FORDÍTÁSTUDOMÁNY MA ÉS HOLNAP

Szerkesztette
ROBIN EDINA
ZACHAR VIKTOR

L'Harmattan

Szakmai lektor: Klaudy Kinga

ISBN 978-963-414-403-8

Melyek a (szak)fordító és a fordításkutató munkáját segítő legfontosabb nyelvi korpuszok?

Seidl-Péch Olívia
E-mail: olivia@inyk.bme.hu

Kivonat: A számítástechnika fejlődésének köszönhetően napjainkra a nyelvi korpuszok a (szak)fordítók és fordításkutatók munkájának fontos segédeszközévé váltak. Az utóbbi évtizedekben a fordításoktatás is egyre nagyobb hangsúlyt helyezett a korpuszalapú megközelítésre. A korpuszok segítségével rákereshetünk egyes szavak és/vagy mondatrészek szövegkörnyezetben való előfordulására, információkat kaphatunk a kívánt kifejezés szinonimáiról, jellemző szókapcsolatairól. A korpuszok használata elengedhetetlen a (szak)fordítások terminológiai konzisztenciáját biztosító adatbázisok létrehozásához, valamint a (szak)fordító maga is párhuzamos korpuszokat hoz létre, ahogy a fordítási környezetben fordítási memóriát használ. A fordítástudomány számára a korpuszalapú megközelítés lehetővé tette a célnyelvi szövegek szövegépítő sajátosságainak új vizsgálati szempontok alapján történő leírását, illetve a deskriptív kutatási kérdések alapjául szolgáló objektív szövegelemzési adatok kinyerését. A cikk kitér a (szak)fordítók és a fordításkutatók számára meghatározó korpuszfajtákra, valamint bemutatja a saját kutatási céloknak megfelelő korpusz összeállításának legfontosabb kritériumait.

Kulcsszavak: fordításkutatás, fordításoktatás, nyelvi korpuszok, (szak)fordítás, szövegelemzés

1. Bevezetés

A fordítóipar számára mára már nemcsak a fordítandó szövegmennyiség ugrásszerű növekedése és a fordítására szánt idő folyamatos csökkenése jelent állandó kihívást, hanem számolni kell a technikai fejlődésnek köszönhető új ismeretekkel és eszközökkel, illetve az egyre specializálódó szakismeretek leírását szolgáló új kifejezések és szókapcsolatok megjelenésével is. További nehézséget jelent az eddig nem megszokott nyelvpárok közötti fordítás igényének térnyerése (pl. magyar–ír), illetve az úgynevezett kis nyelvek (pl. magyar–észt) közötti szótárok hiánya. A fordítóipar fejlődését szemlélve viszont megállapíthatjuk, hogy a XXI. század a (szak)fordítással kapcsolatos számos területen kínál markánsan új lehetőségeket. Ezek hátterében többnyire a számítástechnika és az internetes technológia robbanásszerű fejlődése áll.

A fordítói munkát igen előnyösen támogatják az egyre felhasználóbarátabb *fordítási környezet*et biztosító szoftverek, valamint az egyre szélesebb rétegek számára elérhető fordítástámogató eszközök. Ezek közül manapság szinte minden (szak)fordító gyakrabban használja a hagyományos szótáraknál az egy-, két, vagy többnyelvű *glosszárium*okat, illetve a közvetlen felhasználású *elektronikus szótár*akat (vagyis az olyan szótárakat, ahol a felhasználó maga kereshet) (vö. Seidl-Péch és Pálinkás 2015). Ugyanakkor az is általánosan megfigyelhető, hogy a (szak)fordítók folyamatosan bővítik saját témaspecifikus *terminológiai adatbázis*aikat és *fordítómemóriá*jukat. Utóbbi a forrásnyelvi és a célnyelvi szöveg egymásnak megfeleltetett fordítási egységeit, az úgynevezett szegmenspárokat tartalmazza. A fordítómemóriát a fordító egyrészt a fordítás folyamata során generálja, másrészt tevékenysége hatékonyságának fokozása érdekében szövegpárhuzamosítással is tovább tudja bővíteni. A nem hagyományos fordítástámogató eszközök között említést kell tenni az *előfordító rendszer*ekről is, amelyek a forrásnyelvi szöveget szótárazzák ki, valamint listázzák az ismeretlen szavakat. Az előfordító rendszerek különösen a szakfordítások esetében töltenek be egyre nagyobb szerepet, mivel komoly idő- és pénzmegtakarítást tesznek lehetővé.

A fordítói szakma gyakorlatát segítő mindezen kétségtelenül pozitív változások mellett ugyanakkor figyelmet érdemel az a tény is, hogy a fordítástámogató nyelvi eszközök között említendő nyelvi korpuszok jelentősége, illetve haszna sok fordító és fordításkutató számára még ma sem teljesen egyértelmű. Mindez annál is inkább meglepő, mivel sokszor éppen ezek a nyelvi korpuszok biztosítják az eddig felsorolt alkalmazások háttértámogatását, illetve egyre több annotált nyelvi korpusz válik az internet segítségével széles rétegek számára is elérhetővé. Ugyanakkor kutatók és kutatócsoportok is egyre gyakrabban élnek azzal a technikai fejlődésnek (pl. felhő alapú tárolás) köszönhető lehetőséggel, hogy saját kutatási céljuknak megfelelő korpuszt hozzanak létre. Az utóbbi években a Mona Baker nevével fémjelzett korpuszalapú fordítástudományi megközelítés (Baker 1995) egyre növekvő térnyerése figyelhető meg az ELTE BTK Nyelvtudományi Doktori Iskola Fordítástudományi Doktori Programjában fokozatot szerzett hallgatók disszertációinak esetében is (Seidl-Péch 2011, Polcz 2012, Lengyel 2013, Mohácsi-Gorove 2014, Robin 2014, Sato 2014, Somodi 2014, Bozsik 2015, Kovács 2015, Makkos 2015, Nagy 2015, Szijj 2015), valamint ez adott lendületet a tanszéken épített Pannónia Korpusz létrejöttéhez (Robin et al. 2016).

2. A nyelvi korpuszok összeállítása

Ahogyan Lüdeling és Kytö meghatározása is tükrözi, a nyelvi korpuszok összeállításánál számos kritériumot figyelembe kell venni:

> A korpusznyelvészetet gyakran tekintik a nyelvészet egy viszonylag új területének, amely a „mindennapos" nyelvhasználat empirikus tanulmányozásával foglalkozik számítógépek és elektronikus korpuszok segítségével. A „korpusz" elsősorban írott, vagy beszélt nyelvi szövegek gyűjteménye. Ugyanakkor, ha ezt a kifejezést a modern nyelvtudománnyal összefüggésben használjuk, az számos további ismérvre is utal, így többek között az elektronikus olvashatóságra, a mintavételre és reprezentativitásra, a végleges méretre, valamint arra a szándékra, hogy a korpusz az adott nyelvváltozat hiteles referenciája legyen. (Lüdeling és Kytö 2008: 5; saját fordítás)[1]

A korpuszba gyűjtött szövegekre jellemző legfontosabb ismérvek közé tartozik a *természetes nyelvi előfordulás*, mivel a korpusz „egy nyelv adott állapotára vagy változataira jellemző természetes módon keletkezett nyelvi szövegek gyűjteménye" (Sinclair 1991: 171; saját fordítás)[2]. A korpuszban található szövegek tehát már kivétel nélkül mind rendelkezésre állnak a kutatás megkezdése előtt, és nem a nyelvész hozza őket létre, állításainak alátámasztására. Ugyanakkor a nyelvészeti kutatások célkitűzésének megfelelően az egyes korpuszok összeállításának kritériumai eltérhetnek egymástól, mivel a nyelvészeti kutatás céljából létrehozott korpusz „olyan nyelvi anyagok gyűjteménye, amelyet meghatározott nyelvészeti kritériumok alapján válogattak és rendeztek össze, azzal a céllal, hogy [a korpusz] az adott nyelv mintája legyen" (Sinclair 1996: 4; saját fordítás)[3].

A mintavétel alapján tehát az egyes korpuszoknak az adott nyelv vagy nyelvváltozat vertikális és/vagy horizontális rétegződését kell a teljesség igényével megjelenítenie. Ha a mintavétel tekintetében a korpusz elegendően nagy és *kiegyensúlyozott*, akkor az adott nyelv vagy nyelvváltozat *reprezentatív* korpuszáról beszélünk. Ez utóbbi elengedhetetlen feltétele annak, hogy a korpuszban talált adatok alapján levont következtetések biztosítsák a korpuszalapú nyelvészeti kutatások *validitását*, illetve lehetővé tegyék a (szak)fordítás során a legmegfelelőbb célnyelvi lexikai elem és/vagy célnyelvi minta kiválasztását. A nyelvi korpusz reprezentativitásának köszönhetően a korpusz tehát információkat szolgáltat egy-egy kifejezés, illetve szókapcsolat forrás- vagy célnyelvi környezetben történő használatáról, akárcsak az adott műfajra

és/vagy regiszterre jellemző mondatszintű és szövegtípus függő tulajdonságokról (vö. Seidl-Péch 2011).

2.1. Korpuszépítési kritériumok

Amint arra már az előző fejezetben idézett Sinclair gondolat (1991) is utalt, minden korpuszban szereplő szövegnek általános és közös jellemzője, hogy ezek a szövegek a nyelv természetes megjelenései, tehát a nyelvész vagy a korpusz összeállítójának közbeavatkozása nélkül keletkezett *valódi nyelvhasználati minták*, amelyek e tulajdonságuknak köszönhetően jellemzők a nyelv egy adott állapotára vagy változatára.

A korpusztervezés legfontosabb szempontjai a korpuszalapú kutatási eredmények érvényességét és *megbízhatóságát* szolgálják. A korpusz reprezentativitásának és ezáltal a korpuszon futtatott lekérdezések validitásának érdekében a nyelvészeti korpuszokat összeállító kutatóknak több alapvető kritériumot is figyelembe kell venniük a szöveggyűjtés során, mivel a korpuszba gyűjtött „szövegeket valamilyen szempont szerint válogatják és rendezik" (Magyar Nemzeti Szövegtár). A szövegek felvétele a korpuszba tehát nem véletlenszerű jelleggel történik, mivel a nyelvi korpuszok esetében az egyik legfontosabb elvárás a *tervszerű adatgyűjtés*. Ez a gyakorlatban annyit jelent, hogy a korpusz összeállítója az adatgyűjtés megkezdése előtt számba veszi azokat a jellemzőket (a szövegek külső és belső tulajdonságait), amelyek a korpusz reprezentativitása szempontjából meghatározók. Ezen explicit módon megfogalmazott, a szövegek gyűjtése szempontjából releváns nyelvészeti kritériumok vonatkozhatnak a szövegek megjelenésének korára, illetve területére, a szöveget befogadó közösségre vagy a szövegtípusra, valamint fókuszálhatnak egy-egy lexikai, grammatikai, stilisztikai vagy szövegtani jelenség vizsgálatára (pl. *Mazsola – a magyar igei bővítményszerkezet vizsgálata*).

A korpuszban összegyűjtött szövegekkel szemben támasztott általános követelmény továbbá, hogy ezek a szövegek *gépileg olvasható* formában álljanak rendelkezésre, mivel ennek hiányában nem lehetséges a korpusz egyes kifejezéseire, szókapcsolataira vagy szövegmintáira vonatkozó lekérdezések automatikus futtatása. A korpuszok *reprezentativitása* és *kiegyensúlyozottsága* egymással szorosan összefüggő attribútumok, ugyanis a reprezentativitás értelmében a korpusz teljes mértékben lefedi az adott nyelv vagy nyelvváltozat egy adott időszakban jellemző használatát vagy annak az egyes időszakokon átívelő fejlődését. A reprezentativitás ugyanakkor nem valósulhat

meg maradéktalanul a mintavétel kiegyensúlyozása nélkül. Ennek értelmében a korpuszban tárolt szövegek a korpuszra jellemző alapvető ismérvek szempontjából azonos nagyságú (szószámú) alkorpuszokat alkotnak, mint például a *Corpus of Contemporary American English* különböző műfajú írott (szépirodalmi szövegek, folyóirat cikkek, újságcikkek, tudományos cikkek) és beszélt nyelvi szövegeket tartalmazó alkorpuszai. A korpusz reprezentativitása teszi lehetővé a korpuszalapú kutatások esetében az eredmények *általánosíthatóságát*. A korpusz ugyanakkor „nemcsak tárháza a szövegeknek, hanem tartalmazza azok bibliográfiai adatait, bejelöli a szerkezeti egységeket (bekezdés, mondat)" is (Magyar Nemzeti Szövegtár).

2.2. Korpusztervezési szempontok

A korpusztervezés során felmerülő egyik legfontosabb döntés a korpuszba gyűjtött szövegek időbeli keletkezésére vonatkozik, mivel a korpusz reprezentálhatja egy nyelv vagy nyelvváltozat időbeli fejlődését vagy egy adott időszakra jellemző használatát. E dichotómia mentén beszélhetünk *szinkrón* (pl. *Magyar Nemzeti Szövegtár*) és *diakrón* (pl. *Corpus Diacrónico del Español, Corpus del Español, Magyar Történeti Korpusz*) korpuszokról, bár ez utóbbi típus sokkal kevésbé elterjedt napjainkban. Ez a tendencia többek között a következő okokkal magyarázható: (i) a korpuszok létrehozásának magas költségigénye (elsősorban az anyaggyűjtéssel és rendszerezéssel töltött munkaórák száma miatt), (ii) a szinkrón kutatásokat támogató felhasználói érdekek (korpuszalapú lexikográfia, fordítástámogatás) elsődlegessége, valamint (iii) a szövegállomány digitalizáltságának mértéke.

A szövegek időbeli keletkezése mellett a korpusz keletkezésének időintervalluma is meghatározó adat. Így a korpusz a mintavétel módja szerint lehet statikus vagy dinamikus. Míg a *statikus korpusz* (pl. *Brown Corpus*) a létrehozása óta változatlan, addig a *dinamikus korpusz* (pl. *Collins Corpus*) folyamatosan bővül. A *monitor* korpusz a statikus és dinamikus korpusz ötvözése, miszerint a korpusz bővítése az eredeti statikus korpusz felépítésének struktúráját és arányait megtartva folytatódik (pl. *Corpus of Contemporary American English*).

További szövegválasztási szempontot testesít meg a gyűjtött szövegek írott vagy beszélt nyelvi volta. A korpusztervezők e tekintetben jellemzően előnyben részesítik a gyűjtés és feldolgozás szempontjából is egyszerűbbnek ígérkező, illetve keresettebb *írott nyelvi korpuszok* (pl. *DDR-Korpus, HG-1 korpusz, Szeged Treebank*) létrehozását, ugyanakkor az utóbbi évtizedekben

megnövekedett az érdeklődés a *beszélt nyelvi korpuszok* (pl. *Budapesti Szociolingvisztikai Interjú, Bea – Magyar Spontán Beszéd Adatbázis, Corpus de Français Parlé Parisien des années 2000, EUR-ACCOR*) iránt. A beszélt nyelvi korpuszok előtérbe kerülését egyrészt az egyre nagyobb számú automatikus beszédfelismeréssel és -produkcióval (vö. Kiefer Ferenc et al. 2006: 758) foglalkozó alkalmazott nyelvészeti kutatás indokolja, ugyanakkor e korpuszok kutathatóságát lényegesen egyszerűsíti az éppen ilyen kutatások segítségével kifejlesztett és egyre jobb minőségben megvalósítható automatizált transzkripciós munkafázis. Egy korpuszon belül természetesen lehetséges az egyazon nyelvre jellemző írott és beszélt nyelvi szövegek párhuzamos gyűjtése is (pl. *Corpus of Contemporary American English*), ám ebben az esetben mindenképpen ügyelni kell a korpusz kiegyensúlyozottsága érdekében az írott és a beszélt nyelvi alkorpuszok szószámának megközelítően azonos nagyságára, mivel kizárólag ebben az esetben biztosítható a korpusz reprezentativitása az adott nyelv vagy nyelvváltozat teljességére.

Ezen a helyen mindenképpen meg kell említeni az anyaggyűjtés szempontjából radikálisan új szempontot képviselő *multimodális korpuszokat* (pl. *AMI Meeting Corpus, DEFI-REPERE Project, HucomTech Project, SmartKom Project*), amelyek esetében a korpuszban tárolt adatállomány videofelvételekből és ezek transzkripciójából áll. Az ember–gép interakciót vizsgáló kutatások számára meghatározó a multimodális korpuszok megjelenése, amelyek segítségével például lehetőség nyílik a mimika, a gesztusok, a testtartás tanulmányozására (vö. Kipp et al. 2009). A multimodális korpuszok összeállítása két alapvető nehézséget vet fel: (i) egyrészt a beszélt nyelvi korpuszokhoz hasonlóan a korpuszok összeállítására, a transzkripcióra és annotálásra fordítandó relatív nagy időigényt (ez magyarázza egyelőre ezen korpuszok kis számát, illetve a relatív korlátozott mintavételt), (ii) másrészt a nonverbális tartalom kódolására szolgáló annotációs standard hiányát (Abuczki és Ghazaleh 2013: 87). A multimodális korpuszok között külön figyelmet érdemelnek az elsősorban a kétezres évekre jellemző jelnyelvi korpuszok (pl. *American Sign Language Linguistic Research Project Corpus, JelEsély Projekt*), melyek külön jelentőséggel bírnak a jelnyelv egységesen elfogadott írásbeliségének hiánya miatt. A jelnyelvi korpuszok annotátorai számára további nehézséget jelent a korpuszok szegmentálása (a jelnyelvi elem kezdetének és végének pontos meghatározása), a jelnyelvi produkció standardizálatlansága (országonként, közösségenként változó kézformák), valamint a hangzó nyelvre való fordítás (a jelnyelv sajátos mondatszerkesztési és jelentésalkotási stratégiáinak következtében) (vö. Bartha et al. 2016).

A jelnyelvi kommunikáció standardizációjának elősegítése érdekében a hamburgi *DGS-Korpus-Projekt* távlati célkitűzése például egy német jelnyelvi szótár létrehozása a projekt 15 éves futamidejének végére.

A korpuszba gyűjtött szövegek tekintetében további fontos ismérvet jelent, hogy a korpusz kizárólag *teljes szövegek*et (pl. *Magyar Generatív Történeti Szintaxis*) vagy kizárólag *szövegrészletek*et (pl. *Hunglish Corpus*) tartalmaz-e. Ez utóbbi esetben további fontos kritérium a szövegrészletek hosszának meghatározása is – úgy mint mondat/bekezdés/fejezet/absztrakt stb. (pl. *GENIA corpus*).

Egy adott nyelven belül a szövegek létrejöttének körülménye szerint megkülönböztetjük továbbá a csak eredeti forrásnyelvi szövegekből álló *autentikus nyelvi korpuszokat* (pl. *British National Corpus, El Corpus de Referencia del Español Actual, Magyar nyelvű néprajzi keresőrendszer, TIGER Corpus*), vagy csak az adott nyelvre fordított célnyelvi szövegeket tartalmazó *fordítási korpuszokat* (pl. *Translational English Corpus),* illetve a nyelvtanulóktól gyűjtött nyelvi adatokat tartalmazó *tanulói korpuszokat* (pl. *Arabic Learner Corpus, British Academic Written English, Corpus Écrit de Français Langue Étrangère, Spanish Learner Language Oral Corpora)*.

2.3. Korpuszstruktúra

A nyelvészeti korpuszok alapvetően három különböző típusba sorolható adatot, információt tartalmaznak: primér adatokat, metaadatokat és a nyelvészeti annotációt. A *primér adatok*on nem csupán a korpuszban található szöveg- és hangfájlok, digitalizált képi adatok, videofelvételek gyűjteményét értjük, hanem (amennyiben szükséges) ezeknek az automatikus gépi kereshetőség érdekében történő transzkripcióját is. A *metaadatok* tartalmazzák a primér adatokra vonatkozó összes kiegészítő információt, így a címet, a szerzőt, a keletkezés évét, a kiadót, a nyelvet, a témakört, a műfajt stb. A metaadatok rögzítését különböző standardok alapján végzik, amelyek közül az írott nyelvi korpuszok esetében az egyik legelterjedtebb a TEI (Text Encoding Initiative) alapú adatrögzítés.

A *nyelvészeti annotáció* szempontjából megkülönböztetünk *elemzetlen* és *elemzett* korpuszokat. Az utóbbiak esetében az annotáció tartalmazza a szövegekre vonatkozó szószintű, szintaktikai, szemantikai és pragmatikai információkat. A nyelvészeti elemzés számára szükséges információkat az úgynevezett jelölő (*mark-up*) nyelvek (HTML–Hyper Text Markup Language, SGML–Standard Generalized Markup Language,

XML–Extensible Markup Language) segítségével kódolják platformfüggetlenül. A jelölőelemek (*tagek*) igen szigorú szintaxisát szintén a TEI ajánlásai szabályozzák.

A nyelvészeti annotáció a dokumentumok tartalmi elemeinek megjelölésére szolgál a szövegbe ágyazott metaadatok segítségével. A feldolgozásra szánt szövegek strukturált kódolása történhet kézzel vagy automatikusan. Az annotálást ma már többnyire nyelvészeti elemző programok (pl. *Magyarlánc*) végzik, majd a pontosság érdekében esetleg utólagos humán beavatkozással egyértelműsítik. Az elemzés során megtörténik a bekezdés, mondat és szó szintű szegmentálás, amit a szemantikai kapcsolatok jelölése (*parsing*), valamint a pragmatikai információk azonosítása (*turn-taking*) egészít ki. A szó szintű elemzéshez elengedhetetlen a lemmatizálás (azonos szótőből származó alakok meghatározása), a morfológiai elemzés (a szótő és a toldalékok azonosítása), amit a szavak szófaji egyértelműsítése (*part of speech tagging*) követ. Az annotáció megjelenhet a szövegbe ágyazva vagy egy külső fájlban (*standoff*).

3. Korpusztipológia

A korpuszfajták megkülönböztetése történhet egyrészt a (i) *mintavétel módja* szerint, (ii) a korpuszban *összegyűjtött anyag* alapján, illetve a (iii) korpusz *felhasználásának módja* szerint. Ahogy az előző fejezet már utalt rá, a mintavétel módja szerint a korpusz lehet statikus, dinamikus vagy monitor korpusz, a korpuszban összegyűjtött anyag alapján pedig írott és beszélt nyelvi, illetve multimodális korpusz. A korpuszban feldolgozott anyag meghatározza továbbá, hogy a korpusz mely nyelvterületet reprezentálja. Napjainkra az európai nyelvek nemzeti korpuszai mellett (pl. *Albanian National Corpus, Bulgarian National corpus, Czech National Corpus, Corpus del Español del Siglo XXI, KorpusDK, National Corpus of Polish, PAISÀ Project, Romanian Balanced Corpus, Russian National Corpus, Slovak National Corpus, Stockholm Internet Corpus*), megjelentek már a többi kontinens nyelveinek reprezentatív korpuszai is (pl. *Arabic Newswire Part 1, Australian National Corpus, Balanced Corpus of Contemporary Written Japanese, CALLHOME Egyptian Arabic Speech, Corpus of Spontaneous Japanese, Emille Project, Lancaster Corpus of Mandarin Chinese, Quranic Arabic Corpus, Turkish National Corpus, UCLA Written Chinese Corpus*). Itt érdemelnek mindenképpen említést a már kihalt nyelveket reprezentáló korpuszok (pl. *Electronic Text Corpus of Sumerian Literature*), illetve a kihalással fenyegetett nyelvek írott-

és beszélt nyelvi szövegeit és/vagy multimodális fájljait tartalmazó korpuszok (pl. *Archive of the Indigenous Languages of Latin America*), amelyek kimagasló jelentőséggel bírnak e kultúrák nyelvi emlékeinek megőrzésében.

A nemzeti *egynyelvű korpuszok* (*monolingual corpora*) mellett számon tarunk *többnyelvű korpuszokat* (*mutilingual corpora*) is, amelyek kettő (pl. *COMPARA*) vagy több (pl. *Linguistic Corpus of the University of Vigo*) különböző nyelven előállított, autentikus szövegeket tartalmazó egynyelvű alkorpuszokból állnak. A többnyelvű korpuszok esetében az összehasonlíthatóság érdekében mindenképpen törekedni kell arra, hogy a szövegmintákat azonos nyelvészeti kritériumok (megjelenés ideje, műfaj stb.) szerint gyűjtsék.

A korpusz felhasználásának módja szerint szokás különbséget tenni a már említett szinkrón és diakrón korpuszok, illetve a már szintén említett autentikus, fordítási és tanulói korpuszok között. A felhasználás módja szerinti tipizálás alapján megkülönböztetjük továbbá a *párhuzamos korpuszokat* (*parallel corpora*), amelyek az autentikus szövegeket és azok fordítását tartalmazzák. A párhuzamos korpusz lehet egy-, két- vagy többirányú (pl. *English–Norwegian Parallel Corpus, QCRI AMARA Corpus*), vagy akár mondatillesztéssel párhuzamosított (pl. *Aligned Hansards of the 36th Parliament of Canada, European Parliament Proceedings 1996-2001*). Az egyirányú párhuzamos korpusz a korpuszban található kétnyelvű szövegek esetében csak az egyik irányban tartalmaz fordított szövegeket, míg a kétirányú párhuzamos korpusz mindkét nyelv esetében tartalmaz a másik nyelvről az adott nyelvre fordított szövegeket.

A párhuzamos korpuszok kutatása elsődlegesen a fordítástudomány számára meghatározó, mivel megvilágítják az adott nyelven keletkezett autentikus és fordított szövegek közötti különbségeket, segítve a fordítási univerzálék kutatását (vö. Baker 1995), valamint a célnyelvi szövegalkotásra jellemző fordítási szövegminták kiszűrését (vö. Johansson 2003). A szerver alapú fordítási szolgáltatás előtérbe kerülésének köszönhetően mára már a fordítási piac szereplői is jelentős párhuzamos szövegkorpuszokkal rendelkeznek, amelyek piaci értéke elsősorban a fordítások automatizálása (előfordítás, fordító memória találatok) szempontjából képvisel piaci értéket.

A nyelvészeti kutatások számára meghatározók még az *összehasonlítható korpusz*ok (*comparable corpora*), amelyek meghatározása körül kisebb terminológiai bizonytalanság figyelhető meg. A szakirodalom részben az olyan korpuszokat tekinti összehasonlíthatónak, amelyek több különböző nyelven tartalmaznak azonos kritériumok szerint (szövegtípus, témakör, keletkezés

ideje stb.) gyűjtött autentikus szövegeket. A fordítástudomány (Baker 1995: 23) ugyanakkor azokat a fordítási korpuszokat nevezi összehasonlító korpuszoknak, amelyek egy adott nyelven tartalmaznak autentikus forrásnyelvi szövegeket és egy vagy több másik nyelvről az adott nyelvre fordított célnyelvi szövegeket (pl. *International Corpus of English*). Ezek a korpuszok csak a célnyelvi szövegminták gyűjteményei, és nem tartalmazzák a párhuzamos korpuszokhoz hasonlóan azok forrásnyelvi megfelelőit.

4. A korpuszok használata

A korpuszok alkalmazása napjainkban igen széles lehetőségeket kínál, az egyes korpuszok alkalmazhatóságának alapvetően a korpusz összeállításának szempontjai szabnak határt, azaz egy adott korpusz többféle kutatási terület számára is reprezentatív lehet. Segítségükkel lehetséges egy adott nyelv esetében a lexikon és/vagy a nyelvtan leírása, amire többek között a számítógépes nyelvészet (gépi fordítás) tart igényt. A korpuszok hasznosak a nyelvészet számos más területén is (pl. beszédkutatás, diskurzuselemzés, korpusznyelvészet, összehasonlító nyelvészet, szinkrón és diakrón nyelvleírás, szociolingvisztika), a nyelv- és fordítás oktatása, valamint a lexikográfia számára (pl. *Collins Corpus, Magyar Történeti Korpusz*). Terjedelmi keretek miatt jelen írás csak érinteni tudja a fordítástámogatás és fordításkutatás korpuszalapú támogatását, a többi területtel pedig nem tud részletesebben foglalkozni.

A korpuszok segítségével végzett leggyakoribb lekérdezések részben statisztikai jellegűek (pl. betűgyakoriság, szövegben előforduló szavak listája (*token*) és előfordulásuk gyakorisága, szótári szavak száma (*type*), type/token arány, lexikai sűrűség, szó-, mondat- és bekezdéshossz), részben utalnak az egyes lexikai elemek szövegben betöltött szerepére (kulcsszókeresés) vagy a szövegre jellemző több egységből álló szerkezetekre (n-gram, klaszter, korkordancia). Mindezen elemzések közös tulajdonsága, hogy az információ lekérdezése pontos, gyors és jó minőségű, azaz az elemző szoftverek rövid idő alatt megtalálják a korpuszra vagy az egyes szövegekre jellemző összes előfordulást.

4.1. Korpuszok a (szak)fordítói gyakorlatban

A fordítást támogató eszközöknek – annak alapján, hogy igényelnek-e emberi interakciót – három típusát szokás megkülönböztetni: (i) az auto-

matikus gépi fordítást (*fully automatic machine translation*), (ii) az elő- és/vagy utószerkesztéssel támogatott gépi fordítást (*human aided machine translation*) és (iii) az emberi fordítás gépi támogatását (*machine aided human translation*). A fordítás gépi támogatása (*computer-assisted translation, computer-aided translation tools*) kapcsán már a bevezetőben szó volt az elektronikus glosszáriumokról, szótárakról, terminológiai adatbázisokról, a fordítómemóriáról és az előfordító rendszerekről.

Az automatikus gépi fordítórendszerek részben beépített szótárakra és nyelvészeti ismeretekre épülő szabályalapú rendszerek (*rule based machine translation*), részben nagyméretű háttér adatbázisokra támaszkodó korpuszalapú statisztikai rendszerek (*statistics based machine translation*). Hasznosak lehetnek a fordítók számára az interneten elérhető korpuszok egy-egy lexikai elem jellemző előfordulásának, szövegkörnyezetének tanulmányozására is, mivel a korpuszok a szótárakkal ellentétben a lexikai elemeket nem önmagukban, hanem valós szövegkörnyezetükben jelenítik meg.

4.2. Korpuszok a fordításkutatói gyakorlatban

A korpuszalapú kutatások a fordítástudományban – ahogy azt az első fejezetben említett doktori disszertációk is példázzák – hozzájárultak ahhoz, hogy a preskriptív megközelítéssel szakítva egyre inkább deskriptív jellegűvé váljon a tudományág, és a célnyelvi szövegek kutatását helyezze előtérbe. A korpuszalapú kutatások jellemzője, hogy a fordítás folyamatára, illetve eredményére fókuszálnak, és segítségükkel kimutatható, mely lexikai elemeket, nyelvi jelenségeket vagy szövegmintákat részesítik előnyben a fordítók. Mivel e kutatások középpontjában a valós nyelvi produktumok állnak, a kutatók a statisztikailag szignifikáns megoldásokkal foglalkoznak, és le tudják írni ezek műfaj-, célnyelv-, nyelvpárspecifikus és/vagy fordítóra jellemző karakterisztikumait.

A párhuzamos és a többnyelvű korpuszok alkalmasak a fordítóképzésben a rutinos fordítók megoldásainak elemzésére, valamint lehetővé teszik a célnyelvi elemek jellemző lexikai kapcsolódásainak és szerkezeti mintáinak megfigyelését az autentikus nyelvi környezetben. Napjainkban számos újabb terület kerül a kutatás fókuszába, így megjelentek például a fordítómemóriával kapcsolatos kutatások (Yamada 2011), a klinikai (Jiménez-Crespo 2015) vagy az audiovizuális szövegek fordításával (Bańos et al. 2013) foglalkozó kutatások, vagy akár a szocio- és idiolektusok, valamint a dialektusok fordításához (Haddow et al 2013), illetve a genderkutatáshoz kapcsolódó vizsgálatok (Hayeri 2014).

5. Összegzés

Összegzésként elmondhatjuk, hogy bár a korpusznyelvészet viszonylag fiatal kutatási paradigmát kínál, jelentősége a XXI. századi fordítóipar és fordításkutatás számára ugrásszerűen megnövekedett. A korpuszok egyrészt a fordítási tevékenység gépi támogatásában töltenek be fontos szerepet, másrészt új módszerek és kutatási területek felfedezésére inspirálják a fordítástudományt. A korpuszalapú fordításkutatás számára további lehetőséget jelent az egyéni kutatási célokat szolgáló korpuszok összeállítása, illetve az annotálást és lekérdezést végző szoftverek számának növekedése és minőségének javulása.

Végjegyzet

[1] Corpus linguistics today is often understood as being a relatively new approach in linguistics that has to do with the empirical study of "real life" language use with the help of computers and electronic corpora. In the first instance, a "corpus" is simply any collection of written or spoken texts. However, when the term is employed with reference to modern linguistics, it tends to bear a number of connotations, among them machine readable form, sampling and representativeness, finite size, and the idea that a corpus constitutes a standard reference for the language variety it represents. (Lüdeling, A., Kytö, M. 2008: 5)

[2] „a collection of naturally-occurring language text, chosen to characterize a state or a variety of a language" (Sinclair 1991: 171)

[3] „a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" (Sinclair 1996: 4)

Irodalom

Abuczki, Á., Ghazaleh, E. B. 2013. An overview of multimodal corpora, annotation tools and schemes. *Argumentum* 9. évf. Debreceni Egyetemi Kiadó. 86–98.

Baker, M. 1995. Corpora in Translation Studies. An Overview and Suggestions for Future Research. *Target*. Vol. 7. No. 2. 223–245.

Bańos, R., Bruti, S., Zanotti, S. 2013. Corpus linguistics and Audiovisual Translation: in search of an integrated approach. *Perspectives*. Vol. 21. 483-490.

Bartha Cs., Varjasi Sz., Holecz M. 2016. A magyar jelnyelvi korpusz létrehozásának és annotálásának kihívásai. In: Tanács, A., Varga V., Vincze V. (szerk.) *XII. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem. 207–219.

Bozsik Gy. 2015. *A lexikai kohézió és az énekelhetőség vizsgálata operaszövegek fordításában*. Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/bozsikgyongyver/diss.pdf

Haddow, B., Huerta, A. H., Neubarth, F., Trost, H. 2013. Corpus development for machine translation between standard and dialectal varieties. *Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*. Bulgaria. 7–14.

Hayeri, N. 2014. *Does gender affect translation? Analysis of English talks translated to Arabic.* Elérhető: https://repositories.lib.utexas.edu/bitstream/handle/2152/25082/HAYERI-DISSERTATION-2014.pdf?sequence=1&isAllowed=y

HuConTech Project. Theoretical fundamentals of human-computer interaction technologies project (TÁMOP-4.2.2-08/1/2008-0009).

Jiménez-Crespo, M. 2015. Explicitation in Medical Translation: A Corpus Study. *Meta*. Vol. 60. No. 2. 343.

Johansson, C. 2003. *Visioner och verkligheter. Kommunikationen om företagets strategier.* Elérhető: http://uu.diva-portal.org/smash/get/diva2:162504/FULLTEXT01.pdf

Kiefer F. et al. 2006. *Magyar nyelv. Akadémiai kézikönyvek*. Budapest: Akadémiai Kiadó.

Kipp, M., Martin, J.-C., Paggio, P., Heylen, D. (eds) 2009. *Multimodal Corpora. From Models of Natural Interaction to Systems and Applications.* Berlin Heidelberg: Springer Verlag.

Kovács M. 2015. *A frazeológiai univerzálék fordítási aspektusai és üzenetközvetítő szerepe európai uniós kontextusban.* Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/kovacsmarietta/diss.pdf

Lengyel I. 2013. *A fordítási hiba fogalma funkcionális megközelítésben*. Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/lengyelistvan/diss.pdf

Lüdeling, A., Kytö, M. (eds) 2008. *Corpus Linguistics An International Handbook. Volume 1.* Berlin–New York: Walter de Gruyter.

Magyarlánc. Elérhető: http://www.inf.u-szeged.hu/rgai/magyarlanc

Makkos A. 2015. *Összehasonlítható kompetenciák anyanyelvi és fordított szövegekben. Az anyanyelvi fogalmazási kompetencia és a fordítási kompetencia összefüggései egyetemi hallgatók magyar nyelvű, összehasonlítható szövegei alapján*. Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/makkosaniko/diss.pdf

Mohácsi-Gorove A. 2014. *A minőség fogalma a fordítástudományban és a lektorálás mint minőségbiztosítási garancia.* Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/mohacsigoroveanna/diss.pdf

Nagy J. 2015. *A kommunikatív dinamizmus (relatív) egyensúlya a fordításban.* Doktori értekezés. Kézirat. Elérhető: https://edit.elte.hu/xmlui/bitstream/handle/10831/32440/dissz_Nagy_J%E1nos_nyelvtud.pdf;jsessionid=6A4C5AC4617F2FA25A75A5D97F1D15C5?sequence=1

Polcz K. 2012. *Konvencionálisan indirekt beszédaktusok az angol–magyar filmfordításban.* Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/polczkaroly/diss.pdf

Robin E. 2014. *Fordítási univerzálék a lektorált szövegekben*. Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/robinedina/diss.pdf

Robin E. et al. 2016. Fordítástudomány és korpuszkutatás: bemutatkozik a Pannonia Corpus. *Fordítástudomány* 18. évf. 2. sz. 5–26.

Sato, N. 2014. *A vállalati és üzleti tolmács kettős lojalitása a magyar–japán és a japán–magyar interperszonális kommunikációban.* Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/satonoriko/diss.pdf

Seidl-Péch O. 2011. *Fordított szövegek számítógépes összevetése.* Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/seidlpecholivia/diss.pdf

Seidl-Péch O., Pálinkás M. 2015. Lépést tud-e tartani a műszaki szaklexikográfia a technikai fejlődéssel? In: Bocz Zs. (szerk.) *Porta Lingua – 2015. A XXI. századi szakmai, szaknyelvi kommunikáció kihívásai: tanári és tanulói kompetenciák. Cikkek, tanulmányok a hazai szaknyelvoktatásról és -kutatásról.* Budapest: *SZOKOE*. 145–157.

Sinclair, J. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, J. 1996. *EAGLES preliminary recommendations on corpus typology.* EATCWG-CTYP/P. Pisa: ILC-CNR.
Elérhető: http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

Szijj M. 2015. *Műfordítás nem anyanyelvre. Magyarok a magyar irodalom spanyol fordításában.* Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/szijjmaria/diss.pdf

Somodi J. 2014. *Megszólítások pragmatikája japán-magyar összevetésben. A japán apellatív megszólítások fordításának vizsgálata magyar filmszövegekben.* Doktori értekezés. Kézirat. Elérhető: http://doktori.btk.elte.hu/lingv/somodijulia/diss.pdf

The Text Encoding Initiative. Elérhető: http://www.tei-c.org/index.xml

Yamada, M. 2011. The effect of translation memory databases on productivity. In: Pym, A. (ed.) *Translation research projects 3.* Tarragona: Universitat Rovira i Virgili.

Hivatkozott magyar nyelvű korpuszok

Budapesti Szociolingvisztikai Interjú (BUSZI)	http://www.nytud.hu/buszi/
Bea – Magyar Spontán Beszéd Adatbázis	http://nytud.hu/adatb/bea
HG-1 korpusz	http://corpus.hungram.unideb.hu/
Hunglish Korpusz	http://szotar.mokk.bme.hu/hunglish/search/corpus
JelEsély Projekt	http://jelesely.hu/web/?q=hu/node/4
Magyar Generatív Történeti Szintaxis	http://omagyarkorpusz.nytud.hu/hu-intro.html
Magyar Nemzeti Szövegtár	http://mnsz.nytud.hu
Magyar nyelvű néprajzi keresőrendszer	http://maszeker.huminf.u-szeged.hu:8081/maszeker-web/
Magyar Történeti Korpusz	http://www.nytud.hu/hhc/
Mazsola – a magyar igei bővítményszerkezet vizsgálata	http://corpus.nytud.hu/mazsola/
Pannonia Corpus	https://www.facebook.com/PannoniaCorpus/

Szeged Treebank	http://rgai.inf.u-szeged.hu/index.php?lang=en&page=SzegedTreebank

Hivatkozott nem magyar nyelvű korpuszok

Alcohol Language Corpus - ALC	https://www.phonetik.uni-muenchen.de/Bas/BasALCeng.html
Albanian National Corpus	http://web-corpora.net/AlbanianCorpus/search/
Aligned Hansards of the 36th Parliament of Canada	http://www.isi.edu/natural-language/download/hansard/
AMI Meeting Corpus	http://groups.inf.ed.ac.uk/ami/corpus/
American Sign Language Linguistic Research Project Corpus	http://www.bu.edu/asllrp/
Arabic Learner Corpus	http://www.arabiclearnercorpus.com
Arabic Newswire Part 1	https://catalog.ldc.upenn.edu/LDC2001T55
Archive of the Indigenous Languages of Latin America	http://www.ailla.utexas.org/site/welcome.html
Australian National Corpus	https://www.ausnc.org.au
Balanced Corpus of Contemporary Written Japanese	http://pj.ninjal.ac.jp/corpus_center/bccwj/en/
British Academic Written English	http://www2.warwick.ac.uk/fac/soc/al/research/collections/bawe
British National Corpus	http://www.natcorp.ox.ac.uk
Brown Corpus	http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/
Bulgarian National corpus	http://dcl.bas.bg/bulnc/en/
CALLHOME Egyptian Arabic Speech	https://catalog.ldc.upenn.edu/LDC97S45
Czech National Corpus	https://www.korpus.cz
Collins Corpus	https://www.collins.co.uk/page/The+Collins+Corpus
COMPARA	http://www.linguateca.pt/COMPARA/
Corpus de Français Parlé Parisien des années 2000	http://cfpp2000.univ-paris3.fr/Corpus.html
Corpus del Español	http://www.corpusdelespanol.org
Corpus del Español del Siglo XXI	http://web.frl.es/CORPES/view/inicioExterno.view;jsessionid=F004423F8EF3F6E3E4B94A893BD0087C
Corpus Diacrónico del Español	http://corpus.rae.es/cordenet.html

Corpus Écrit de Français Langue Étrangère	http://projekt.ht.lu.se/cefle/
Corpus of Contemporary American English	http://corpus.byu.edu/coca/
Corpus of Spontaneous Japanese	http://pj.ninjal.ac.jp/corpus_center/csj/en/
DEFI-REPERE Project	http://www.defi-repere.fr/index.php?id=7
DDR-Korpus	https://www.linguistik.hu-berlin.de/de/forschung/abgeschlossene-projekte/ddr-korpus
DGS-Korpus-Projekt	http://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html
El Corpus de Referencia del Español Actual	http://www.rae.es/recursos/banco-de-datos/crea
Electronic Text Corpus of Sumerian Literature	http://etcsl.orinst.ox.ac.uk
Emille Project	http://www.emille.lancs.ac.uk
English-Norwegian Parallel Corpus	http://www.helsinki.fi/varieng/CoRD/corpora/ENPC/
EUR-ACCOR	http://www.cstr.ed.ac.uk/research/projects/artic/accor.html
European Parliament Proceedings 1996-2001	http://www.statmt.org/europarl/
GENIA corpus	http://www.geniaproject.org
International Corpus of English KorpusDK	http://www.ucl.ac.uk/english-usage/projects/ice.htm http://ordnet.dk/korpusdk_en?set_language=en
Lancaster Corpus of Mandarin Chinese	http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/
Lancaster-Oslo/Bergen Corpus (LOB)	http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html
Linguistic Corpus of the University of Vigo	http://sli.uvigo.es/CLUVI/index_en.html#correo
National Corpus of Polish	http://nkjp.pl/index.php?page=0&lang=1
QCRI AMARA Corpus	http://alt.qcri.org/resources/qedcorpus/
Quranic Arabic Corpus	http://corpus.quran.com

Romanian Balanced Corpus	http://metashare.elda.org/repository/browse/romanian-balanced-corpus-rombac/0a7dd85edc7311e5aa0b00237df3e35873a0d662435d42dd94fba48c29dc0065/
Russian National Corpus	http://www.ruscorpora.ru/en/index.html
SmartKom Project	http://www.smartkom.org
Spanish Learner Language Oral Corpora	http://www.splloc.soton.ac.uk
Slovak National Corpus	http://korpus.juls.savba.sk/index_en.html
Stockholm Internet Corpus	http://www.ling.su.se/english/nlp/corpora-and-resources/sic
TIGER Corpus	http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
Translational English Corpus	http://www.alc.manchester.ac.uk/translation-and-intercultural-studies/research/projects/translational-english-corpus-tec/
Turkish National Corpus	http://www.tnc.org.tr
UCLA Written Chinese Corpus	http://www.lancaster.ac.uk/fass/projects/corpus/UCLA/