Korpuszalapú fordításkutatás: lehetőségek és nehézségek
Fókuszban a korpuszépítés és a korpuszalapú elemzés

Seidl-Péch Olívia

olivia@inyk.bme.hu

Budapesti Műszaki és Gazdaságtudományi Egyetem

Idegen Nyelvi Központ

Kivonat: A fordítástudomány számára releváns kutatási módszerek között már az ezredfordulót megelőzően is hangsúlyosak voltak a korpuszalapú kutatások, amelyek szerepe napjainkra tovább növekedett. A fordítástudomány alkalmazott nyelvészeti beágyazottsága már a kezdetektől kezdve szükségessé tette a módszertani kérdések tisztázását (vö. Károly 2002), melyhez hasonlóan napjainkban a tudományágon belül a korpuszalapú kutatások előtérbe kerülését eredményező kutatási paradigmaváltás és az azt lehetővé tevő technológiai fejlődés miatt válik szükségessé a korpuszalapú módszer jellemzőinek és kritériumainak áttekintése. A hatékony és lekérdezési eredményeit tekintve megbízható kutatások alapvető tulajdonságai között kerülnek előtérbe a kutatási kérdéseknek megfelelő, valamint a vizsgált sokaságot kiegyensúlyozottan reprezentáló minta, illetve a minta adekvát rendezését megvalósító korpuszdesign. A korpuszalapú kutatások elméleti kereteit biztosító módszertani alapelvek (vö. Laviosa 1998a) mellett tehát a mintavételi kritériumok tisztázása válik elsődlegessé, amelynek jelentőségét növeli a könnyen hozzáférhető, nagy adatmennyiségek megjelenése. A kutatások számára tehát hangsúlyozottan elsődleges (a valóban óriási kínálatból) a reprezentatív minta összeállításának szükségessége. A mintavételi szempontrendszer összeállításához, illetve a kutatások tervezéséhez kínál egyfajta összefoglalást a jelen tanulmány.

Kulcsszavak: korpuszépítés, korpuszalapú fordításkutatás, kvantitatív kutatás, korpusztervezés, mintavétel

Seidl-Péch Olívia: Korpuszalapú fordításkutatás: lehetőségek és nehézségek. Fókuszban a korpuszépítés és a korpuszalapú elemzés. In: Robin E., Seidl-Péch O. (szerk.) 2020. *Fókuszban a fordított és a tolmácsolt szöveg: korpuszalapú fordításkutatás Magyarországon.* Segédkönyvek a nyelvi közvetítésről I. Budapest: ELTE BTK Fordítástudományi Doktori Program, MANYE Fordítástudományi Szakosztály. DOI: https://doi.org/10.36252/Nyelvikozvsegedkonyv1.4

1. Bevezetés

A korpuszalapú kutatások megjelenése óta eltelt évtizedek tapasztalatai alapján ma már bátran megállapíthatjuk, hogy a szakterület méltán foglal el kiemelkedően fontos helyet a nyelvészeti kutatások között (McCarthy és O'Keefee 2010). A korpuszok legfontosabb ismérve, miszerint a korpuszban összegyűjtött szövegek a természetes nyelvhasználat gépileg olvasható formában tárolt mintái (vö. Bowker és Pearson 2002: 9), meghatározó jelentőséggel bír az empirikus nyelvészeti kutatások számára.

A kutatók a kutatási célokat szolgáló tudatos gyűjtés és mintavétel során az adott nyelvre vagy nyelvváltozatra jellemző szövegeket rendezik korpuszokba és azok alkorpuszaiba, mindig ügyelve arra, hogy a vizsgálatokat és a lekérdezéseket a lehető legkiegyensúlyozottabb korpuszok segítségével végezzék. Az mára már nyilvánvaló, hogy – a nagyon kevés és csak bizonyos kutatási témákra jellemző kivételtől eltekintve (például egy szerző munkásságának vizsgálata) – a számítástechnikai eszközök megnövekedett kapacitása ellenére sem lehetséges a korpuszban az adott nyelvre vagy nyelvváltozatra jellemző összes nyelvi minta rögzítése. Ennek következtében egyre inkább előtérbe kerül a korpuszok kiegyensúlyozottságának és reprezentativitásának kérdése (Seidl-Péch 2018), mivel csak ezek figyelembevételével zárható ki a kutatások eredményeinek véletlenszerűsége. Másképpen fogalmazva, csak a körültekintően meghatározott kritériumok alapján válogatott mintán végzett korpuszalapú kutatások eredményei tekinthetők tudományosan megalapozottnak. Ellenkező esetben a lekérdezések validitása megkérdőjelezhető akkor is, ha a kutatás esetleg egy több millió szövegszavas korpusz lekérdezésével valósul meg. A korpuszok kiegyensúlyozottsága és reprezentativitása tehát elengedhetetlen feltétele a kutatási eredmények érvényességének és megbízhatóságának. Megállapíthatjuk továbbá, hogy a nyelvi korpusz kiegyensúlyozottsága és reprezentativitása teszi a korpuszt alkalmassá arra, hogy segítségével információkat gyűjtsünk egy-egy kifejezés, illetve szókapcsolat forrás- vagy célnyelvi környezetben történő műfajra és/vagy regiszterre jellemző használatáról, vagy tágabban értelmezve az adott nyelv térben, időben és modalitásban behatárolt változatáról, amely megmutatja a korpusz összetételének megfelelő, műfajra,

szövegtípusra, regiszterre, illetve az autentikus és/vagy célnyelvi nyelvhasználatra jellemző használati mintákat (vö. Seidl-Péch 2018).

2. A korpuszalapú módszer jelentőségének növekedése a fordításkutatásban

Már az 1950-60-as években, a nyelvészeti fordítástudomány megjelenésével (vö. Klaudy 2004) megkezdődött a fordítási folyamat **összes nyelvi és nem nyelvi** változójának függvényében a fordítás eredményének, folyamatának és funkciójának leírása (vö. Holmes 1975). Az irodalmi paradigmával való szakítás következtében megerősödött a fordítástudomány nyelvészeti beágyazottsága, ezzel együtt pedig megjelentek a társadalomtudományi kutatásokat jellemző módszerek és eredmények. Az alkalmazott nyelvészeti kutatások sorába illeszkedő leíró fordítástudományi vizsgálatok között a korpuszalapú paradigma kezdetekben a fordított és a nem fordított (autentikus), illetve a forrásnyelvi és a célnyelvi szövegek nyelvi jellemzőinek kontrasztív leírását részesítette előnyben, mely kutatások kedveztek a fordítási univerzálék megfogalmazásának és a nyelvpárspecifikus fordítói magatartás vizsgálatának (vö. Károly 2003). Klaudy korszakolása alapján (Klaudy 2004, 2005) az 1970-80-as éveket a nyelvészeten belüli (például szövegnyelvészet, pragmatika, számítógépes nyelvészet) és kívüli (például pszichológia, szociológia, irodalomtudomány, filozófia) határtudományokkal való összefogás, míg az 1990-es éveket az európai integráció következtében a fordítások iránt megnövekedett társadalmi igény megjelenése és felerősödése (például az EU nyelvpolitikája, multinacionális vállalatok igényei, lokalizáció) jellemezte.

Az ezredfordulótól kezdve a fordítandó dokumentumok számának robbanásszerű növekedése szükségessé tette a nyelvtechnológiai alkalmazások integrálását a fordítási folyamatba (például gépi adatbázisok, párhuzamos korpuszok, a fordító számítógépes segédeszközei). Napjaink fordítója munkája során lépten nyomon találkozik korpuszokkal (például online korpuszok, gépi fordítás), illetve maga is hoz létre speciális korpuszokat (például terminológiai adatbázis, fordítómemória, párhuzamos dokumentumok). Mindezek hozzájárultak ahhoz, hogy a 2000-es évektől kezdve egy olyan **újabb paradigmavál-**

tás következzen be a fordításkutatásban, amelyet az empirikus alapokon nyugvó korpuszalapú tudományos megközelítés jellemez (Seidl-Péch 2019). A korpuszalapú kutatások előtérbe kerülését a nagy adatállományok tárolásának és kezelésének egyszerűbbé válása tette lehetővé. Az eredményközpontú kutatások fókusza így már igen széles skálán mozoghat a célnyelvi szövegtulajdonságok leírásától a fordítói stílus vizsgálatáig, illetve a fordítóképzés tapasztalataitól a tolmácsoláskutatásig. A korszakra ugyanakkor egyre inkább jellemző a költségigényesebb folyamatközpontú kutatások számának növekedése is (például tolmácsolásnál és fordításnál egér- és/vagy szemmozgáskövető vizsgálat, tolmácsolásnál levegővétel/szünet- vagy elakadásvizsgálat), amelyekre a technológiailag teljesen megújult tolmácsszakmának (vö. Fantinuoli 2018.) óriási szüksége is van. A korpuszalapú kutatások kvantitatív eredményeit ugyanakkor a kutatók mindinkább szükségesnek érzik kvalitatív vizsgálatokkal kiegészíteni (lásd például Robin 2018), amely gyakorlatot egyrészt a korpuszok kiegyensúlyozottságának és reprezentativitásának megkérdőjelezhetősége teszi szükségessé, másrészt annak felismerése, hogy a kvantitatív eredmények sok esetben csak az intuíció megerősítésére elegendőek, de nem adnak valós válaszokat a felmerülő kutatási kérdésekre (Seidl-Péch és Robin 2019).

Az új módszertan a kutatási témakínálat bővülését is elősegítette. Egyes kutatások a fordítási normát kísérelik meg körüljárni, és a fordítási mintázatot (például Balaskó 2005) írják le, továbbá előtérbe került a fordítások szociokulturális jellemzőinek nyelvészeti és interkulturális beágyazottságának vizsgálata is. A kutatások egy másik csoportja az új technológiák mentén végez vizsgálatokat (például audiovizuális fordítás: Bańos et al. 2013; közösségi fordítás: Malaczkov 2020; távtolmácsolás: Castagnoli–Niemants 2018, Devaux 2018, Seresi 2016) , felerősödött a szaknyelvi fókuszú szövegek iránti érdeklődés (például jogi szaknyelv: Vincze 2018; EU-s kontextus: Jablonkai 2009; műszaki szaknyelv: Nagy 2020; orvosi szaknyelv: Mány 2020), illetve általánossá vált az angolon kívüli nyelvek fordításközpontú vizsgálatának integrálása (például ázsiai és arab nyelvek: Bakaja 2020; török: Aksan és Aksan 2018; kis nyelvek: Karakanta et al. 2018).

A korpuszalapú fordításkutatás eredményeinek hasznosítása egyre több szakterület számára kínál releváns tartalmakat. A kutatási eredmények fontosak mind az autentikus,

mind a célnyelvi szövegprodukció leírása, az anyanyelvhasználat tudatosítása, a fordításnyelv sajátosságainak feltárása (Balaskó 2007) és a célnyelvi jellemzők célnyelvi normától eltérő használatának megvilágítása szempontjából(Seidl-Péch 2011). Az eredmények közvetlenül és közvetve is felhasználhatók a fordításoktatásban és a nyelvtechnológiai alkalmazások használatakor (például gépi fordítás, online szótárak).

3. Korpuszépítés

A korpuszalapú kutatások eredményeinek érvényessége és megbízhatósága szempontjából a korpusztervezés kiemelkedő jelentőséggel bír (vö. Robin et al. 2017), mivel a korpusz(ok)ba rendezett szövegek válogatásának szisztematikussága alapján dönthető el, hogy a korpuszban tárolt mintára vonatkozó megállapítások általánosíthatók-e a vizsgálni kívánt szövegtípusra. A mintavétel tekintetében elvárható tehát az egyes kutatásoktól, hogy a vizsgált korpusz(ok)ba rendezett szövegek vagy szövegrészletek az adott nyelv vagy nyelvváltozat vertikális és/vagy horizontális rétegződését a teljesség igényével reprezentálják (vö. Seidl-Péch 2017). Természetesen a mintavétel nem zárja ki a kutatás leszűkítését egy bizonyos modalitás/szövegtípus/korszak/szerző vagy fordító vizsgálatára, ugyanakkor a kutató felelőssége egy olyan megbízható kritériumrendszer kidolgozása, amely alapján a korpuszba válogatott szövegek bekerülési esélye megegyezik az adott korpuszba éppen fel nem vett, de a korpusz felépítése (korpusztervezés) szempontjából szintén beválogatható szövegek bekerülési esélyeivel.

A szisztematikus, de egyidejűleg véletlenszerű mintavétel kritériumának teljesítése igazi kihívás elé állítja a kutatókat, hiszen a minta végessége miatt mindenképpen szükséges a módszertani lépések részletes igazolása. A korpuszalapú kutatások általánosíthatósága tehát nagy mértékben függ attól, hogy a tervezés során a kutatók eleget tesznek-e a minőségbiztosítási kritériumoknak. Az adatgyűjtés folyamán ennek megfelelően a kutatás elméleti kereteinek megfelelő szövegek sokaságából úgy kell kiválasztani a korpuszba felvett elemeket, hogy a szelekció során az összes potenciálisan választható szöveg egyenlő eséllyel szerepeljen a mintában. Erre még akkor is kiemelt figyelmet kell fordítani, ha

másrészről evidenciának tekintjük azt a tényt, hogy a minta sohasem képezheti le teljes mértékben azt a sokaságot, amelyet reprezentálni hivatott (vö. Dörnyei és Csizér 2012). A kutatónak ugyanakkor számolnia kell több olyan, a kutatás tervezését és lefolytatását befolyásoló tényezővel, amelyek korlátozzák a mintavétel véletlenszerűségét, mivel a vizsgált populációt érintő adottságok mindenképpen befolyásolják a mintavételt.

3.1. Mintavételi nehézségek

Amint az előzőekből is kitűnik, a kvantitatív nyelvészeti kutatások a korpuszba összegyűjtött mintára vonatkozó mennyiségileg is feldolgozható információk alapján kívánnak a minta által reprezentált sokaság tulajdonságaira rámutatni. Ennek következtében egy bizonyos nyelv vagy nyelvváltozat adott vertikális és/vagy horizontális rétegződésének megfelelő nyelvi adathalmaz, azaz sokaság vizsgálatának esetében elvárható, hogy a sokaságot reprezentáló korpusz elemzése alapján bemutatott megállapítások a korpuszon kívül is érvényesek maradjanak, azaz ne csak az adott mintát jellemezzék, hanem a minta által reprezentált sokaságot is. Ezen elvárásnak való megfelelés teljesülése jelenti a korpuszalapú kutatások esetében a legnagyobb mintavételi nehézséget.

Ezzel kapcsolatban már a téma tárgyalása elején le kell szögezni, hogy téves az a napjainkban egyre inkább terjedő felfogás, miszerint az egyre nagyobb minta egyre jobb mintavételt eredményez. Kétségtelen tény, hogy a korpuszok reprezentativitása szempontjából fontos szerepet játszik a kutatási kérdések nagy adatmennyiségeken való tesztelése, ugyanakkor nem lehet figyelmen kívül hagyni ezen nyelvi adatok összeválogatásának szempontjait sem. Ez utóbbiak hozhatók összefüggésbe a korpuszok kiegyensúlyozottságával.

Továbbá azt is figyelembe kell venni a korpuszalapú nyelvészeti kutatásoknál, hogy a minta kiválasztásának alapjául szolgáló nyelvi adatok száma a legtöbb vizsgálat esetében végtelen nagyságú, és ennek következtében nem beszélhetünk matematikai módszerekkel pontosan körülírható mintavételi eljárásról, hanem sokkal inkább a kutatási szempontrendszer alapján praktikus elvek mentén összeállított korpuszokról. A kutatónak

meg kell elégednie az ideális mintavétel helyett a kutatási céloknak megfelelő, a vizsgált szempontok alapján rétegzett mintával, amelynek hiányosságaira a mintavétel bemutatásánál mindenképpen reflektálnia kell.

A korpuszalapú kutatások bemutatásának amúgy is kiemelten fontos része a korpuszok összeállításának és a korpuszban tárolt szövegek és/vagy szövegrészletek kiválasztási metódusának leírása. Ennek hiányában az olvasóban számos kétely fogalmazódik meg a minta alapján kapott eredmények **érvényességére** vonatkozóan. Annak ellenére, hogy a korpuszalapú kutatások elsődleges célja a lekérdezési eredmények és a belőlük levonható következtetések tárgyalása, ezek a kutatások nem értelmezhetők az adatgyűjtő eszköz részletes bemutatása nélkül, amely az eredmények érvényességét támasztja alá a mért változók szakirodalmi áttekintéséhez hasonlóan. Az adatgyűjtési megfontolások és lépések részletes és alapos bemutatása biztosítja többek között a kutatás **megismételhetőségét**, illetve a mért eredmények **összevethetőségét**. Ha például a bemutatott mintavétel alapján valaki egy későbbi időpontban megismétli az adott kutatást, akkor az első kutatás eredményeivel megegyező eredmények igazolni tudják az előző kutatás **megbízhatóságát** (vö. Dörnyei és Csizér 2012).

A kutatási eredmények összevetésére akkor kerülhet sor, ha egy következő kutatás az előbbi mintavételére támaszkodva pusztán egyetlen kritérium alapján változtatja meg a minta összetételét. Ilyen lehet például egy újabb nyelvpár vagy egy másik szövegtípus esetében az adott vizsgálat megismétlése. Ugyanakkor igen gyakori problémát okoz, amikor egyes kutatók egy korábbi kutatás mintavételi kritériumai közül egyszerre többet is megváltoztatva kívánnak a korábbi kutatás eredményeire reflektálni, illetve amikor egy kutatáson belül az alkorpuszok összeállítása több tulajdonság esetében sem halad ugyanazon **mintavételi kritériumrendszer** szisztematikus végigvitele mentén. Ilyen esetekben nem állapítható meg teljes bizonyossággal, hogy az eltérések mely változók mentén jöttek létre, és ebből következően nem vonhatók le egyértelmű következtetések.

További bizonytalanságot okozhatnak a mintavétel során tapasztalható belső aránytalanságok, amikor egy-egy szövegtípus, nyelvpár, szerző, téma, műfaj stb. valamilyen okból kifolyólag (például könnyebben vagy nehezebben elérhető szövegek) felül-

vagy alulreprezentált a korpuszban. Ilyen esetben nehezen bizonyítható, hogy a korpusz összetételének esetleges megváltoztatásával nem módosulnának-e a lekérdezések eredményei, ezért a kutatás során feltárt összefüggések sem tekinthetők **érvényesnek**.

3.2. Kvantifikálható szempontrendszer

A kutatás másik lényeges jellemzője a kutatási kérdés(ek) megfogalmazásának szükségessége a kutatás megkezdése előtt, amelyek természetesen nem zárják ki, hogy a kutatás közben újabb és újabb feltárandó kérdések merüljenek fel. Ugyanakkor a vizsgálandó kérdések meghatározása a kutatás elején és ezzel párhuzamosan a feltételezett eredmények hipotézisek formájában való megfogalmazása elengedhetetlen annak számbavételéhez, hogy a tervezett kutatás valóban elvégezhető-e kvantitatív kutatási módszerekkel. Másként fogalmazva a kutatás tervezési szakaszában el kell dönteni, hogy a vizsgálni kívánt kérdés esetében a kvantitatív lekérdezés és az azt lehetővé tevő nagy adatmennyiség gyűjtése a célszerű módszer-e, vagy előnyösebb a kérdés kisebb mintán végzett kvalitatív vizsgálata. Míg a nagyszámú nyelvi minta vizsgálata alapján végzett kvantitatív kutatások többnyire hipotézisek megfogalmazásával, előfeltételezések alapján keresik a választ egy-egy nyelvi minta működésének jellemzőire és gyakoriságára, addig a kisebb mintán végzett kvalitatív kutatások az adott minta működésének ok–okozati összefüggéseit is fel tudják tárni.

A korpuszalapú vizsgálatok igen fontos jellemzője, hogy a vizsgálatok tárgyai a valós nyelvi előfordulások, így a lekérdezések eredményei a tényleges nyelvi produktum (például szövegkutatások) vagy folyamat (például elakadásvizsgálat a tolmácsoláskutatásban) elemzését és feltárását teszik lehetővé. Ugyanakkor a kutatónak a kutatási kérdéseket mindenképpen úgy kell megfogalmaznia, hogy az eredményeket számszerűsíthető adatok formájában tudja lekérdezni és elemezni. A fordítás-/tolmácsoláskutató vizsgálhatja például, hogy a fordítók/tolmácsok mely nyelvi jelenséget (például lexikai elemet, szókapcsolatot) használják gyakrabban vagy ritkábban, mint az anyanyelvi beszélők, vagy éppen

mely elemek használata nem jelenik meg a fordított/tolmácsolt célnyelvi szövegekben a forrásnyelvi stimulus hiányában (például egyedi nyelvi elemek, Dankó 2017).

A korpuszalapú fordítástudományi vizsgálatok esetében végzett leggyakoribb leíró statisztikai lekérdezések a **szövegszavak számát** (az összes szövegszó gyakoriságtól független számát), a korpuszban szereplő különböző **szótári szavak számát** (szótípus) és a **szótípus/szövegszó arányt** (a korpuszra jellemző lexikai változatosságot) vizsgálják (vö. Laviosa 1998b). Jellemzőek továbbá a **betűgyakorisági** listák (az adott nyelvre vagy szerzőre jellemző betűeloszlás), a **gyakoriság szerinti szólisták** (a szöveg(ek)ben gyakrabban és kevésbé gyakran – akár csak egyszer – előforduló szavak), a szöveg feldolgozhatósága és az egyszerűsítés szempontjából meghatározó **átlagos szó- és mondathossz** (a szövegben található összes betű/szó száma elosztva az összes szó/mondat számával) lekérdezései, a **kulcsszavak** szűrése (egy hosszabb szöveg szólistájához viszonyítva a vizsgált szöveg szólistájában gyakrabban előforduló szavak), illetve a klaszterek vagy **N-gramok** elemzése (a szövegben szereplő több egységből álló szerkezetek), ahol az N helyére kerülő szám határozza meg, hogy a szövegben szereplő szavak hány egységes előfordulását vizsgáljuk. Ez utóbbi segíthet például feltárni a terminusjelölteket, illetve a szövegben szereplő formulaszerű elemeket (Nagy 2019). A gyakorisági listák és a kvantitatív elemzések segítségével vizsgált felszíni jelenségek jó alapot kínálnak a mélyebb szerkezeti jellemzők feltárásához, a kvantitatív kutatások eredményei alapján megkezdett kvalitatív kutatások lefolytatásához.

Az egyszerű statisztikai lekérdezéseken túl a korpusz felszíni jegyei további elemzéseket is lehetővé tesznek, amennyiben a korpusz annotált, azaz a nyelvészeti elemzés számára érdekes jelenségek megjelölésére metanyelvi többletinformációt tartalmaz. A kézi vagy gépi annotálás céljára általában az úgynevezett jelölő (Mark-up) nyelveket használják, melyek közül a HTML, SGML, XML (Hyper Text Markup Language, Standard Generalized Markup Language, Extensible Markup Language) a legelterjedtebbek. Az annotálásra használt jelölőelemek (tagek) szabadon bővíthetők, de használatuk a TEI (The Text Encoding Initiative) által szabályozott szigorú szintaxishoz kötött. Az annotációként megjelenő többletinformációk (pl. bekezdések, mondathatárok, szótövek és szófajok jelö-

lése) alkalmassá teszik a vizsgált korpuszt/szöveget többek között grammatikai (például Sass et al. 2011) vagy szintaktikai (például Seidl-Péch 2011) összefüggések feltárására is.

4. Korpuszalapú elemzés

A korpuszalapú fordításkutatást jellemző kvantitatív vizsgálatok legfontosabb tulajdonsága talán abban összegezhető, hogy a hipotézisek megerősítése vagy cáfolata nagy szövegállományok korpuszalapú lekérdezésével történik. A lekérdezési eredmények elsősorban a korpuszban gyakran és jellemzően előforduló mintázatok kutatását támogatják, bár nem lehetetlen a ritkábban előforduló mintázatok vizsgálata sem. Ez utóbbiak esetében a kutatási kérdések megjelenése általában nem a korpusz tanulmányozásához köthető, hanem azok már korábban felmerültek, és a korpusz inkább csak a már meglévő hipotézisek tesztelésére szolgál (például egyedi nyelvi elemek vizsgálata, terminuseloszlás vizsgálata). Ugyanakkor a gyakran előforduló, jellemző mintázatok esetében egyáltalán nem ritka, hogy maguk a korpuszvezérelt lekérdezések során nyert kvantitatív eredmények hívják fel a kutatók figyelmét egy-egy tipikus nyelvi jelenség létére (például konkordancia vizsgálatok).

Az utóbbi időben elterjedt az a hibás felfogás is a kutatók körében, hogy a kvantitatív vizsgálatok eredményei annál megbízhatóbbak, minél nagyobb tokenszámú (szóközzel határolt szavak száma) korpuszon történnek a lekérdezések. Mivel a korpuszok méretéből még nem lehet egyértelműen következtetni a mintavétel pontosságára, vagyis hogy a korpusz mennyire pontosan reprezentálja a vizsgálni kívánt nyelvet/nyelvváltozatot, ezért a diszciplínát jellemző paradigmaváltás következtében általános elvárássá vált a korpuszalapú kvantitatív vizsgálatok eredményei tekintetében a **statisztikai szignifikancia vizsgálata** is (Bisiada 2017: 242). Ez utóbbi segít annak alátámasztásában, hogy a minta vizsgálata során észlelt mérési eredmények nem a véletlenek összejátszásából, a mintavétel hibájából vagy valamely mérési hibából következnek, hanem valóban jellemzik a vizsgálni kívánt nyelvet/nyelvváltozatot. Ha ugyanis a statisztikai szignifikancia vizsgálatok alapján a lekérdezések során kapott eredmények nem szignifikánsak, akkor

önmagukban ezen eredmények alapján még nem lehet a korpuszban reprezentált nyelvre/nyelvváltozatra egyértelmű következtetéseket levonni.

A kutatási céloknak megfelelően felépített reprezentatív és kiegyensúlyozott korpuszok önmagukban még nem garanciái a kutatás sikerének. A kutatás bemutatásánál azt is célszerű igazolni, hogy az egymást követő lépések milyen **módszertani elveket** követnek (vö. Dörnyei és Csizér 2012). A kutatáshoz kapcsolódó minőségbiztosítási folyamat támaszkodhat egy már korábban elvégzett sikeres kutatásra, amelynek eredményei már igazolódtak, vagy magának a kutatásnak egy kisebb mintán történő kipróbálására. Mindkét esetben fontos szerepet játszik a jelenlegi kutatás kontextusának részletes bemutatása, amely alapján eldönthető, hogy a példaként használt vagy előzetesen elvégzett pilot kutatás megfelel-e módszertanilag a jelenlegi kutatás kontextusának. Ebből arra is lehet következtetni, hogy az aktuális kutatás az alkalmazott módszertan segítségével valóban a vizsgálni kívánt kérdésekre ad-e választ. Már a pilot projekt során is mindenképpen meg kell arról győződni, hogy a kutatási kérdések és a mérési eszköz összhangban vannak-e, tehát a kutatás során valóban a feltett kérdésekre kapunk-e válaszokat, illetve hogy milyen a mérési eredmények minősége. A kvantitatív korpuszalapú kutatások **érvényessége** tehát arra vonatkozik, hogy a kapott eredmények valóban alátámasztják vagy cáfolják-e a hipotéziseket, míg a **megbízhatóságuk** az elemzés megismételhetőségére (vö. Károly 2002), a következetes osztályozásra és a számítások helyességére enged következtetni (vö. Dörnyei 2011: 48–78).

A kutatás **belső érvényessége**, azaz annak teljesülése, hogy a vizsgált függő változókban valóban a független változók okoznak-e változást, nagy mértékben függ attól, hogy a kutatás figyelembe veszi-e az összes olyan változót, amely befolyásolhatja a mért eredményeket. Az egyes lekérdezések során a kutatóknak ügyelniük kell arra is, hogy ezen változók közül mindig csak egyet változtassanak meg az eredmények kiértékelhetősége érdekében. A belső érvényességgel rendelkező kutatások esetében a reprezentatív és kiegyensúlyozott korpuszokból nyert adatok **külső érvényességet** is mutatnak, azaz a megállapítások érvényesek lesznek arra a nyelvre/nyelvváltozatra, amelyet ezek a korpuszok reprezentálnak.

5. Összegzés

Az ezredfordulót követő módszertani paradigmaváltás számos fordítástudományi kutatás esetében a korpuszalapú megközelítést helyezte előtérbe, amelyek között az eredményorientált kutatások mellett egyre nagyobb számban jellennek meg az előkészítés szempontjából sokkal nagyobb körültekintést igénylő folyamatorientált kutatások is. Az elmúlt évtizedek gyakorlatából kiindulva, a korpuszalapú fordítástudományi kutatások előkészítésének és lefolytatásának metódusa egyfajta purifikációs folyamaton esett át. Részben a rendelkezésre álló eszközök technikai kapacitásának növekedése, részben az újabban megjelenő kutatási szempontok integrálása irányította a kutatók figyelmét a körültekintőbb adatgyűjtés fontosságára és a lekérdezések statisztikai validálására. Az ezek hatására gyökeresen megújuló fordítástudományi korpuszalapú kutatások szakítani látszanak a területet jellemző kezdeti átgondolatlanságokkal és felszínességgel: a második generációs kutatásokat egyre inkább összeköti egyfajta következetes metodológiai koncepció, amely a kutatási eredmények megbízhatóságáért is felel.

Irodalom

Aksan, Y.; Aksan, M. 2018. Linguistic corpora: A view from Turkish. In: Oflazer, K., Saraçlar, M. (eds) *Turkish Natural Language Processing. Theory and Applications of Natural Language Processing.* Springer International Publishing. 301–327. https://doi.org/10.1007/978-3-319-90165-7_14

Bakaja Z. 2020. Az indiai vallási irodalomban használatos beszélőjelölő igéjének magyar fordítása. In: Robin E., Seidl-Péch O. (szerk.) *Fókuszban a fordított és a tolmácsolt szöveg: korpuszalapú fordításkutatás Magyarországon.* Segédkönyvek a nyelvi közvetítésről I. Budapest: ELTE BTK Fordítástudományi Doktori Program, MANYE Fordítástudományi Szakosztály. https://doi.org/10.36252/Nyelvikozvsegedkonyv1.9

Balaskó M. 2005. Korpusznyelvészeti vizsgálatok és fordításnyelvi minták (angol és magyar tudományos szövegek anyaga alapján). Doktori értekezés. Budapest: ELTE.

Balaskó M. 2007. A fordításnyelvről, avagy a flamand szőnyeg láthatatlan szálairól. In: Heltai P. (szerk.) 2007. *Nyelvi modernizáció. Szaknyelv, fordítás, terminológia. A XVI. MANYE Kongresszus előadásai.* Gödöllő. 2006. április 10−12. (A MANYE Kongresszusok előadásai 3.) Pécs–Gödöllő: MANYE–Szent István Egyetem. 159–166.

Bańos, R., Bruti, S., Zanotti, S. 2013. Corpus linguistics and Audiovisual Translation: in search of an integrated approach. *Perspectives* Vol. 21. 483–490. https://doi.org/10.1080/0907676x.2013.831926

Bisiada, M. 2017. Universals of editing and translation. In: Hansen-Schirra, S., Czulo, O., Hofmann, S. (eds) *Empirical modelling of translation and interpreting*. Berlin: Language Science Press. 241–275. https://doi.org/10.5281/zenodo.1090972

Bowker, L., Pearson, J. 2002. *Working with specialized language. A practical guide to using corpora*. London: Routledge. https://doi.org/10.4324/9780203469255

Castagnoli, S., Niemants, N. 2018. Corpora worth creating: A pilot study on telephone interpreting. *InTRAlinea* Különkiadás. Elérhető: http://www.intralinea.org/specials/cbis

Dankó Sz. 2017. Alulreprezentált célnyelvi elemek a fordításban, avagy a „szokott" esete *Fordítástudomány* 19. évf. 1. szám 75–84.

Devaux, J. 2018. Technologies and role-space: How videoconference inter- preting affects the court interpreter's perception of her role. In: Fantinuoli, C. (ed.) *Interpreting and technology*. Berlin: Language Science Press. 91–117. https://doi.org/10.5281/zenodo.1493297

Dörnyei, Z. 2011. Research Methods in Applied Linguistics. Quantitative, Qualitative, and Mixed Methodologies. Oxford: Oxford University Press.

Dörnyei, Z., Csizér, K. 2012. How to Design and Analyze Surveys in Second Language Acquisition Research. In: Mackey, A., Gass, S. M. *Research Methods in Second Language Acquisition: A Practical Guide.* New Jersey: Wiley-Blackwell Publishing. 74–94. https://doi.org/10.1002/9781444347340.ch5

Fantinuoli, C. 2018. Interpreting and technology: The upcoming technological turn. In: Fantinuoli C. (ed.) *Interpreting and technology*. Berlin: Language Science Press. 1–12. http://doi.org/10.5281/zenodo.1493289

Holmes, J. S. 1975. The Name And Nature Of Translation Studies. In: Holmes, J. S. 1988. *Translated!: Papers on Literary Translation and Translation Studies*. Amsterdam: Rodopi. 66–80.

Jablonkai, R. 2009. „In the light of": A corpus-based analysis of lexical bundles in tow EU-related registers. *WoPaLP* Vol 3.

Karakanta, A., Dehdari, J., van Genabith, J. 2018. Neural machine translation for low-resource languages without parallel corpora. *Mach Tranlat* No. 32. 167–189. https://doi.org/10.1007/s10590-017-9203-5

Károly K. 2002. Az alkalmazott nyelvészeti kutatások néhány alapvető módszertani kérdéséről. *Alkalmazott Nyelvtudomány* 2. évf. 1. szám. 77–87.

Károly K. 2003. Korpusznyelvészet és fordításkutatás. *Fordítástudomány* 5. évf. 2. szám. 18–26.

Klaudy K. 2004. Fordítástudomány az ezredfordulón. *Publicationes Universitatis Miskolciensis* Vol. 9. 157 – 170. Elérhető: http://efolyoirat.oszk.hu/02100/02137/00001/pdf/EPA02137_ISSN_1219-543X_tomus_9_fas_1_2004_hun_eng_157-170.pdf

Klaudy K. 2005. Párhuzamos korpuszok felhasználása a fordításkutatásban. In: Lanstyák I. és Vančoné Kremmer I. (szerk.). *Nyelvészetről – változatosan. Segédkönyvek egyetemisták és a nyelvészet iránt érdeklődők számára*. Dunaszerdahely: Gramma Nyelvi Iroda. 153–185.

Laviosa, S. 1998a. The Corpus-based Approach: A New Paradigm in Translation Studies. In: *Meta: Translators' Journal* Vol. 43. No. 4. 474–479. http://doi.org/10.7202/003424ar

Laviosa, S. 1998b. Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta: Translators' Journal* Vol. 43. No. 4. 557–570. https://doi.org/10.7202/003425ar

Malaczkov Sz. 2020. A vonatkozó mellékmondat használata a feliratokban: TED-előadások magyar nyelvű összehasonlítható korpuszának vizsgálata. In: Robin E., Seidl-Péch O. (szerk.) *Fókuszban a fordított és a tolmácsolt szöveg: korpuszalapú fordításkutatás Magyarországon.* Segédkönyvek a nyelvi közvetítésről I. Budapest: ELTE BTK Fordítástudományi Doktori Program, MANYE Fordítástudományi Szakosztály. https://doi.org/10.36252/Nyelvikozvsegedkonyv1.10

Mány D. 2020. Idegen szavak félautomatikus elemzése autentikus és fordított orvosi szövegekben: fordítási stratégiák angolról franciára és magyarra fordított betegtájékoztatók tükrében. In: Robin E., Seidl-Péch O. (szerk.) *Fókuszban a fordított és a tolmácsolt szöveg: korpuszalapú fordításkutatás Magyarországon.* Segédkönyvek a nyelvi közvetítésről I. Budapest: ELTE BTK Fordítástudományi Doktori Program, MANYE Fordítástudományi Szakosztály. https://doi.org/10.36252/Nyelvikozvsegedkonyv1.6

McCarthy, M., O'Keeffe, A. 2010. Historical perspective: what are corpora and how have they evolved? In: McCarthy, M., O'Keeffe, A. (eds) *The Routledge Handbook of Corpus Linguistics.* New York: Routledge. 3–13. https://doi.org/10.4324/9780203856949.ch1

Nagy A. L. 2020. A szógyakoriság gépi vizsgálata műszaki szövegeken: terminusok és műszaki formulaszerű elemek. In: Robin E., Seidl-Péch O. (szerk.) *Fókuszban a fordított és a tolmácsolt szöveg: korpuszalapú fordításkutatás Magyarországon.*

Segédkönyvek a nyelvi közvetítésről I. Budapest: ELTE BTK Fordítástudományi Doktori Program, MANYE Fordítástudományi Szakosztály. https://doi.org/10.36252/Nyelvikozvsegedkonyv1.8

Robin E. 2018. *Fordítási univerzálék és lektorálás.* Budapest: Eötvös József Kiadó. http://www.eltereader.hu/kiadvanyok/robin-edina-forditasi-univerzalek-es-lektoralas/

Robin, E., Götz, A., Pataky, É., Szegh H. 2017. Translation Studies and Corpus Linguistics: Introducing the Pannonia Corpus. In: *Acta Universitas Sapientiae Philologica* Vol. 9. No. 3. 99–116. https://doi.org/10.1515/ausp-2017-0032

Sass B., Váradi T., Pajzs J., Kiss M. 2011. *Magyar igei szerkezetek. A leggyakoribb vonzatok és szókapcsolatok szótára.* Budapest: Tinta Könyvkiadó.

Seidl-Péch O. 2011. *Fordított szövegek számítógépes összevetése*. In: Bocz Zs., Sárvári J. (szerk.) 2013. *Válogatott cikkek, tanulmányok, 2010–2013*. Budapest: BME GTK Idegennyelvi Központ. 369–386.

Seidl-Péch O. 2018. Melyek a (szak)fordító és a fordításkutató munkáját segítő legfontosabb nyelvi korpuszok? In: Robin E., Zachar V. (szerk.) *Fordítástudomány ma és holnap*. Budapest: L'Harmattan Kiadó. 175–191.

Seidl-Péch O. 2019. *Korpusznyelvészeti kutatások a fordítástudomány szolgálatában*. Elhangzott: Nyelvi Közvetítés a digitalizáció korában. Miskolc: Miskolci Egyetem, BTK. (2019. január 23.)

Seidl-Péch O.; Robin E. 2019. *Alkalmas-e a korpuszalapú módszer a terminuskezelési stratégiák kutatására?* In: Dróth, J. (szerk.) *Korpusz és kontrasztivitás a szakfordítás oktatásában és gyakorlatában.* Budapest: Károli Gáspár Református Egyetem – L'Harmattan Kiadó. 93–104.

Seresi M. 2016. *Távtolmácsolás és távoktatás a tolmácsképzésben*. Budapest: ELTE Eötvös Kiadó.

Vincze V. 2018. A Miskolc Jogi Korpusz nyelvi jellemzői. In: Szabó M., Vinnai E.: *A törvény szavai. Prudentia Iuris* Vol. 33. 9–36