkül említünk néhányat: *British National Corpus* (www.natcorp.ox.ac.uk), *Corpus of Contemporary American English* (www.americancorpus.org), *TIGER Project* (www.ims.uni-stuttgart.de/projekte/TIGER/), *Magyar Nemzeti Szövegtár* (http://corpus.nytud.hu/mnsz/), *Szeged Treebank 2.0* (www.inf.u-szeged.hu/projectdirs/hlt/).

A **többnyelvű korpuszok** (multilingual corpora) két vagy több különböző nyelven előállított egynyelvű korpuszt tartalmaznak, azaz csak autentikus (nem fordított) szövegek gyűjteményei. Ilyen többnyelvű korpusz például a *MLCC Multilingual and Parallel Corpora* (www.elda.org/catalogue/en/text/ W0023.html).

A **párhuzamos korpuszok** (parallel corpora) ezzel szemben az autentikus szövegek mellett ezek fordítását is tartalmazzák. A napjainkban már több nyelvpárra elkészített párhuzamos korpuszok lehetnek egy- vagy kétirányúak. A *Parallel Corpora in Uppsala* például csak eredeti svéd szövegekből és azok fordításaiból áll (http://xml.coverpages.org/etap-over.html), míg az *English–Norwegian Parallel Corpus* tartalmaz angol és norvég ere-

deti szövegeket is, és azok fordítását a másik nyelven (http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/). A fordítástudomány számára különösen érdekesek a **mondatillesztéssel párhuzamosított** korpuszok, mint az *Aligned Hansards of the 36th Parliament of Canada* (http://www.isi.edu/natural-language/download/hansard/), és a *European Parliament Proceedings* 1996–2001 (http://www.statmt.org/europarl/).

Az **összehasonlító korpuszok** (comparable corpora) valamely nyelv autentikus szövegeinek több nyelvre való fordítását is tartalmazzák, mint például az *International Corpus of English* (www.ucl.ac.uk/english-usage/ice/). A **fordítási korpuszok** (translational corpus) ezzel szemben a különböző nyelvekről csak az adott nyelvre fordított szövegek gyűjteményei, mint a *The Translational English Corpus (TEC)* (www.ucl.ac.uk/english-usage/ice/).

2.2. A fordításkutatás számára fontos korpuszok

A fordítási korpuszok különösen alkalmasak a fordított szövegek sajátosságainak vizsgálatára. Az egynyelvű korpuszokkal történő egybevetésük eredményeként magyarázatot kaphatunk a fordításízűség szövegszinten kimutatható okaira, illetve feltárhatjuk az adott nyelvpárra és fordítási irányra jellemző fordítási mintázatot.

Speciális használati lehetőségük miatt a párhuzamos és az összehasonlító korpuszok fontosak még a fordítástudomány számára, mivel alkalmasak a fordítási szövegminták tanulmányozására (Johansson 2003) és a fordítási univerzálék feltárására (Baker 1995). Ezek a korpuszok kulcsfontosságúak továbbá a nem angol korpusznyelvészet (ázsiai nyelvek, kis nyelvek) számára is.

McEnery and Xiao (2007) a párhuzamos és az összehasonlító korpuszok előnyeként kiemeli, hogy (i) általuk olyan új szempontokra figyel fel a komparatív nyelvészet, melyek nem merülnek föl az egynyelvű korpuszok vizsgálatakor; (ii) kibővítik az egyes nyelvek kulturális és nyelvi jellemzőinek különbségeire vonatkozó ismereteinket; és (iii) megvilágítják a célnyelvi és forrásnyelvi szövegek, illetve az autentikus és fordított szövegek közötti különbségeket. E korpuszok alkalmazása a fordítástudományon kívül számos területen hasznos, így a lexikográfiában, nyelvoktatásban stb.

3. Saját korpusz építése

A fordításkutatók körében egyre népszerűbb az egyetemi kutatóhelyek által létrehozott, és elektronikusan elérhető korpuszok helyett a saját kutatási célokat jobban kiszolgáló, egyénileg épített korpuszok vizsgálata. A korpuszépítés maga több egymásra épülő részfeladatra bontható, kezdve a tervezés és szöveggyűjtés fázisával, amit a tárolás követ, majd végül az annotálás, azaz a nyelvészeti elemzés eredményeit tartalmazó többletinformáció feltüntetése.

A korpusztervezés kapcsán két alapvető szempont merül fel: a reprezentativitás igénye és az osztályozás, ami a korpuszok összehasonlíthatóságát biztosítja. Mivel a korpusz tervezőjének elsődleges célja, hogy garantálja kutatási eredményeinek érvényességét és megbízhatóságát (vö. Kennedy 1998), az eredmények általánosíthatósága érdekében ügyelnie kell a korpusz reprezentativitására. A reprezentatív korpusz a mindennapi nyelvhasználatnak megfelelő arányban tartalmazza a szövegtípus, a műfaj, a szövegfunkció, a szövegalkotó neme, a térbeli és időbeli lefedettség stb. szempontjából különböző jellegű szövegeket, amiért a szakirodalom kiegyensúlyozott (well-balanced) korpusznak is hívja. A korpusz reprezentativitása összefügg a korpusz méretével és a mintavétel meghatározásával is.

A korpusz méretét a szövegszavak számával szokás jelölni. Jelenleg az egyik legnagyobb online kereshető korpusz az 1990 és 2009 között a Brigham Young Universityn Mark Davis vezetésével több mint 400 millió szövegszavassá bővített *Corpus of Contemporary American English* (www.americancorpus.org). Bár a számítástechnika tárolási kapacitásának ugrásszerű növekedése következtében egymás után születnek a több millió szavas korpuszok, a kutatási céltól és a lekérdezéshez rendelkezésre álló eszközöktől függően sok esetben előnyösebb a hipotézisek vizsgálatát kisebb korpuszokon megkezdeni.

4. A korpusztervezés fő szempontjai

A korpusz **méret**ének meghatározása mellett a tervezésénél dönteni kell arról, hogy a korpusz a teljes vagy a részszövegek gyűjteményét tartalmazza-e, illetve hogy statikus vagy dinamikus korpuszt (monitor korpusz) kívánunk-e létrehozni (vö. Kennedy 1998). Míg a teljes szövegekből felépített korpusz esetében nehezebb megszerezni a szerzői jogi hozzájárulást, addig a részszövegek gyűjtése azonos korpuszméret mellett több szerző szövegrészletének tárolását teszi lehetővé. A szerzői jogi hozzájárulás szükséges mind

az elektronikus formában történő tároláshoz, mind a szövegek későbbi kutatásban való felhasználásához. A szövegek letöltését ma már megkönnyíti számos elektronikusan elérhető adatbázis, mint a *Digitális Irodalmi Akadémia*, *Magyar Elektronikus Könyvtár*, *The database of translations of Hungarian literary works* (http://translations.bookfinder.hu/indexn.htm), *The Online Books Page* (University of Pennsylvania), *Electronic Text Collections in Western European Literature*, *La biblioteca telematica*, *Literatur im Netz* stb.

A statikus korpusz változatlan formában tartalmazza a meghatározott szempontok alapján válogatott szövegeket, míg a dinamikus korpusz a belső arányok megtartása mellett folyamatosan bővül, ezért alkalmas a nyelv változásának nyomon követésére.

A **szövegválasztáskor** több szempontot kell a korpusz összeállítójának figyelembe vennie aszerint, hogy kutatása során milyen nyelvű, autentikus vagy fordított szövegeket kíván vizsgálni, és a vizsgált szövegekkel milyen időperiódust szeretne felölelni. A kiválasztott szövegeknek illeszkedniük kell a vizsgálat tárgyához (a forrásnyelvi vagy a célnyelvi szöveg vizsgálata, illetve összehasonlító elemzés) és az eredmények tervezett felhasználásához (leíró fordítástudomány, oktatás, szótárkészítés stb.). A szövegválasztást befolyásolják a szöveg csak elemzéssel meghatározható belső tulajdonságai és a szövegalkotás jellemzőiként számon tartott külső tulajdonságok (nyelvi összetétel, szinkrón/diakrón időperiódus, a szerző/ terület/ közzététel helye/ befogadó stb. szerinti tematikus szempontok, a szövegtípus, valamint a szövegek beszélt vagy írott nyelvi volta).

5. Egyszerű korpuszalapú elemzések

A korpuszok elemzéséhez használt számítógépes programokra az információlekérdezés pontossága, gyorsasága és jó minősége jellemző (Szirmai 2005). A ma rendelkezésre álló eszközökkel házilagosan is könnyen lekérdezhető statisztikák, kulcsszókeresések, konkordanciák segítenek a korpusz lexikális jellemzőinek feltérképezésében, ami további kutatási hipotézisekre inspirálhat.

A **szószám** (token) a szóközzel határolt szavakat mutatja, azaz az összes előforduló szövegszót, függetlenül attól, hogy ugyanaz a szó hányszor szerepel az egész korpuszban vagy az egyes szövegegységekben. A **szóalak/típus** (type) a duplikátumok kiszűrése után a korpuszban szereplő kü-

lönböző szótári szavak számáról tájékoztat. A **type/token arány** a korpuszra jellemző lexikai változatosságról ad számot. Mivel a szöveghosszal nő az ismétlések száma, az összehasonlíthatóság érdekében standardizálták a type/token arány számítását, és a szöveg ezer tokenes egységeiben mért eredmények átlagát vetítik a teljes szövegre. Készíthetünk továbbá kis-és nagybetűs **betűgyakorisági** listát, alfabetikus elrendezésű vagy gyakoriság szerinti **szólistát**, vagy vizsgálhatjuk az **átlagos szó- és mondathosszt**. Ez utóbbi a mondathatár-felismerés miatt nehezen automatizálható, és így utólagos kézi javítást igényel.

A statisztikák készítésénél tapasztalható **gondok** közül meg kell még említeni az egymástól független jelentésű, de azonos hangalakú, homonim szavakat (pl.: *Az ország egyik leglátogatottabb műemléke az egri **vár**. / Peti Katira **vár**.*), a ragozott alakok szótári tőre való visszavezetésének, lemmatizálásának problematikáját (pl.: *alma, almát, almás*), a többszavas kifejezések egy szótári egységként való kezelésének nehézségét (pl.: *házi feladat*), az azonos szó különböző írásmódjainak egységes kezelését (pl.: *szőlő – szöllö*), és a kategóriakülönbségek felismerését, mint például a kis- és nagybetűs írásmód hatására módosuló jelentést (pl. *Attila* = alkalmi öltöny, *Attila* = ffi név).

A **kulcsszókereső** program a vizsgált szöveg szólistáját hasonlítja egy hosszabb szöveg szólistájához, és a rövidebb listában gyakrabban előforduló szavakat jelöli meg kulcsszavakként. A **klaszter** vagy **N-gram** a szövegben szereplő több egységből álló szerkezetek vizsgálatát segíti elő. Az *n* helyére kerülő szám határozza meg, hogy a szövegben szereplő szavakat hányas egységekre bontjuk. Például a *Pisti minden délben eszik levest.* mondatot 3-gram esetében a következő egységekre bonthatjuk: *Pisti minden délben*, *minden délben eszik*, és *délben eszik levest.* A **konkordancia (concordance)** az „adott szó vagy kifejezés szövegben szereplő összes előfordulását szövegkörnyezetében bemutató lista” (Szirmai 2005). Az eredményeket a csomóponttól jobbra vagy balra levő első/második.../ötödik szó alapján listáztathatjuk ki a programmal. Sinclair a konkordanciára mint a korpusznyelvészet egyik legfontosabb vizsgálati eszközére tekint, mivel a szöveg mintázataira hívja fel a kutatók figyelmét (vö. Sinclair 1991).

6. Összegzés

A tanulmány megpróbált áttekintést nyújtani a számítógéppel támogatott, illetve a korpusznyelvészeti eszközökkel végzett fordítástudományi vizsgálati lehetőségekről. Mivel ma már korszerű technológia segíti a nyelvészek munkáját, tanulmányunkkal népszerűsíteni is kívántuk felhasználhatóságát a fordítástudomány kutatói körében.

Irodalom

Baker, M. 1993. Corpus Linguistics and Translation Studies. Implications and Applications. Baker., M., G. Francis, E. Tognini-Bonelli (eds). 1993. *Text and Technology: In Honour of John Sinclair*, 233–250. Amsterdam: Benjamins

Baker, M. 1995. Corpora in Translation Studies. An Overview and Some Suggestions for Future Research. *Target* 7/2. 223–243.

Baker, M. 1996. Corpus-based Translation Studies: The Challenges that Lie Ahead. Harold Somers (ed). 1996. *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam & Philadelphia: John Benjamins.

Laviosa, S. 1997. How Comparable Can 'Comparable Corpora' Be? *Target* 9(2): 289–319.

Laviosa, S. 1998. The Corpus-based Approach: A New Paradigm in Translation Studies. *Meta* 43/4. 474–479.

Laviosa, S. 2002. *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam and Atlanta: Rodopi.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*, Harlow: Addison Wesley Longman Ltd.

Kenny, D. 2001. *Lexis and Creativity in Tarnslation. A Corpus-based Study.* Mancsester: St. Jerome.

McEnery, A., Xiao, Z. 2007. Parallel and comparable corpora: What is happening? M. Rogers and G. Anderman (ed.). 2007. *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Multilingual Matters.

Olohan, M. 2004. *Introducing Corpora in Translation Studies*. Routlegde.

Szirmai M. 2005. *Bevezetés a korpusznyelvészetbe. A korpusznyelvészet alkalmazása az anyanyelv és az idegen nyelv tanulásában és tanításában.* Budapest: Tinta Könyvkiadó.

Sinclair, J. 1991. *Corpus, concordance, collocation.* Oxford: OUP.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Teubert. W., Cermáková, A. 2007. *Corpus Linguistics: A Short Introduction.* Continuum.

Tymoczko, M. 1998. Computerized Corpora and the Future of Translation Studies. *Meta* 43/4. 652–660.