Durst Péter[1] – Szabó Martina Katalin[2]
– Vincze Veronika[3] – Zsibrita János[4]

A „HUNLEARNER" MAGYAR TANULÓI KORPUSZ FEJLESZTÉSE ÉS VÁRHATÓ HOZADÉKAI[5]

Abstract

HunLearner is a new corpus that includes essays written by learners of Hungarian. Here, we give a comprehensive review of its construction and the possibilities it may offer in language teaching. Following a general description of learner corpora, we present the results of analyses that were based on data got from HunLearner and outline directions for future research. As compiling a learners' dictionary is definitely a promising area where these results may be used, its theoretical background is described in detail and we also show how our corpus can contribute to this research area.

Keywords: *learner corpus, computational linguistics, learners' dictionary*
Kulcsszavak: *tanulói korpusz, számítógépes nyelvészet, tanulói szótár*

1. Bevezetés

Tanulmányunkban a tanulói korpuszok felhasználási lehetőségeit mutatjuk be a magyar mint idegen nyelv szempontjából, továbbá a lehetőségek szemléltetése érdekében összefoglaljuk ezen a területen eddig elért eredményeinket. A tanulói korpuszok nyelvtanulóktól gyűjtött nyelvi adatokat tartalmaznak, amelyeket megfelelő számítógépes eszközökkel elemezve új felismerésekre juthatunk. A tudományos eredményeken kívül, illetve azok felhasználásával a nyelvtanulást segítő gyakorlati előnyökre is szert tehetünk, így például segítséget kaphatunk tananyagok szerkesztésében, de a nem túl távoli jövőben akár olyan program is készülhet, amely a nyelvtanulók egyes hibáit is képes lesz javítani. A magyar nyelv számítógépes feldolgozása gazdag mor-

[1] Durst Péter, PhD, Szegedi Tudományegyetem, Hungarológia Központ, durst.peter@gmail.com
2 Szabó Martina Katalin, Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék, szabomartinakatalin@gmail.com
[3] Vincze Veronika, PhD, MTA-SZTE Mesterséges Intelligencia Kutatócsoport, vinczev@inf.u-szeged.hu
[4] Zsibrita János, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, zsibrita@inf.u-szeged.hu
[5] A jelen kutatás részben a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

fológiája miatt igen összetett feladat, így a hibás formákat is tartalmazó nyelvtanulói adatok elemzése hatványozottan nehéznek tűnik. Ugyanakkor az eredmények akár túl is mutathatnak a magyar mint idegen nyelv tanításának korlátain.

A tanulmányban használjuk a *köztes nyelv* (Selinker 1972) fogalmát, melynek jelentését széles körű használata miatt itt nem tartjuk szükségesnek kifejteni. A *hiba* terminussal kapcsolatban fontosnak tartjuk megjegyezni, hogy ebben a tanulmányban általánosságban a célnyelvitől eltérő formák megjelölésére fogjuk használni, mert a nyelvészeti szakirodalomban ismert distinkció a nyelvtudás hiányosságából fakadó, szisztematikusan visszatérő *hiba (error)* és a figyelmetlenség miatt elkövetett, alkalmi jellegű *tévesztés (mistake)* között az adott kutatási és nyelvfeldolgozási helyzetben nem alkalmazható.

2. A nyelvtanulói korpuszok és a HunLearner

2.1. A nyelvtanulói korpuszokról általában

A nyelvtanulói korpusz fogalmát pontosan meghatározó definíciót nehéz találni a szakirodalomban, így annak tekinthető valójában minden írott vagy hangzó formában elérhető nyelvi adat, amely nyelvtanulóktól származik. A mai technikai feltételek fényében azonban csak olyan nyelvi adathalmazt érdemes tanulói korpusznak tekinteni, amelyet számítógépes eszközökkel lehet elemezni – tehát digitális formában elérhető (vö. Szirmai 2005:16-19). Ez a feltétel számos kérdést vet fel, hiszen ha a nyelvtanulók nem eleve digitális formában készítik el a később elemezni szánt szöveget (például egy fogalmazás formájában), akkor a kézírást, illetve a hangzó anyagot is megfelelő módon át kell írni, ami rendkívül alapos előkészítést és sok munkát igényel.

Bár nyilvánvalóan a legtöbb nyelvtanulót számláló angol nyelvnek van a legtöbb és a legkiterjedtebb tanulói korpusza, örvendetes módon már számos más nyelv esetében is sikerült megvalósítani ilyen vállalkozást. Példaként említhetjük a morfológiailag összetettebb cseh (Hana et al. 2010), valamint a morfológiai és tipológiai szempontból is érdekes finn nyelv tanulói korpuszát (Jantunen 2011). Ezek nem pusztán azoknak az érdeklődését kelthetik fel, akik az adott nyelvvel foglalkoznak, hiszen a korpusz általános jellemzői, az adatok gyűjtésének és kezelésének módjai, továbbá a számítógépes eszközök használata szinte minden más kutató figyelmét megragadhatják. Így mindenképpen érdemes megemlíteni, hogy a finn tanulói korpuszt a más országokban finnül tanuló egyetemi hallgatóktól gyűjtötték, ami egy rendkívül hatékony és kooperatív hozzáállást feltételez a projekt minden résztvevőjétől. A cseh korpusz pedig több alkorpuszával, hangzó és írott szöveget tartalmazó, több nyelvtudási szintet is átfogó összetettségével, valamint kiemelkedően szerteágazó hibakódolási módszerével hívja fel magára a figyelmet.

Egy nyelvtanulói korpusz alapvető jellemzői közé tartozik, hogy hangzó és/vagy írott anyagot tartalmaz. A hangzó anyagokat tartalmazó korpuszok készítésekor az

adatgyűjtés módszerének körültekintő meghatározása mellett az átírást is részletesen szabályozni kell, ami jelentősen megnehezíti a munkát. A kézzel írott anyagok számítógépes feldolgozásához is szükség van egy részletes útmutatóra, amely kitér például a hibás vagy a kiolvashatatlan részek megfelelő jelölésére is. Előzetes döntés és természetesen a lehetőségek kérdése is, hogy milyen nyelvtudási szintű nyelvtanulókat vonnak be az adatgyűjtésbe, de szerencsére ezzel kapcsolatban viszonylag könnyen javíthatók a hiányosságok, hiszen a tanulói korpuszok is folyamatosan bővíthetők. A nyelvi adatokon túl a legtöbb esetben az adatközlők személyes adatai is rögzítésre kerülnek, ezek segítségével ugyanis tovább bővül a statisztikai értékelés lehetősége (például az életkor, az anyanyelv vagy más idegen nyelvek ismeretének figyelembevételével).

A tanulói korpuszok kezelésének legfontosabb kérdése az annotálás, valamint a hibák keresése, kódolása és esetleges javítása. Az annotálás lényege, hogy a szöveghez – a megfelelő nyelvi elemzést követően – az alkotóelemeire vonatkozó információt adunk hozzá, amelynek segítségével azután például listákat vagy statisztikákat készíthetünk. Ezt szintén lehet manuálisan is végezni, de egy nagyobb korpusz esetében a manuális feldolgozás rendkívül idő- és munkaigényes. Magyar nyelvű szövegek elemzéséhez is elérhetők olyan számítógépes eszközök, amelyek kiváló pontossággal végzik el a szövegek mondatokra, szavakra és morfémákra történő felbontását és elemzését (lásd a 2.4. bekezdésben), és ugyan ezek eredetileg a sztenderd magyar nyelvváltozat feldolgozására készültek, kiváló eredményeket lehet velük elérni a tanulói korpuszok elemzésében is. Az annotáció tehát tartalmazhat például egy morfológiai elemzést, amelynek alapján ki lehet listázni a korpusz szövegében előforduló összes főnevet, igét vagy az összes helyhatározóragos szóalakot – akár az előfordulás gyakorisága szerint is.

A hibák keresése és kódolása szintén végezhető manuálisan és számítógépes eszközökkel is, bár az előbbi az annotátorok képzése és a több személy által egységesen végzett munka igénye miatt igen körülményes. A hibák kategorizálásához használt kódrendszer kidolgozása alapos előkészítő munkát igényel, amelynek során figyelembe kell venni az elemzés célját, valamint a számítógépes eszközök kínálta lehetőségeken túl azok korlátait is. Az elemzésnek ez a mozzanata köti össze a leíró nyelvészet által megalkotott fogalmakat a gépi elemzés lehetőségeivel, így gyakran van szükség kompromisszumokra, esetleg új kategóriák felállítására. A hibák felkutatása és kódolása után nyílik lehetőség a hibakódok segítségével különböző elemzések elvégzésére, amelyek eredménye mind a nyelvészeti kutatómunkában, mind pedig a nyelvoktatásban is jól használható. Egyes esetekben a hibákat még javítják is, ehhez azonban gyakorlatilag elengedhetetlen az emberi beavatkozás.

2.2. Egy amerikai tanulói korpusz

A magyar mint idegen nyelv elemzéséhez eddig két tanulói korpusz született: a jelen dolgozatban bemutatott HunLearneren kívül az egyesült államokbeli Indiana egyetem kutatói publikálták nemrégiben a témához kapcsolódó eredményeiket (Dickinson–Ledbetter 2012). Az Indiana egyetemen vannak magyar nyelvórák, így az adatokat is az

ott tanuló diákoktól gyűjtötték. Összesen 14 írás szerepel a korpuszban, amelyek mindegyike 10–15 mondat hosszú, tartalmuk pedig különféle témákban írt naplóbejegyzés. A tanulmány egyik szerzője maga is haladó szintű nyelvtanuló, így az annotálást is ő végezte egy magyar anyanyelvű lektor segítségével, azonban a sikeres együttműködést megkérdőjelezi, hogy magában a publikált tanulmányban is több helytelen magyar mondat szerepel helyesként feltüntetve. Ez a tanulmány egy konkrét nyelv – adott esetben a magyar – sajátosságainak figyelembevételével inkább mégis a hibakódolás egy elméleti megközelítését mutatja be, hiszen láthatjuk ugyan egy többszintű hibakódolási rendszer alapelveit, de a szöveg szegmentálása, annotálása és hibajavítása is manuálisan történt, a hibakódok rendszere pedig nincs elég részletesen kidolgozva ahhoz, hogy jól használható statisztikai elemzéseket lehessen végezni segítségükkel. Az indianai korpusz feldolgozását bemutató tanulmány ettől függetlenül – a tanulói korpuszok feldolgozására vonatkozó általános tapasztalatokra alapozva – tartalmaz olyan lényeges megállapításokat, amelyeket érdemes figyelembe venni. Az annotációban kategóriák és szintek különböztethetők meg, az egyes szinteken elvégzett javítások sorba vannak rendezve, de azon belül a kategóriák nincsenek rangsorolva.

A kategóriák lefedik a lehetséges hibák teljes skáláját, így a helyesírási *(Character),* morfológiai *(Morpheme),* grammatikai viszonyokat *(Relation)* magába foglaló és a mondat szintű *(Sentence)* területeken is lehetséges a kódolás, valamint a javítás. A hibák annotálásának rendszerét itt nem részletezzük, mindössze a morfológiai hibák kategóriáit mutatjuk be. A morfológiai hibákat alapvetően két kategóriába osztják: egyeztetési hibák *(Agreement)* és szóképzési hibák *(Derivation).* Az egyeztetési hibákon belül megkülönböztetik a személy *(Person),* a szám *(Number),* az eset *(Case)* és a határozottság *(Definiteness)* jelölésével kapcsolatos hibákat, míg a szóképzésen belül a kihagyást *(Omission),* a beszúrást *(Insertion)* és a sorrendet *(Ordering)* lehet jelölni hibaként.

Ezzel a hibakódolási rendszerrel ugyan minden típusú hibát tudnak valamilyen módon kódolni, de véleményünk szerint jobban használhatók a kódolás eredményei, ha ennél részletesebben megkülönböztetik a hibákat. Ha a tanulmányban közölt adatok alapján górcső alá vesszük a morfológiai elemzést, akkor felmerül bennünk, hogy lényeges lenne például a szótövekkel vagy a kötőhangok használatával kapcsolatos hibák elkülönítése (esetleg még a különböző tőtípusok és a különböző hangrendű szavak szerint is). Az angol nyelv alaktani jellemzőit tekintve sokkal részletesebb morfológiai elemzésnek számít persze már az esetek jelölése is (illetve a hibakódolásban a morfológiai jelölés elmulasztásának feltüntetése), azonban a magyarral kapcsolatban ennek még csekély az információértéke.

Természetesen nem szabad figyelmen kívül hagyni az amerikai tanulmányban használt annotáció egyik alapelvét – amelyet egyébként éppen a fentebb is említett cseh korpusz szerzőire (Hana et al. 2010) hivatkozva alkalmaznak –, miszerint elengedhetetlen kompromisszumokat kötni, és csak az adott projekt lehetőségeinek megfelelő, megbízhatóan annotálható tulajdonságokkal foglalkoznak.

2.3. A HunLearner magyar nyelvtanulói korpusz

A Szegedi Tudományegyetemen indult projektben a HunLearner korpusz feldolgozása során is az indianai korpusz építésében alkalmazott alapelvet követtük és követjük, azaz igyekszünk a megbízhatóan annotálható nyelvi sajátságokra fókuszálni, azonban ennek megvalósítását az amerikai kutatóktól eltérően képzeljük el. Már a legelső egyeztetések során egyhangúan úgy döntöttünk, hogy mindig csak egy-egy részfeladat megvalósítására koncentrálunk, azt azonban a technikai lehetőségek és az elemzési célok figyelembevételével a lehető legjobban kidolgozzuk. Ez a megközelítés igen hasznosnak bizonyult, mert az eddig elvégzett munka során számos olyan apró, ám lényeges probléma merült fel, amelyek megoldása elengedhetetlen a jól használható statisztikai eredmények kinyeréséhez, valamint az egyéb távlati célok (pl. automatikus hibajavítás) megvalósításához.

Bár korpusznak tekinthető valójában minden összegyűjtött írás, beadott dolgozat vagy hangfelvétel, érdemi megállapításokhoz előre meghatározott feltételek mellett gyűjtött nyelvi produkció szükséges. Saját korpuszunk építését a Zágrábi Egyetem hungarológia szakos hallgatóinak beadványaival kezdtük, így az első két alkorpuszban a magyar nyelv nehézségeiről szóló írások és a külföldi munkavállalásról szóló beadványok találhatók. Ezután kezdtük el a szövegek szisztematikus gyűjtését, így a HunLearnerben az imént említetteken kívül csak olyan írásban készült anyagok szerepelnek, amelyek egységesen megfelelnek a projekt kezdetekor lefektetett követelményeknek: terjedelmük kb. 1500 karakter, kb. egy óra alatt készültek szótár, nyelvkönyv és egyéb segítség nélkül, csak a nyelvtanuló saját nyelvtudása és készségei alapján, elektronikus formában, az összes magyar ékezetes karaktert tartalmazó billentyűzettel. Egy rövid indulási szakasztól eltekintve két meghatározott téma közül lehetett választani, így a válaszadók „Egy szimpatikus ember" vagy „Magyarországról és a magyarokról" címmel írtak fogalmazásokat. A korpusz bővítésekor jelenleg is ezeket a feltételeket kérjük betartani. Elengedhetetlen volt ilyen egyszerű, viszonylag könnyen teljesíthető követelményeket szabni, ellenkező esetben ugyanis túl nagy feladatot jelentene a fogalmazás az egyébként is önkéntes módon résztvevő nyelvtanulóknak és a tanáraiknak. A tanár kollégák időnként vállalták, hogy az írásokat egy kurzus egyik otthon elkészítendő feladataként kérték be, majd kijavították és megbeszélték a hallgatókkal. A fogalmazásokat eddig csak felsőoktatási intézmények legalább A2-B1 nyelvtudási szinttel rendelkező magyarul tanuló hallgatói írták, de valójában sem az anyaggyűjtés helyével, sem pedig a nyelvtudási szinttel kapcsolatban nincsenek szigorú megkötések. Természetesen a későbbiekben lehetséges a korpusz bővítése hangzó anyagokkal és kézzel írt beadványokkal is, azonban ezek megfelelő átírása egyelőre meghaladja a lehetőségeinket.

Az adatok felvételekor a válaszadókra vonatkozó személyes adatokat is rögzítettünk, így a nyelvi anyag elemzésekor figyelembe lehet venni az életkort, az anyanyelvet, a többi idegen nyelv ismeretét, a Magyarországon eltöltött időt és a magyar tanulásával töltött idő hosszát – az eddigi statisztikákban azonban ezek a változók még nem szerepelnek.

A HunLearner korpusz jelenleg 1427 mondatot és mintegy 22 000 tokent tartalmaz. Az anyagot különböző számítógépes eszközökkel elemezve (lásd a 2.4. fejezetben) eddig három nagyobb kérdéskört vizsgáltunk meg: a főneveknél megfigyelhető morfológiai hibákat, a határozott tárgyas ragozás használatának egyes jellemzőit, valamint elemeztük és összehasonlítottuk MID-tananyagok olvasmányainak a szövegét a tanulói korpusz szövegével. Ezeknek az elemzéseknek az összefoglalása a 3. fejezetben szerepel. A határozott tárgyas ragozás vizsgálata már átvezet a mondatszintű elemzések területére, ami projektünk következő nagyobb állomása lesz.

2.4. A HunLearner automatikus elemzése

A számítógépes nyelvészet fejlődésének köszönhetően ma már számos nyelv automatikus feldolgozására nyílik lehetőség különféle nyelvi elemző eszközök segítségével. Ezek az eszközök a bemenetül kapott szöveget első lépésben mondatokra bontják, majd a mondatokat további alkotóelemekre – szavakra, illetve írásjelekre – tagolják. Ezt követően a szófaji egyértelműsítés során az egyes szavakhoz az aktuális mondatkörnyezetnek megfelelő szófaji és részletes morfológiai elemzést rendelnek. Ezután az egyes szavak közti szintaktikai kapcsolatok megállapítására kerül sor, azaz minden mondathoz hozzárendeljük annak szintaktikai elemzését. Így a nyers szövegtől eljuthatunk annak morfológiailag és szintaktikailag annotált változatához, teljes egészében automatikus úton.

A *magyarlanc* nevű programcsomag (Zsibrita–Vincze–Farkas 2013) magyar nyelvű szövegek automatikus elemzésére képes a szövegek mondatra bontásától kezdve egészen a szintaktikai (függőségi) elemzésig. Az elemző nemzetközi mércével mérve is kielégítő pontosságot ér el sztenderd magyar szövegeken mind a szófaji egyértelműsítést, mind a függőségi elemzést tekintve[6], így vizsgálatainkban is ezt az eszközt alkalmaztuk. Elemzéseink kiindulópontját tehát a *magyarlanc* által elemzett szövegek jelentik.

3. Az eddig elvégzett részfeladatok

3.1. A főnevek morfológiájának elemzése

A korpuszt a *magyarlanc* elemzővel (Zsibrita–Vincze–Farkas 2013) automatikusan átnéztük, majd az ismeretlennek minősített szavakat további vizsgálatnak vetettük alá. Természetesen ki kellett szűrni az ismeretlennek minősülő szóalakok közül az idegen szavakat és a tulajdonneveket, majd a továbbiakban csak a főnevekre koncentráltunk. A hibás szóalakokat a *hunspell* helyesírás-ellenőrző (Trón et al. 2005) segítségével javí-

[6] A magyarlanc kísérleteinkben használatos változata szófaji egyértelműsítésben 96%-os pontosságra (accuracy), függőségi elemzésben pedig 93%-os pontosságra (az ULA metrika szerint) képes, a Szeged Dependencia Treebank adatbázison tanítva és kiértékelve (vö. Zsibrita–Vincze–Farkas 2013).

tottuk, de azokban az esetekben, ahol több lehetőséget is ajánlott a program, kézzel kellett kiválasztani a kontextusba illőt. Az automatikus javítás nagyon eredményesnek bizonyult, mert amennyiben a *hunspell* által javasolt első helyes szóalakot választottuk, akkor 81,86%-os pontosságot értünk el az összes javított szóalak figyelembevételével, ami az összes ismeretlen szóalak 49%-ának felel meg. Eredményeink arra utalnak, hogy már egyszerű módszerekkel is jelentősen, körülbelül felére lehet csökkenteni a hibás szóalakok számát egy nyelvtanulók által írt szövegben, ez pedig igen ígéretesnek mutatkozik a tanulói szövegek automatikus feldolgozására nézve.

A morfológiai hibák osztályozására egy saját kategóriarendszert és egy ennek megfelelő kódrendszert hoztunk létre az általános nyelvtanári tapasztalat, valamint a magyar mint idegen nyelv vonatkozásában készült hibaelemzések alapján (Durst 2010). A kódok négy karakterből állnak, melyek közül az első a szótővel, a második a hasonulással, a harmadik a hangrenddel, kötőhangokkal és a toldalékok allomorfjaival, a negyedik pedig a toldalékok számával kapcsolatos hibák típusát jelzi. Az automatikus hibakódolás lehetővé tette az egyes hibatípusok számszerűsítését, így meg tudtuk állapítani a tő- és toldaléktévesztések arányát, illetve a hasonulási és hangrendi problémák arányát is. Az eredmények szerint a leggyakoribb hibatípus a tőtévesztés (85%) volt, különös tekintettel az ékezetek nem megfelelő használatára (28%). A toldaléktévesztések közül pedig a hibás kötőhang volt a leggyakoribb (29%).

3.2. A határozott tárgyas ragozás elemzése

A határozott tárgyas ragozás használatának elemzése előtt mindenképpen fel kell hívni a figyelmet arra, hogy a számítógépes eszközök nem tudnak úgy „gondolkodni", mint egy leíró nyelvészettel foglalkozó szakember vagy mint egy házi feladatot javító nyelvtanár. Szükségszerűen előfordulnak olyan esetek, amelyeknek helye lenne ugyan a statisztikában, de azonosításuk nem megoldható. A jelen elemzésben a határozott tárgyi tömbök fő típusait tudtuk csak figyelembe venni, így olyan, viszonylag ritka eseteket nem is állt szándékunkban bevonni, mint például a *valamennyi* névmás tárgyi szerepben. A tárgyi alárendelő mondatok és az explicit módon nem megjelenő tárgyak (amikor a tárgyra csak magával a határozott ragozással utalunk) pedig az automatikus azonosításukkal kapcsolatos nehézségek miatt nem szerepelnek. Csak érdekességként jegyezzük meg, hogy a magyarlancban nem megoldott az intranzitív igék azonosítása sem, így határozott ragozású és határozott tárggyal álló intranzitív igék elméletileg megjelenhetnek a statisztikában, de ennek igen csekély a valószínűsége (vö. **futom az almát*).

A HunLearner korpusz szövegeit a *magyarlanc* szoftverrel automatikusan elemeztük, majd a morfológiai és szintaktikai elemzés alapján összegyűjtöttük azokat az eseteket, amelyekben eltérés mutatkozott a tárgy típusa által indikált és a tényleges igeragozás között. Az elemzés következő fázisában azokat a morfológiailag többértelmű igealakokat is kizártuk, ahol a határozott és határozatlan ragozás egybeesik (pl. múlt idő E/1. alakban, vö. *olvastam*), itt ugyanis nem eldönthető, hogy a nyelvtanuló határozott vagy határozatlan ragozást kívánt-e használni.

Összesen 2423 igét vizsgáltunk, és 372 esetben volt helytelen a határozatlan / határozott tárgyas ragozás közti választás. Ezek közül 117 olyan eset fordult elő, amelyben volt tárgya az igének a mondatban, és a nyelvtanuló nem a helyes ragozást választotta. Ebből a 117 hibás esetből még kiszűrtük azokat az igéket, amelyek morfológiailag nem többértelműek alanyi és tárgyas ragozás között (vagyis például az E/1. múlt időt), így végül 87 esetet vetettünk alá további vizsgálatoknak. Az eredmények szerint a leggyakoribb hibaforrás a határozott névelős köznévi tárgy: ez határozott ragozást váltana ki, azonban a hibák 17%-ában határozatlan ragozású igével szerepel együtt. Két másik gyakori hiba a mutató névmási tárgy és a névelőtlen köznévi tárgy, melyek a hibák 13–13%-ában a nem megfelelő ragozású igével fordulnak elő. Az eredmények egyben azt is mutatják, hogy jóval több a határozott tárgy-határozatlan igealak típusú tévesztés (59%), mint a határozatlan tárgy-határozott igealak típusú.

3.3. MID-tananyagok elemzése és összevetése a tanulói korpusszal

Hat MID-tankönyv szövegét elemeztük és vetettük össze a HunLearner tanulói korpusz anyagával. Az elemzésben a következő tankönyvek szerepeltek (megjelenésük sorrendje szerint): *Halló, itt Magyarország; Hungarolingua 1.; Lépésenként magyarul 1.; Új színes magyar nyelvkönyv 1.; Hungarian the Easy Way 1-2., MagyarOK 1.* A *Hungarian the Easy Way* a többi tankönyvtől eltérő módon három részben tartalmazza hozzávetőleg ugyanazt a nyelvismereti anyagot, így ebből a sorozatból az első részt és a második rész felét vontuk be az elemzésbe. A tankönyvek anyagát részben a szerzők bocsátották rendelkezésünkre digitális formában, részben pedig a SZTE BTK Hungarológia mesterképzés hallgatói vitték számítógépre. Az alábbiakban közölt rövid összefoglaló bővített változata hangzott el a Károli Gáspár Református Egyetemen 2013. december 14-én „A magyar mint idegen nyelv napja" című rendezvényen tartott előadás keretén belül. Az elemzések több olyan sajátosságra is rávilágítanak, amelyeket eddig intuitív módon tudhattunk, azonban a kvalitatív eredmények objektív megvilágításában most már akár hivatkozásszerűen is felhasználhatunk. Az igealakok megoszlása nem meglepő: a tankönyveknél összességében az E/1 és E/3 személyű igealakok dominálnak (E/1: 26%, E/2: 7%, E/3: 47%), míg a többes számú alakok közül egyértelműen kiemelkedik a harmadik személy (T/1: 8%, T/2: 1%, T/3: 12%).

A tananyagok szövegét is alávetettük a határozott tárgyas ragozással kapcsolatos elemzésnek, így nagyon jól megfigyelhető, hogy az olvasmányokban alapvetően a tulajdonnevek (39%), a határozott névelővel álló köznevek (32%), a birtokos szerkezetek (13%) és a mutató névmások (9%) szerepelnek határozott tárgyként. Vélhetően nem a tananyagoknak a későbbi nyelvtudásban játszott közvetlen szerepét mutatja, hanem inkább a bennük érvényesülő helyes szemléletét támasztja alá az, hogy a tanulói korpusz elemzéséből származó adatokban is hasonló arányokat fedezhetünk fel: itt a tulajdonnevek (63%), a határozott névelővel álló köznevek (16%), a birtokos szerkezetek (7%) és a mutató névmások (10%) szerepelnek határozott tárgyként.

A számítógépes elemzés alkalmat nyújt még a szókincs gyakoriságának megfigyelésére is. Az olvasmányokban előforduló főnevek gyakorisági listája a tulajdonnevek kiszűrése után is viszonylag nagy változatosságot mutat. Van ugyan néhány olyan szó, amelyik szinte mindegyik tananyagban szerepel a leggyakoribb szavak között *(ember, gyerek, egyetem, óra)*, azonban a leggyakoribb főnevek inkább az adott könyvre jellemző szituációkhoz kapcsolódnak, így a néhány száz szót tartalmazó alapszókincsen belül viszonylag nagy változatosságot láthatunk. Az igéknél is változatos a kép, azonban itt már jóval több egyezés van: a *van, megy, dolgozik, lakik, beszél* igéken túl is találunk még néhányat, amelyek majdnem minden tananyagban előfordulnak láthatóan nagyobb gyakorisággal.

A statisztikákat elnézve talán meglepő azzal szembesülni, hogy még a leggyakrabban előforduló főnevekkel is mindössze körülbelül húsz alkalommal találkozik a nyelvtanuló egy tankönyv olvasmányain belül. Az igék esetében már vannak olyanok (például *van, jön, megy, szeret, tanul, beszél, kér),* amelyek tankönyvtől függően harmincszor-negyvenszer vagy akár nyolcvanszor-kilencvenszer is előfordulnak. Megállapíthatjuk tehát, hogy az olvasmányok legfeljebb az igék esetében tekinthetők alkalmasnak a szókincs rögzítésére, a főnevek esetében ez a funkció sokkal inkább az olvasmányokhoz kötődő feladatokra hárul.

4. További lehetőségek: egynyelvű szótárak

4.1. Az egynyelvű nyelvtanulói szótárak használatának előnyei az idegennyelv-tanulási folyamatban

A fentiekből egyértelműen látható, hogy a számítógépes eszközök már most is jól használhatók a tanulói korpuszban előforduló hibák elemzésére és sok esetben javítására is, de ezt a felhasználási lehetőséget mindenképpen érdemes még továbbfejleszteni, valamint az elemzés körét más nyelvi szintekre is ki kell terjeszteni. Az igei vonzatkeretek számítógépes vizsgálatában már jelentős eredmények születtek (Vincze 2014), továbbá részben a tanulói korpuszt is elemeztük ebből a szempontból (Vincze et al. 2013), de a későbbiekben szeretnénk még részletesebben is megvizsgálni, hogyan lehet automatikus eszközökkel tovább csökkenteni a hibás vonzatkeretek számát. A hibák elemzésén és automatikus javításán túl azonban vannak még további lehetőségek is, amelyek közül kiemelkedik a tanulói szótárak készítése. A HunLearner korpusz létrehozásának és fejlesztésének egyik végső célja az, hogy vizsgálati anyagot teremtsünk egy egynyelvű magyar nyelvtanulói szótár létrehozásához. Bár számos szerző hangsúlyozza az egynyelvű nyelvtanulói szótárak alkalmazásának hasznát az idegen nyelvek tanulása során, hasonló szótár a magyar mint idegen nyelv vonatkozásában eleddig még nem készült.

Egynyelvű nyelvtanulói szótárnak nevezzük azt a szótártípust, amely a kifejezetten egy adott nyelvet idegen nyelvként tanulók igényeinek kielégítését célozza, és az egyes szótári címszóknál megadott információkat is az adott célnyelven közli (De Cock–Gran-

ger 2005: 72). Az egynyelvű nyelvtanulói szótárakat számos sajátság különbözteti meg a kétnyelvű szótáraktól, valamint az anyanyelvűeknek készült egynyelvű szótáraktól. Egyrészt az egynyelvű nyelvtanulói szótár szókészlete szűkebb, hiszen csupán a magasabb frekvenciájúnak, ezáltal a nyelvtanulók számára fontosabbnak tartott szavakat tartalmazza. Másrészt részletesebb információt nyújt az adott nyelvi kifejezés morfológiai, szintaktikai, valamint szemantikai sajátságait, viselkedését illetően (De Cock Granger 2005: 72), és a magyarázatban a nyelvtanulók igényeihez alkalmazkodva az egyszerű, a nyelvtanuló számára érthető megfogalmazásra törekszik. Harmadrészt az egyes nyelvi kifejezések használati sajátságait tipikus példákkal szemlélteti, tehát nyelvi kontextusba ágyazza, valamint egyes esetekben képekkel, ábrákkal, rajzokkal is illusztrálja (H. Gouws 2004: 269). Mindemellett a szótári szócikkek tartalma bizonyos egynyelvű nyelvtanulói szótárak esetében még további kiegészítő információkkal is bővül. Bizonyos szótárak megadják például a kifejezések antonimáit (Lee 1998: 456), de találunk példát arra is, ahol a szótárkészítők a szótári kifejezéseknek az adott kultúrabeli vonatkozásait tartották fontosnak. Ez utóbbi esetet példázza a *Longman English Dictionary of Language and Culture* (Summers 1993), amelyben a szerkesztők ún. „kulturális megjegyzés"-ben (Cultural Note) információt közölnek az adott nyelvet anyanyelvként beszélő népnek az adott kifejezéshez kapcsolódó sajátos asszociációiról.[7]

Számos kutató, köztük Berwick és Horsfall (1996: 12), valamint H. Gouws (2004: 274) is hangsúlyozza az egynyelvű nyelvtanulói szótárak használatának előnyét a kétnyelvű nyelvtanulói szótárakéhoz képest, miszerint az előbbi nem leegyszerűsített, nyelvek közötti „egy az egyben" megfeleléseket ad a nyelvtanulóknak, hanem, a fentebb bemutatott jellemzőknek köszönhetően, a kifejezések nyelvi sajátságainak pontosabb megismerését és mélyebb megértését támogatja a kétnyelvű szótárakkal szemben. A kétnyelvű szótárak használata során gyakran jelentkezik ugyanis az a probléma, hogy, bár a szótár készítői több ekvivalenst is közölnek az adott nyelvi kifejezés megfelelőjeként, nem mutatnak be olyan nyelvi környezetet, illetve nem adnak meg elegendő és megfelelő minőségű információt ahhoz, hogy a nyelvtanuló kiválaszthassa a szótárban megadott elemek közül a számára az adott esetben megfelelő ekvivalenst (Szabó 2012). Ugyanakkor azt is érdemes szem előtt tartani, hogy az egynyelvű nyelvtanulói szótár csupán a nyelvtanuló megfelelő szintű nyelvismerete esetén nyújthat igazán hathatós segítséget (H. Gouws 2004: 274; Holi Ali 2012: 3).

Mivel az egynyelvű nyelvtanulói szótárak alapvető törekvése az, hogy azokat a szavakat tartalmazza, amelyek a nyelvtanuló számára (az adott nyelvi szinten) a legszükségesebbek, az egynyelvű nyelvtanuló szótárak készítői a kezdetektől különböző szövegkorpuszok statisztikai adataira támaszkodnak (Девель 2004: 131). Ennek okán a dolgozat következő fejezetében a korpuszokról, különös tekintettel a nyelvtanulói korpuszokról, valamint azok szótárkészítésbeli hasznáról szólunk részletesebben.

[7] Az egynyelvű nyelvtanulói szótárak szisztematikus áttekintésére ebben a munkában nincs mód. E szótártípust részletesebben tárgyalja többek között Cowie (1999).

4.2. A nyelvtanulói korpuszok és felhasználhatóságuk az egynyelvű nyelvtanulói szótárak készítésében

Régóta foglalkoztatja a nyelvészeket a kérdés, miszerint hogyan lehetséges az idegennyelv-tanulás szempontjából a legcélravezetőbb, azaz a nyelvtanuló számára a legszükségesebb nyelvi kifejezéseket tartalmazó szótár megalkotása (Девель 2004: 1). Már a 20. század elejétől sorra jelentek meg azok a szótárak, amelyek elsődleges célkitűzése, hogy szövegstatisztikai vizsgálatok eredményeit alapul véve a legmagasabb frekvenciájú nyelvi elemeket rendszerezzék, és ezáltal egy ún. lexikai minimumot adjanak a nyelvtanuló kezébe. Ugyanakkor, ahogyan azt Lee (1998: 455) is hangsúlyozza a korpuszok statisztikai adatainak használhatósága kapcsán, az anyanyelvi szöveganyagok vizsgálata alapján a legmagasabb frekvenciájúnak ítélt lexika közlése az idegennyelv-tanulás szempontjából igencsak problematikusnak tekinthető. Egy olyan speciális szótárban ugyanis, mint az egynyelvű nyelvtanulói szótár, szükség lehet viszonylagosan ritkább előfordulási aránnyal rendelkező nyelvi kifejezések reprezentálására is.

Az egynyelvű nyelvtanulói szótárak készítésének céljából végzett vizsgálatok sokáig kizárólagosan anyanyelvű beszélőktől származó nyelvi produktumok analízisét jelentették; a nyelvtanulói szövegek adatainak lexikográfiai hasznosítása kifejezetten újkeletűnek tekinthető a szótártudományban. Bár az egynyelvű nyelvtanulói szótárak szerkesztése során kétségtelenül nagy segítséget nyújtanak az anyanyelvi beszélők által létrehozott szövegekből álló korpuszok, hiszen értékes információval szolgálnak az adott nyelvnek mind a lexikai és grammatikai, mind a kollokációs sajátságainak tekintetében, emellett autentikus nyelvi példák forrásául is szolgálnak (Девель 2004: 3). Ahogyan azt De Cock és Granger (2005: 72) is megemlíti, egy jól funkcionáló egynyelvű nyelvtanulói szótár készítéséhez nélkülözhetetlen a nyelvtanulói korpuszadatok figyelembe vétele is. A nyelvtanulók nyelvi produktumaiból álló korpusz rámutat ugyanis mindazokra a problémákra, amelyekkel az adott nyelvet idegen nyelvként tanulók küzdenek, lehetővé téve ezzel egy problémacentrikusabb, s ezáltal hatékonyabb egynyelvű nyelvtanulói szótár megalkotását (Rundell 1999: 47). Emellett a nyelvtanulói szövegek vizsgálata segít detektálni a nyelvtanulók számára az adott nyelvi szinten nélkülözhetetlen lexikát, valamint támpontként szolgál a lexikográfusnak ahhoz, hogy a szótári szavakhoz adott magyarázat és a példák csupán olyan kifejezéseket tartalmazzanak, amelyek a nyelvtanulók számára az adott nyelvi szinten viszonylag könnyen érthetők.

Az első, már nyelvtanulói korpuszra épülő egynyelvű nyelvtanulói szótár, a *Longman Language Activator* megjelenése 1993-ra datálható (De Cock–Granger 2005: 72). Azóta természetesen több, nyelvtanulói korpuszon alapuló egynyelvű nyelvtanulói szótár is napvilágot látott, ezek többsége azonban az angol nyelvet tanulók igényeit igyekszik kielégíteni. Mindemellett a viszonylagosan újnak tekinthető, 1998-ban megjelentetett koreai egynyelvű nyelvtanulói szótár is csupán anyanyelvi beszélők szövegkorpuszának adataira épül (Lee 1998).

Számos tényező befolyásolja azt, hogy egy adott nyelvtanulói korpusz milyen eredményességgel alkalmazható egy egynyelvű nyelvtanulói szótár szerkesztése so-

rán: a korpusz mérete és reprezentativitásának foka, valamint az, hogy rendelkezik-e a nyelvtanulói hibák kódolásával, és ha igen, milyen minőségben (De Cock–Granger 2005: 74–75). Ami a korpusz méretét illeti, De Cock és Granger (2005: 75) hangsúlyozza, hogy már a viszonylag kis méretű (körülbelül 100 ezer szövegszós) korpuszok is képesek hathatós segítséget nyújtani az egynyelvű nyelvtanulói szótárak készítése során. A korpusz reprezentativitása kapcsán fontos hangsúlyozni, hogy a nyelvtanulói korpuszt mind a nyelvtanulók, mind a benne foglalt szövegek sajátságai egyaránt meghatározzák. Ennek következtében a korpusz lexikológiai felhasználása során érdemes figyelembe venni egyrészt a korpuszszövegek típusát és keletkezési sajátságait, másrészt a nyelvtanulók életkorát, nyelvismereti szintjét, valamint anyanyelvi hátterét is. A HunLearner korpuszban szereplő szövegek a 2.3. bekezdésben említett követelmények szerint készülnek, továbbá a nyelvtanulók több lényeges adatát is rögzítjük, bár ezeket az adatokat egyelőre nem vontuk be az elemzésbe. A korpusz szélesebb körű feldolgozása esetén ezeket a metaadatokat is szeretnénk figyelembe venni. A nyelvtanulói hibák kódolása azért fontos, mert ez teszi lehetővé a lexikográfus számára a nyelvtanulók nehézségeinek szisztematikus detektálását, és ezáltali figyelembe vételét a szótárkészítés folyamatában. De Cock és Granger (2005: 79–80) két nagy csoportra osztja a nyelvtanulói hibákat aszerint, hogy milyen jellegű információval szolgálnak a lexikográfus számára. Az egyik csoportba a helyesírási, a lexikai, a lexiko-grammatikai, valamint a regiszterbeli hibákat sorolja, míg az anyanyelvi nyelvhasználattól eltérő gyakoriságú használat problémáit külön kategóriába tartozóként kezeli. Mindezek az információk Rundell (1999: 47) alapján két formában tükröződhetnek az egynyelvű nyelvtanulói szótárban: implicit és explicit módon. Implicit formában közvetíti a szótár a nyelvtanulói korpuszból kinyert információt akkor, ha a problémásnak tekinthető nyelvi sajátságot igyekszik jól érthető, alapos magyarázattal bemutatni, ezzel segítve a nyelvtanulót a helyes használat felé. Explicit módon törekszik a szótár az adott nyelvtanulói hiba kiküszöbölésére, amennyiben nem csupán a korrekt használatot mutatja be, de egyes, magas frekvenciájúnak ítélt hibás alakok esetében explicit módon fel is hívja a figyelmet a problémára.

5. Összefoglalás

A magyar tanulói korpusz létrehozása, fejlesztése és a kutatásokban történő felhasználása egy új és igen hasznosnak ígérkező szempontot jelent a magyar mint idegen nyelv vizsgálatában. Elméleti megfigyeléseken túl számos gyakorlati haszna is lehet, hiszen például egynyelvű szótárak elkészítéséhez egyedülálló segítséget tud nyújtani, de a távlati felhasználási lehetőségek között még a tanulói szövegek automatikus javítása is szerepelhet. A tanulói korpusz véleményünk szerint jól integrálható az eddigi kutatási gyakorlatba, hiszen ahogy azt Sylviane Granger, a korpusznyelvészet egy elismert kutatója is megjegyzi, a korpusz általában inkább csak kiegészíti és nem helyettesíti az eddig használt adatforrásokat (Granger et al. eds. 2002: 4).

Irodalom

Berwick, G. – Horsfall, P. 1996. Making Effective Use of the Dictionary. *PATHFINDER* 28. Bedfordbury: Centre for Information on Language Teaching and Research.

Cowie, A.P. 1999. *English Dictionaries for Foreign Learners: A History.* Oxford: Oxford University Press.

De Cock, S. – Granger, S. 2005. Computer Learner Corpora and Monolingual Learners' Dictionaries: the Perfect Match. *Lexicographica* 20. 72–86.

Dickinson, Markus–Ledbetter, Scott 2012. Annotating Errors in a Hungarian Learner Corpus. *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012).* Istanbul, Turkey. http://jones.ling.indiana.edu/~mdickinson/papers/dickinson-ledbetter12.pdf

Durst Péter 2010. A magyar mint idegen nyelv elsajátításának vizsgálata – különös tekintettel a főnévi és igei szótövekre, valamint a határozott tárgyas ragozásra. Bölcsészdoktori értekezés. Kézirat. Pécs

Granger, Sylviane 2002. A Bird's-eye View of Computer Learner Corpus Research, In: Granger S. – Hung J. – Petch-Tyson S. ed(s). *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching.* Amsterdam & Philadelphia, Benjamins, Language Learning and Language Teaching 6, p. 3-33.

H. Gouws, Rufus 2004. Monolingual and Bilingual Learners' Dictionaries. *Lexikos* 14. 264–274.

Hanah, Jirka – Rosen, Alexandr – Škodová, Svatava–Štindlová, Barbora 2010. LAW IV,10 *Proceedings of the Fourth Linguistic Annotation Workshop. Association for Computational Linguistics Stroudsburg*, PA, USA. 11–19.

Holi Ali, H. I. 2012. Monolingual Dictionary Use in an EFL Context. *English Language Teaching* 5/7. 2–7.

Jantunen, Jarmo Harri 2011. Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. Lähivõrdlusi. Lähivertailuja 21. Tallinn, *Estonian Association for Applied Linguistics (EAAL)*, 86–105.

Lee, S. 1998. Compiling a Monolingual Learner's Dictionary on Corpus Linguistic Principles: the Case of YLDCK., in EURALEX ,98 PROCEEDINGS. 453–457. http://www.euralex.org/elx_proceedings/Euralex1998_2/Sangsup%20LEE%20Compiling%20a%20Monolingual%20Learners%20Dictionary%20on%20Corpus%20Linguistic%20Principles%20the%20Case%20of%20YLDCK.pdf

Rundell, M. 1999. Dictionary use in production. *International Journal of Lexicography* 12/1. 35-53.

Selinker L. 1972. Interlanguage. *IRAL* 10, 209-230.

Summers, D. 1993. *Longman English Dictionary of Language and Culture.* Harlow, Essex, England: Longman

Szabó Martina Katalin 2012. A *bárki* és az *akárki* névmások fordítási kérdéseinek vizsgálata a magyarról oroszra történő fordítás tükrében. „A Tudomány Támogatásáért a Dél-Alföldön" Alapítvány és a Magyar Tudományos Akadémia Szegedi Akadémiai Bizottságának közös pályázatára írt, díjazott pályamunka.

Szirmai Mónika 2005. *Bevezetés a korpusznyelvészetbe.* Budapest: Tinta Kiadó
Trón Viktor – Németh László – Halácsy Péter – Kornai András – Gyepesi György – Varga Dániel 2005. Hunmorph: open source word analysis. In: *Proceedings of ACL*. Prága, Csehország: Association for Computational Linguistics.
Vincze Veronika 2014. Valency frames in a Hungarian corpus.*Journal of Quantitative Linguistics* 21/2. 153-176.
Vincze Veronika – Zsibrita János – Durst Péter – Szabó Martina Katalin 2013. HunLearner: a magyar nyelv nyelvtanulói korpusza. In: Tanács Attila – Vincze Veronika (szerk.): *IX. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem. 97–105.
Zsibrita János – Vincze Veronika – Farkas Richárd 2013. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In*: Proceedings of RANLP 2013*. Hissar, Bulgaria. 763–771.Девель, Л.А. 2004. Репрезентативность корпусов английского языка (данные учебных одноязычных словарей), in *Труды международной конференции „Корпусная лингвистика – 2004".* Санкт-Петербург, Изд-во Санкт-Петербургского ун-та. 131–137. http://www.corpora.phil.spbu.ru/Works2004/Devel_art.pdf

Az elemzésben szereplő tankönyvek

Durst Péter 2004. *Lépésenként magyarul. Első lépés.* Szeged: Szegedi Tudományegyetem
Durst Péter 2012. *Hungarian the Easy Way 1.* Szeged: Design Kiadó
Durst Péter 2013. *Hungarian the Easy Way 2.* Szeged: Design Kiadó
Erdős József – Prileszky Csilla 2002. *Halló, itt Magyarország ! I.* 4. kiadás. Budapest: Akadémiai Kiadó
Erdős József 2007. *Új színes magyar nyelvkönyv.* Budapest: Balassi Intézet
Hlavacska Edit – Hoffmann István – Laczkó Tibor – Maticsák Sándor 1996. *Hungarolingua 1.*,2. kiadás. Debrecen: Debreceni Nyári Egyetem
Szita Szilvia – Pelcz Katalin 2013. *MagyarOK 1.* Pécs: Pécsi Tudományegyetem