Antal Zsófia

A *KorSzak* tanulói korpusz bemutatása és a *magyar* nemzetiséget jelölő melléknév vizsgálata

1. Bevezetés

A tanulói korpuszok nagyon fontos szerepet játszhatnak egy adott nyelvet idegen nyelvként tanulók nyelvelsajátításának vizsgálataiban. Az elérhető korpuszok és a szövegelemző szoftverek lehetővé teszik, hogy nagy nyelvi mintán szisztematikusan tanulmányozhassuk, miként sajátítják el az új nyelvet a tanulók. Jelen írás a Korpusznyelvészeti és Szakmódszertani Munkacsoport (Baumann et al. 2020: 35) dinamikus *KorSzak* tanulói korpuszát ismerteti részletesebben, feltárja a korpuszban rejlő kutatási lehetőségeket, valamint bemutat egy lehetséges vizsgálatot, jelesül a *magyar* nemzetiséget jelölő melléknévhez kapcsolódó leggyakoribb főnévi kollokációs partnereket elemzi.

2. Tanulói korpuszok

A tanulói korpuszkutatás a korpusznyelvészet meglehetősen fiatal, de rendkívül dinamikus ága, amely az 1980-as évek végén, az 1990-es évek elején kezdett önálló tudományágként megjelenni. A tanulói korpuszok olyan írott és szóbeli megnyilvánulások szövegeinek elektronikus gyűjteményei, amelyeket egy adott nyelvet idegen nyelvként tanulók produktumaiból gyűjtöttek össze (Granger 2004: 124, Szirmai 2005: 34). Nesselhauf (2005: 40) a tanulói korpusz definícióját kiegészíti a szisztematikusság fogalmával is, mellyel arra utal, hogy a korpusz szövegeit megadott kritériumok alapján választják ki (mint például a nyelvtanuló első nyelve, szintje stb.). Nesselhauf (2005: 40) rámutat arra is, hogy a tanulói korpuszok az anyanyelvi korpuszokkal ellentétben irányítottságuk miatt másként autentikusak (vö. Sinclair 1996), hiszen nem az anyanyelvi beszélők természetes nyelvhasználatát tükrözik. A tanulói korpuszoknál a hangsúly a spontán nyelvi megnyilvánulásra kerül. Véleménye szerint a legkevésbé kontrollált, korpuszkompatibilis szövegek az esszék (ahol csak a téma van megadva), illetve a szóbeli interjúk (Nesselhauf 2005: 40).

Az egyik leghíresebb és legismertebb tanulói korpusz, az *International Corpus of Learner English (ICLE)*. A korpusz B2-es és C1-es szintű nyelvtanulók angol nyelvű esszéinek gyűjteménye. Építése 1990-ben, Sylviane Granger, a Leuveni Egyetem professzora kezdeményezésére indult, és egy közel harminc éves együttműködés eredménye számos egyetem között. Az ICLE korpuszépítés elsődleges célja az volt, hogy az alkalmazott nyelvészet és a számítógépes technológia segítségével alaposan megvizsgálják az idegennyelv-tanulók internyelvét, azaz azt az önálló összetett nyelvi rendszert, amelyet a nyelvtanuló a nyelvelsajátítás során alkot, és amelynek bizonyos elemei az anyanyelvből, más elemei pedig a célnyelvből származnak, de tartalmazhatnak olyan elemeket is, amelyek sem az anyanyelvre, sem a célnyelvre nem terjednek ki (Selinker 1972; Granger 1993: 57). A 2002-es első és a 2009-es második kiadás megjelenése óta a korpuszt nemzetközi kutatási projektek széles körében használják, és kulcsszerepet játszik mind a mai napig a tanulói korpuszkutatások népszerűsítésében. A korpusz első változata 2,5 millió szót[1] tartalmazott olyan nyelvtanulóktól, akik 11 különböző anyanyelvi háttérrel rendelkeztek. A második

[1] Jelen tanulmány nem fejti ki a *szó* és a *token* definícióját, hanem az adott szakirodalom terminusát alkalmazza.

változat mind a szavak, mind a nyelvi háttér tekintetében nagyobb volt; 3,7 millió szó és 16 különböző nyelvi háttér. A jelenlegi verzió két lényeges dologban különbözik a korábbiaktól: még nagyobb, mint a korábbi változatok; meghaladja az 5,5 millió szót, és 25 különböző anyanyelvi háttérrel rendelkező nyelvtanulóktól származik, valamint eltérően a korábbi két kiadástól, melyeket CD-ROM-okon adtak ki, a harmadik kiadás webes felületen[2] található meg, amely könnyebb és rugalmasabb hozzáférést tesz lehetővé, és az új alkorpuszok folyamatos beillesztése is könnyedén megvalósítható (Granger–Dupont–Meunier–Naets–Paquot 2020).

Napjaink másik legnagyobb angol nyelvű (tanulói) korpusza[3], a *Trinity Lancaster Corpus (TLC)*, melynek érdekessége, hogy angol anyanyelvű (L1) és angol mint idegen nyelvű (L2) beszélők közötti szóbeli interakciók átírt anyagait tartalmazza. A TLC a Lancasteri Egyetem és a Londoni Trinity College együttműködésével jött létre azzal a céllal, hogy egy olyan nagyméretű korpuszt hozzanak létre, amely különböző nyelvészeti és nyelvi kutatások eszközévé válhat. A korpuszban használt adatokat 2012 és 2018 között gyűjtötték a Londoni Trinity College által kifejlesztett és lebonyolított *Graded Examinations in Spoken English* vizsga keretében. 4,2 millió tokent tartalmaz, melyeket az L2-es vizsgázók és az L1-es vizsgáztatók között lezajlott vizsgabeszélgetésekből, szóbeli interakciókból jegyeztek le és írtak át. Az L2-es adatok több mint 2000 különböző nyelvi és kulturális háttérrel, valamint nyelvi szinttel rendelkező adatközlőtől származnak. A nyelvi mintákon kívül a korpusz további metaadatokat is tartalmaz az L2-es beszélőkről, melyeket kérdőívek segítségével gyűjtöttek össze (Gablasova–Brezina–McEnery 2019).

E tanulói korpuszok számos kutatást és tudományos munkát inspiráltak, illetve folyamatosan segítik az idegennyelvtanításban a mérést és értékelést, valamint a tananyagfejlesztési munkákat. Az angol nyelvű tanulói korpuszok száma folyamatosan növekszik, hiszen a korpuszkutatási módszerek terjedésének köszönhetően és a korpuszok gyakorlati jelentőségét felismerve egyre többen vágnak bele tanulói korpuszépítésbe, akár csak saját használatra is.

Az angolon kívül természetesen más nyelveken is születtek tanulói korpuszok, melyek közül a berlini Humboldt Egyetemen épített német nyelvű hibakódokkal ellátott *Fehlerannotiertes Lernerkorpus (Falko)*[4] talán az egyik legismertebb. A korpusz nyelvtanulók német nyelvű esszéit, leveleit, szépirodalmi írásait, tudományos írásait, folyóiratcikkeit és könyvismertetéseit tartalmazza kezdőtől haladó szintig, valamint számos anyanyelvi alkorpuszt is magába foglal, melyben az egyetem német anyanyelvű hallgatóitól gyűjtöttek anyagokat (Reznicek et al. 2012).

Szintén hibakódokkal ellátott és szélesebb körökben is ismert a cseh *The Learner Corpus of Czech as a Second Language (CzeSL)*[5], mely az első olyan hangzó és írott szöveget és több alkorpuszt (pl. orosz, romani és vietnámi L1) tartalmazó, több nyelvi szintű korpusz, amely flektáló nyelvet vizsgál, valamint többrétegű hibakódolási módszerrel dolgozik (Hana–Rosen–Škodová–Štindlová 2010; Rosen 2016). Itt meg kell jegyezni, hogy a *KorSzak* tanulói korpusz munkacsoport szintén tervezi különböző L1-es alkorpuszok létrehozását a későbbiekben a korpuszépítés folyamán.

Érdemes még megemlíteni a 2007-es *International Corpus of Learner Finnish (ICLFI)*[6] korpuszt is, mely az egyik legnagyobb, balti finn nyelvet feldolgozó tanulói korpusz. A közel egymillió szavas korpusz anyagát 22 különböző anyanyelvű, kezdő, középhaladó és haladó szintű nyelvtanulótól gyűjtötték össze, esszék, elbeszélések, naplóbejegyzések formájában. A

[2] https://corpora.uclouvain.be/cecl/icle/home letöltve: 2021.09.15.
[3] http://corpora.lancs.ac.uk/trinity/search letöltve: 2021.09.20.
[4] https://korpling.german.hu-berlin.de/falko-suche/ letöltve: 2021.09.20.
[5] http://utkl.ff.cuni.cz/dokuwiki/doku.php?id=czesl:czesl letöltve: 2021.09.20.
[6] https://korp.csc.fi/korp/#?cqp=%5B%5D&corpus=iclfi&stats_reduce=word letöltve: 2021.09.05.

szövegeken kívül a korpusz metatextuális információk sorát is tartalmazza az egyes változókról; így például a tanulók koráról és anyanyelvéről, az adatgyűjtő nyelvtanár anyanyelvéről, és a szöveg műfajáról (Jantunen 2011, Jantunen–Brunni 2013: 237). Ez a korpusz nemcsak azért érdekes a számunkra, mert a finn nyelv morfológiailag és tipológiailag is hasonlít a magyarra, hanem mert egy olyan együttműködés eredménye, amely inspirálta a *KorSzak* tanulói korpusz építését is. (Különböző országok egyetemein oktató finn mint idegen nyelvi tanárok gyűjtötték össze a nyelvtanulók írásos produktumait.) Egy párhuzamot kiemelve: a *KorSzak* tanulói korpusz is különböző – magyarországi és külföldi – oktatási intézményekben tevékenykedő magyar mint idegen nyelvi tanárok együttműködésében épül.
Magyar mint idegen nyelvi tanulói korpuszokban nem igazán bővelkedünk. 2012-ben megjelent egy tanulmány, melyben az Indianai Egyetem két kutatója számolt be korpusznyelvészeti vizsgálatairól (Dickinson–Ledbetter 2012). Minikorpuszuk 14 (9 kezdő, 1 középhaladó és 4 haladó) magyarul tanuló hallgató 10–15 soros naplóbejegyzéseit tartalmazza, melyek témáit a tanulók maguk választották. Céljuk a magyar nyelv vizsgálata volt, de a tanulmány inkább a hibakódolás egy lehetséges megközelítésével foglalkozik. A szövegek annotálását és hibajavítását is manuálisan végezték, a hibakódok rendszere nincs alaposan kidolgozva, és ugyan az anyagok feldolgozásában magyar nyelvű lektor segítségét is igénybe vették, a publikált tanulmányba mégiscsak bekerültek helyesnek vélt helytelen mondatok, mely a vizsgálati eredmények relevanciáját igencsak aláássa.

A magyar mint idegen nyelv elemzéséhez készült publikált tanulói korpusz a *HunLearner*[7], mely a Szegedi Tudományegyetem kutatóinak projektjeként jött létre. A korpusz első két alkorpusza a Zágrábi Egyetem 35 magyar szakos hallgatójának írásbeli beadványainak anyagát tartalmazza, melyek a magyar nyelv nehézségeiről, illetve a külföldi munkavállalásról szólnak. Hosszúságuk kb. 1500 karakter, egy óra alatt készültek segédeszközök (szótár, nyelvkönyv) nélkül elektronikusan, magyar billentyűzettel. Az adatközlő hallgatók nagy többségének magyar nyelvtudása B1-es szintű volt, de a többiek is részt vettek legalább egy éves magyar nyelvű képzésben. A korpusz később gyűjtött anyagai is követték a kezdetekben lefektetett alapelveket, a választható témák pedig a következők voltak: „Egy szimpatikus ember” vagy „Magyarországról és a magyarokról.” Az utolsó publikációs adatok alapján a korpusz 1427 mondatot és mintegy 22 000 tokent tartalmaz. A korpusz ismertetése mellett a kutatók a tanulmányaikban vizsgálati eredményeiket is bemutatták: a korpusz anyagában a főneveknél megfigyelhető morfológiai hibákat, a határozott tárgyas ragozás használatának egyes jellemzőit, és összehasonlították MID-tananyagok olvasmányainak szövegét a HunLearner szövegeivel, valamint előrevetítették a végső céljukat, azaz, hogy vizsgálati anyagot teremtsenek egy egynyelvű magyar nyelvtanulói szótár elkészítéséhez (Durst–Szabó–Vincze–Zsibrita 2013, 2014).

3. A KorSzak tanulói korpusz

A 2020 februárja óta épülő dinamikus *KorSzak* tanulói korpusz azzal a céllal jött létre, hogy meghatározott alapelvek mentén folyamatosan bővülő, jól kereshető, jelentős méretű, nyilvános adatbázist hozzunk létre, mely a nyelvészek, a hungarológusok és a MID tanításával foglalkozó szakemberek kutatásai és vizsgálatai számára gazdag forrásanyag lehet. A gyűjtésben eddig Antal Zsófia (PTE), Baumann Tímea (PTE), Erdősi Vanda (PTE), Gergely Viktória (Konstantin Filozófus Egyetem), Lesznicskova Liljana (Szófiai Egyetem), Pelcz Katalin (PTE), Schmidt Ildikó (KRE), Szita Szilvia (Strasbourgi Egyetem) és Zsolcsák-

[7] https://rgai.inf.u-szeged.hu/node/167 (letöltve: 2021.09.29.)

Dimitrova Edina (Szófiai Egyetem) működött közre, de szeretettel várjuk további munkatársak jelentkezését, akik szívesen részt vennének ebben a projektben.

A KorSzak tanulói korpusz két nagy részre különül el; írásbeli és szóbeli nyelvi produktumokra. A folyamatosan bővülő nyelvi adatbázis jelenleg 115 nyelvtanuló 1295 nyelvi produktumát tartalmazza közel 200 000 szószámmal. A cikk megjelenésének pillanatában az írásbeli korpusz a jelentősebb anyag, a maga 1099 írásos szövegalkotásával és közel 125 000 szószámával.

A korpusz írásbeli szövegalkotások alkorpusza a magyart idegen nyelvként tanulók írott beadványaiból épül fel. A szövegek egy része kézírással készült, melyeket úgy gépeltünk be, hogy tartalmilag és – amennyire lehetséges – formailag is hűek maradjanak az eredeti változathoz, tehát tartalmilag és formailag is vizsgálhatók legyenek kutatási szempontból. A tanulói szövegek másik része a digitális oktatás bevezetésének folyományaként elektronikus formában került beadásra, majd pedig változatlanul a korpuszba. Azért, hogy elkerüljük a digitalizálás során esetlegesen fellépő hibalehetőségeket (elgépelések, kihagyások a begépelések során stb.), a jelenléti oktatás visszatérte mellett továbbra is biztosítjuk azokat az online platformokat, amelyeken keresztül a tanulók elektronikus formában adhatják le munkáikat.

Egy jól kereshető adatbázis feltétele a szövegek megfelelő kódolása. A *KorSzak* tanulói korpusz minden szövege egy tizenhárom változós kóddal van ellátva: *PTE NOK_SU_2021_MagyarOK_A2_8 F_HF_CZ_cseh_F_27_1_53*.

Nézzük meg, hogy milyen adatok állnak rendelkezésünkre a korpusz kódrendszeréből!

A kódrendszer első változójából megtudhatjuk, hogy melyik oktatási intézményben került sor a gyűjtésre. A fenti kódban található *PTE NOK* rövidítés a Pécsi Tudományegyetem Általános Orvostudományi Kar Nemzetközi Oktatási Központját jelöli, de az adatbázisban találhatunk szövegeket a Strasbourgi Egyetem (*STR*) hallgatóitól és magántanítványoktól (*PRIVAT*) is.

A második változó a képzési formára utal. A mintakódban megadott *SU* rövidítés a Pécsi Tudományegyetem Általános Orvostudományi Kar Nemzetközi Oktatási Központjának 60, illetve 120 kontaktórás intenzív nyári egyetemét takarja, de ennél a változónál még a következő képzési formákkal is találkozhatunk a korpuszban:

– Magyar 1., Magyar 2. – a Strasbourgi Egyetem Magyar Tanulmányok Tanszékének magyar képzései;
– PREP – a Pécsi Tudományegyetem Nemzetközi Oktatási Központjának két féléves, 800 kontaktórás előkészítő képzése (1 kontaktóra = 45 perc);
– PRIV – egyéni oktatás, igény szerinti óraszámban;
– SEM – a Pécsi Tudományegyetem Nemzetközi Oktatási Központjának 12 hetes, 48 kontaktórás általános nyelvkurzusa;
– SH INTENSIVE – a Pécsi Tudományegyetem Nemzetközi Oktatási Központjának 2 hetes, 48 kontaktórás intenzív kurzusa Stipendium Hungaricum ösztöndíjas hallgatóknak.

A harmadik változó azt az évet (*2021*) vagy tanévet (*2020/2021*) jelöli, amikor a rögzített szöveg készült. A negyedik változó a tankönyvet jelöli (*MagyarOK*), amelyből a nyelvtanuló tanul. Itt meg kell jegyezni, hogy a *KorSzak* korpusz építésének alapkritériuma, hogy a korpusz adatközlői a *MagyarOK* tankönyvcsalád köteteiből tanuljanak különböző képzési formában és különböző nyelvi szinten, aminek oka a *MagyarOK* pedagógiai korpusz és *KorSzak* tanulói korpusz összehasonlíthatósága. Abban az esetben, ha ez a kitétel nem teljesül, természetesen még bekerülhet anyag az adatbázisba, de a tematikának és a nyelvi szintnek mindenképpen azonosnak kell lennie a többi felgyűjtött anyagéval.

Az ötödik változóból a nyelvtanuló nyelvi szintjéről kapunk információt, mely besorolás megegyezik a Közös Európai Referenciakeret nyelvi szintezésével (*A1, A2, B1, B2, C1*). Jelen korpusz 417 A1-es, 354 A2-es, 227 B1-es és 101 B2-es írott szövegproduktumot tartalmaz. A

hatodik változó az adott tankönyv fejezetszámát (*8F*), a hetedik a feladattípust (*HF* – házi feladat, *OM* – órai munka), a tizenkettedik pedig a feladat sorszámát (*1*) jelöli, azaz megtudhatjuk, hogy az írott szöveg a *MagyarOK* tankönyvcsalád melyik kötetének melyik fejezetéből származik, illetve hogy mi volt az esszé témája.

A nyolcadik és kilencedik változóból a nyelvtanulók állampolgárságáról és anyanyelvéről kapunk információt. A cikk megjelenésekor az adatközlők 49 nemzetiséget képviselnek, és anyanyelvként 37 különböző nyelvet jelöltek meg. Legnagyobb számban arab, vietnámi, francia és angol nyelvi háttérrel rendelkeznek. Ugyanaz a nyelv több változatban is szerepel: pl. amerikai angol, brit angol, dél-afrikai angol, ghánai angol, kanadai angol, nigériai angol; portugál, brazil portugál.

A tizedik változó a nemet jelöli (*F* – nő, *M* – férfi). A korpusz adatközlői nemi eloszlásuk alapján nem mutatnak nagy különbséget, hiszen 60 nő és 55 férfi írásos anyagait tudjuk vizsgálni. A tizenegyedik változóból a nyelvtanulók életkoráról szerezhetünk információt. A legfiatalabb adatközlő 19, a legidősebb 77 éves. A korpusz adatközlőinek több mint a fele (61%-a) fiatal felnőtt (20–30 év). Ez nem meglepő, hiszen felsőoktatási intézményekben szervezett nyelvkurzusok nyelvtanulóitól gyűjtöttük az anyagokat.

Az adatbázis anonim, az adatközlők nevét nem közölhetjük, ezért a nyelvtanulókat sorszámmal jelöltük, mely az utolsó helyen szerepel a kódban. Az írásbeli beleegyező nyilatkozatban az adatközlők egyéb személyes adataira is rákérdezünk: iskolai végzettség, L1-en kívüli beszélt nyelvek, valamint a képzési forma mellett a korpusz metaadatai között szerepel az is, hogy az adott képzés online vagy tantermi keretek között valósult meg.

Az írásbeli szövegeket tartalmazó alkorpusz az összehasonlító kutatások tárháza. Lehetőségeket ad olyan összehasonlítások elvégzéséhez is, melyekben azt vizsgáljuk, hogy a különböző képzési formákban részt vevő diákok munkái ugyanazon a szinten, ugyanabban a témában miben különböznek. Például más-e, és ha igen, miben más egy olyan előkészítős hallgatónak az írásosos munkája, aki a *MagyarOK* A1+ kötetet emelt óraszámban 8 hét alatt fejezi be, mint annak a hallgatónak, aki ugyanezt az anyagmennyiséget szemeszteres képzésben heti 2x90 percben két szemeszteren keresztül 24 hét alatt abszolválja? Vajon ugyanazt a lexikai állományt fogja-e használni a két csoport? Jelentkeznek-e különbségek a lexika nagyságában, a szórendi kérdésekben? Vizsgálat tárgyát képezheti az is, hogy miben különböznek a célnyelvi környezetben magyarul tanuló hallgatók írásos produktumai a nem célnyelvi környezetben tanulóéktól. Van-e a szövegekben kimutatható különbség? Tovább gondolva: eredményez-e minőségi változást az, ha valaki például online egyéni oktatásban tanulja a magyar nyelvet, majd pedig részt vesz egy magyarországi intenzív nyári egyetemen vagy hosszabb magyarországi képzésen? Izgalmas kutatási eredményeket hozhat ennek a két képzési módnak az összehasonlítása a magyar nyelv elsajátításának eredményességét illetően. A szociolingvisztikai jellemzők (kor, nem, iskolai végzettség) hatása a nem kivételével a tanulói korpuszkutatásban eddig nem kapott nagy figyelmet (Gablasova–Brezina–McEnery 2017), holott az életkorról, az iskolai végzettségről számos nyelvtanulói korpusz rögzít adatokat. Az iskolázottsági és az életkori adatokat is felhasználhatjuk összehasonlító vizsgálatokra. Mi a különbség például a fiatal és az érettebb nyelvtanulók nyelvhasználata között? Milyen nyelvi mintákat lehet beazonosítani a kognitív és nyelvi érettség és/vagy az iskolai végzettség hatására például a lexikaválasztás, a grammatikai komplexitás és a pragmatikai képességek területén?

A tanulmány 4. fejezetében jómagam is az írásos korpuszt vizsgálom, de néhány sor erejéig térjünk egy kicsit vissza a *KorSzak* tanulói korpusz szóbeli beszédproduktumokat rögzítő alkorpuszára!

A szóbeli beszédproduktum alkorpusz jelenleg a Pécsi Tudományegyetem Nemzetközi Oktatási Központ előkészítős hallgatóinak szóbeli fejezetzáró vizsgáinak, félévvégi és évvégi vizsgáinak, prezentációinak videófelvételeit, valamint az egyéb kurzusokra beadott, videóra

vett szóbeli házi feladatok felvételeit tartalmazza. A korpuszok növekvő száma ellenére még mindig viszonylagosan kevés a hangzó szövegeket reprezentálók száma, pedig a beszélt nyelvi korpuszok vizsgálatával olyan leírásokat kaphatunk a nyelv használatáról és elsajátításáról és azok módjáról, amelyek eltérhetnek az írásművek vizsgálati eredményeitől. Ellentétben az írással, a spontán szóbeli megnyilatkozások nem biztosítják a beszélgetőpartnerek számára azt a lehetőséget, hogy a későbbiekben újra megtekinthessék és átszerkeszthessék mondanivalójukat; gyors javításra csupán csak a szóbeli interakciók alatt van lehetőségük, melyhez rugalmasságra és gyors reakcióra van szükség. Megfigyelhetők a habozások, szünetek, egyéb megakadásjelenségek, azaz a spontán beszédben fellépő különböző hibák. A korpuszgyűjtemény e része kiváló kutatási terepe lehet a kiejtéssel kapcsolatos vizsgálatoknak is. A beszéd különleges helyet foglal el a pszicholingvisztikai kutatásokban is. A vizsgálati eredmények betekintést nyújthatnak a nyelvhasználat, a nyelvelsajátítás és a kognitív folyamatok közötti kapcsolatokba. A szóbeli interaktív kommunikáció gazdag adatokat szolgáltat a korpusz alapú pragmatikai kutatásokhoz, melyek az interperszonális tényezőket vizsgálják a nyelvhasználatban.

4. A KorSzak tanulói korpusz egy lehetséges vizsgálata

A sokszempontú keresésnek alávethető digitalizált korpuszok többek között kiváló lehetőséget adnak a kollokációk, azaz a statisztikailag kimutatható gyakorisággal előforduló szókapcsolatok kutatására is. A korpuszalapú keresés lehetővé teszi, hogy a keresett szavakat több szavas szövegkörnyezetükkel együtt kilistázzuk, elemezzük, és rendszerezett példatárat készítsünk belőlük. Az egyik legkézenfekvőbb vizsgálati terület a nyelvtanulók és az anyanyelvi beszélők nyelvhasználatának összehasonlítása magyar nyelvű korpuszok segítségével (pl. *Magyar Nemzeti Szövegtár*, *Hunglish*, a Sketch Engine-en elérhető *huTenTen12* és a didaktizált *MagyarOK* korpusz).

Az alábbiakban három különböző korpusz (a *huTenTen12*, a *MagyarOK*, a *KorSzak* írásos szövegalkotásokat tartalmazó alkorpusza) segítségével feltérképezem, hogy mely nemzetiséget jelölő mellékneveket használják a leggyakrabban a magyarok, és melyeket a magyar mint idegen nyelvet tanulók; majd pedig megvizsgálom, hogy a mindhárom korpuszban szereplő leggyakoribb nemzetiséget jelölő melléknévnek korpuszonként mi a két leggyakoribb főnévi kollokánsa.

4.1. A leggyakoribb nemzetiséget jelölő melléknevek a vizsgált korpuszokban

A MID tankönyvek tradicionálisan az első leckék egyikében ismertetik meg a nyelvtanulókkal a nemzetiségneveket. A nemzetiségnevek és a létige használatával a nyelvtanuló meg tudja határozni önmagát, saját magát és társait is el tudja „helyezni” a térképen (*Német vagyok. Francia vagyok, Párizsban élek. James amerikai, de most Magyarországon él.*). A nemzetiségnevek tehát lehetnek a névszói állítmány részei, valamint jelzőként (nemzetiséget jelölő melléknévként) is szerepelhetnek.

Ha végignézzük a képzési szabályok szerint csoportosított nemzetiségnevek listáját a *MagyarOK* tankönyvcsalád A1+ kötetében (második fejezet, 22. oldal), megfigyelhetjük, hogy ez a felsorolás – a földrajzi elhelyezkedésünk, Észak-Amerika dominanciája és részben terjedelemi korlátok miatt – Európa és Észak-Amerika központú. A nemzetiségnevek között a *brazil*, az *indonéz*, a *mongol*, az *egyiptomi*, az *indiai*, az *izraeli*, a *kínai* és a *japán* azok, amelyek más kontinenst reprezentálnak. Feltételezhetjük tehát, hogy az általunk kiválasztott két anyanyelvi korpuszban (*huTenTen12, MagyarOK*) is ugyanez a tendencia jelenik meg, de

vajon a tanulói korpuszban (*KorSzak*) is ezek a nemzetiségneveket jelölő melléknevek lesznek-e a leggyakoribbak?

Az első 50 leggyakoribb melléknévre először a Sketch Engine felületéről elérhető magyar nyelvű óriáskorpuszban, a *huTenTen12*-ben kerestem rá. A korpusz az internetes nyelvhasználatot reprezentálja, olyan szövegekből áll, melyeket 2012-ig publikáltak az interneten. 2 572 620 694 szót és 3 161 920 362 tokent tartalmaz. A gyűjtemény 50 leggyakrabban használt melléknéve a következő (1. ábra):

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	nagy	6,150,485	11	saját	1,966,868	21	egész	1,092,356	31	következő	884,736	41	politikai	729,795
2	olyan	5,653,248	12	amilyen	1,943,251	22	adott	1,092,271	32	régi	884,712	42	elmúlt	728,888
3	jó	4,941,229	13	másik	1,915,659	23	európai	1,034,551	33	megfelelő	874,237	43	nemzetközi	722,871
4	új	4,537,008	14	való	1,704,246	24	kevés	1,027,383	34	biztos	873,438	44	rossz	721,383
5	egy	3,524,466	15	fontos	1,454,081	25	magas	1,024,651	35	hasonló	825,223	45	kedves	717,269
6	magyar	3,345,890	16	teljes	1,217,842	26	késő	1,008,430	36	további	820,646	46	amerikai	710,095
7	ilyen	3,247,796	17	szép	1,179,301	27	képes	994,282	37	mai	813,062	47	erős	709,028
8	egyik	2,886,974	18	kettő	1,123,275	28	többi	987,486	38	különböző	810,738	48	utóbbi	704,600
9	kicsi	2,357,857	19	igaz	1,119,829	29	utolsó	970,881	39	szükséges	794,239	49	gazdasági	698,623
10	kis	2,081,594	20	hosszú	1,115,872	30	nehéz	920,511	40	című	743,722	50	érdemes	693,597

1. ábra A huTenTen12 korpusz 50 leggyakoribb mellékneve

Láthatjuk, hogy az általunk vizsgálni kívánt nemzetiséget jelölő melléknevek közül négy szerepel az 50-es listán; a *magyar* hatodikként 3 345 890, az *európai* huszonharmadikként 1 034 551 és az *amerikai* negyvenhatodikként 710 095 példányszámmal. A 100-as listában ötvenkilencedikként még megtalálhatjuk 666 215 példányszámmal a *német*et, illetve kilencvennegyedikként 452 433 példányszámmal az *angol*t.

A szintén a Sketch Engine felületéről elérhető 144 832 szó és 201 079 token számú *MagyarOK* pedagógiai korpusz (amely kifejezetten tanítási céllal készült; Szita 2020: 174–175) esetében a melléknevekre szűrt 50-es gyakorisági lista a következő (2. ábra):

	Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency		Lemma	Frequency
1	olyan	400	11	kis	121	21	hosszú	76	31	rengeteg	58	41	amerikai	46
2	jó	369	12	kedves	112	22	másik	73	32	való	56	42	helyi	45
3	magyar	320	13	fontos	105	23	szinonim	72	33	érdekes	56	43	szuper	44
4	amilyen	287	14	magas	90	24	finom	69	34	tilos	50	44	európai	44
5	nagy	283	15	régi	90	25	késő	68	35	adott	49	45	érdemes	41
6	új	279	16	német	89	26	kevés	67	36	francia	49	46	éves	40
7	kicsi	230	17	ilyen	84	27	friss	66	37	közösségi	47	47	rövid	40
8	egy	162	18	kedvenc	78	28	gyönyörű	61	38	szíves	47	48	híres	40
9	szép	152	19	nehéz	77	29	fáradt	59	39	drága	46	49	pozitív	40
10	egyik	132	20	angol	76	30	könnyű	59	40	üzleti	46	50	idős	40

2. ábra A MagyarOK pedagógiai korpusz 50 leggyakoribb mellékneve

320 példányszámmal a *magyar* nemzetiséget jelölő melléknév a harmadik helyen, 89 példányszámmal a *német* a tizenhatodik helyen, 76 példányszámmal az *angol* a huszadik helyen, 49 példányszámmal a *francia* a harminchatodik helyen, 46 példányszámmal az *amerikai* a negyvenegyedik, 44 példányszámmal az *európai* a negyvennegyedik helyen található. A 100-as listán még az *olasz* is szerepel: 32 példányszámmal a hatvanötödik helyen.

A *KorSzak* tanulói korpusz még nem érhető el korpuszkezelő és korpuszelemző szoftveren, ezért csak manuálisan vizsgálható. Első lépésként a korpusz anyagán keresőprogram segítségével végigfuttattam a korábbi korpuszok első öt-öt leggyakoribb nemzetiséget jelölő melléknevét (*magyar*, *európai*, *amerikai*, *német*, *angol*, *francia*) (1. táblázat). A keresési eredmények alapján a tanulói korpusz esetében a következő sorrendet kaptam: első helyen 351 példányszámmal a *magyar*, második helyen 91 példányszámmal az *angol*, harmadik helyen 66 példányszámmal a *francia*, negyedik helyen 45 példányszámmal a *német*, ötödik helyen 34 példányszámmal az *amerikai* 34, hatodik helyen 24 példányszámmal az *európai* szerepelt. Második lépésként rászűrtem a korpusz adatközlői nemzetiségeire, majd pedig végigfuttattam az ezeket jelölő mellékneveket a keresőben, melynek eredményeként a *vietnámi* nemzetiséget jelölő melléknév 77, a *japán* 44, az *olasz* és a *lengyel* fej fej mellett 33, a *koreai* pedig – hasonlóan az *európai*hoz – 24 alkalommal szerepel a korpuszban.

Megfigyelhetjük tehát, hogy a három eltérő méretű, eltérő adatforrású és eltérő céllal készült korpusz a nemzetiséget jelölő melléknevek gyakoriságában erőteljes hasonlóságot mutat. Egyedül a *KorSzak* tanulói korpusz esetében találunk eltérést, ahol a ranglistán a harmadik helyen a *vietnámi* melléknév található, de ez az eredmény talán annyira nem meglepő, hiszen míg az első két korpusz adatközlői magyar nemzetiségűek, addig a tanulói korpusz adatközlői különböző nemzetiségeket képviselnek, és egy adott nemzetiség létszámának dominanciája a nagy számok törvénye alapján szinte biztosra vehetően visszatükröződik a korpusz lexikájában is.

Gyakoriségi sorrend	*huTenTen12*	*MagyarOK*	*KorSzak*
1.	*magyar*	*magyar*	*magyar*
2.	*európai*	*német*	*angol*
3.	*amerikai*	*angol*	*vietnámi*
4.	*német*	*francia*	*francia*
5.	*angol*	*amerikai*	*német*

1. táblázat: A vizsgált korpuszok öt leggyakoribb nemzetiséget jelölő melléknéve

A következő alfejezetben a mindhárom esetben első helyen szereplő nemzetiséget jelölő melléknevet vizsgálom, ezért itt szeretném megjegyezni, hogy a különböző nemzetiséget jelölő melléknevek tipikus kollokációinak jelentésmezeje eltér egymástól. Az *angol*t és a *német*et jellemzően mindhárom korpuszban a nyelv és nyelvtudás vonatkozásában használjuk:

- *Az előadást angol nyelven a Petőfi 5-ös számú tanteremben lehet meghallgatni. Angol nyelven készülnek a dalok, nem elsősorban Magyarországra tervezünk.* (*huTenTen12*)
- *Amit elvárunk: minimum középfokú végzettség, legalább egy év releváns szakmai tapasztalat, felsőfokú angol nyelvtudás, számítógépes ismeretek, precíz munkavégzés és jó kommunikációs képesség. Tárgyalóképes szintű angol nyelvtudással rendelkezem, emellett német nyelvtudásomat is folyamatosan fejlesztem.* (*MagyarOK*)
- *Angol nyelvű szemináriumokra és előadásokra jár. Szeretem az angol nyelvet, mert szép és könnyű.* (*KorSzak*)
- *Pálfi Adél saját programja használatával tanítja és szeretteti meg a gyerekekkel a német nyelv tanulását. Törekszünk a tantárgyak színvonalas oktatására, a német nyelv olyan szintű elsajátítására, hogy a tovább tanuló diákjaink más idegen nyelvet tanulhassanak a középiskolában.* (*huTenTen12*)

- *Nyelvtanár vagyok: angol és német nyelvet tanítok felnőtteknek. Német szakon tanulok a szegedi egyetemen, de most Berlinben, a Humboldt Egyetemen töltök két félévet ösztöndíjasként.* (*MagyarOK*)
- *Fontos a német nyelvtudás is, mert sok munkához nagy szükség van rá. Sajnos Gázában csak angol és német tanfolyam van.* (*KorSzak*)

Ezzel szemben például az *amerikai* nemzetiséget jelölő melléknevet az anyanyelvi óriáskorpuszban politikai vonatkozásban használjuk:

- *Éppen ő, holott a mai napig Jackson az egyetlen amerikai elnök, aki (még a krízis előtt) visszafizette a teljes államadósságot. Az amerikai katonák a napokban vonultak ki végleg Irakból, a közel-keleti konfliktusnak azonban ezzel még korántsincs vége* (*huTenTen12*),

a pedagógiai és tanulói korpuszban pedig a kultúrával kapcsolatban:

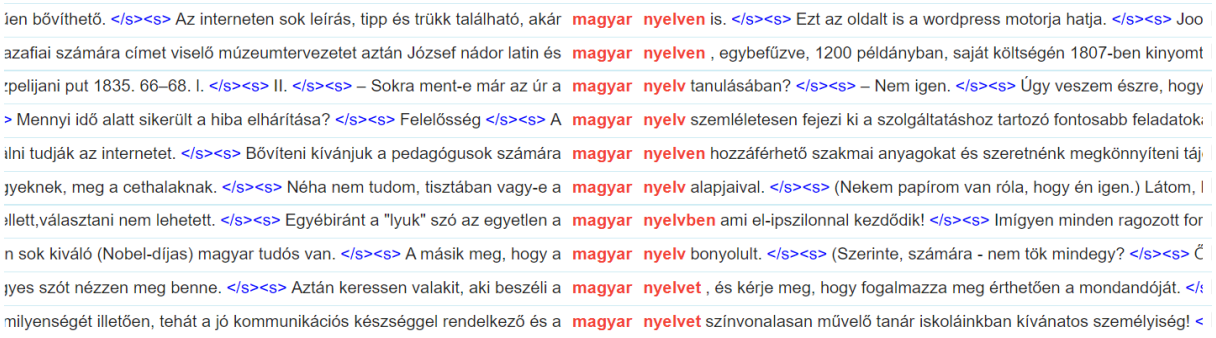
- *A Pulitzer-díjat, az amerikai újságírás legrangosabb díját nem kell bemutatnunk. Maria Callas amerikai operaénekes.* (*MagyarOK*)
- *Amerikai akciófilmeket nagyon szeretem. A német fiatalok menők akartak lenni és adoptálódtak az amerikai kultúrához és a nyelvhez is.* (*KorSzak*)

4.2. A leggyakoribb nemzetiséget jelölő melléknév két leggyakoribb főnévi kollokánsa a korpuszokban

Következő lépésként kiválasztottam a mindhárom korpuszban leggyakoribb nemzetiséget jelölő melléknevet, a *magyar*t, majd megvizsgáltam, hogy melyek a leggyakoribb főnévi kollokánsai a korpuszokban.

A *huTenTen12* óriáskorpusz és a *MagyarOK* pedagógia korpusz esetében a Sketch Engine konkordancia programja segítségével, szisztematikusan elemezve a konkordanciasorokat jutottam el az eredményekhez, míg a *KorSzak* tanulói korpusz elemzésénél manuális vizsgálatot alkalmaztam. A minivizsgálat bemutató jellege miatt minden esetben csak 10 konkordanciasort fogok szemléltetni, amelyek azonban a Sketch Engine szoftvernek köszönhetően kis számuk ellenére is reprezentatívak.

A *huTenTen12* korpusz gyűjteménye alapján a *magyar* nemzetiséget jelölő melléknév két leggyakoribb főnévi kollokánsa a *nyelv* 71 213 találattal (3. ábra) és a *kormány* 45 101 találattal (4. ábra).



3. ábra A huTenTen12 korpuszban a 'magyar' nemzetiséget jelölő melléknév leggyakoribb főnévi kollokánsa

4. ábra A huTenTen12 korpuszban a ’magyar’ nemzetiséget jelölő melléknév második leggyakoribb főnévi kollokánsa

A konkordanciasorok elemzésével megfigyelhetjük, hogy:
– a *magyar* és *nyelv* kollokánsok közvetlenül követik egymást, és nem ékelődik be közéjük másik lexikai elem (ha igen, akkor minősítőjelző: *a magyar hivatalos nyelv*, *a magyar nemzeti nyelv*, *a magyar irodalmi nyelv*);
– jellemzően határozott névelő áll előttük (*a magyar nyelv*, *a magyar nyelvet*, *a magyar nyelvben*);
– a *magyar* mellett a *nyelv* szó három leggyakoribb ragozott formája a tárgyragos *nyelvet* (*a magyar nyelvet és irodalmat magas fokra fejlesztették, a magyar nyelvet anyanyelvükként tisztelhetik, már nem beszélte a magyar nyelvet*), valamint a határozóragos *nyelven* (*magyar nyelven is kapható lesz, a tanítás magyar nyelven folyik, magyar nyelven végeztem*) és *nyelvben* (*a magyar nyelvben nagyon szoros kapcsolatban van, ilyesféle hangot ejtünk a magyar nyelvben a következő szavakban, a magyar nyelvben jól visszajön*);
– a *magyar nyelv* szókapcsolat gyakran birtokos szerkezet része (*a magyar nyelv ápolását, a magyar nyelv szótára, a magyar nyelv szerkezete*);
– a kollokációk gyakran állnak mondatkezdő pozícióban (*A magyar nyelv pl. ebben is egyedülálló: nálunk külön van piros és vörös szín. Magyar nyelven megjelent könyvek. A magyar nyelvben több nevét – köszméte, piszke – is használjuk a gyümölcsnek*), mondatvégi pozícióban viszont ritkán (*Az 1810-ben elindított nyelvújítási mozgalom mintegy 10 000 szóval gyarapította a magyar nyelvet. Békési László munkáját, amely ingyen hozzáférhető, magyar nyelven. Ha egyetértünk az elmélettel, miszerint a nyelv – bonyolult, komplex módon, de – meghatározza a gondolkodást, akkor elgondolkodtató, hogy vajon mit jelent ezen forma hiánya például a magyar nyelvben*).

A *magyar* és *kormány* kollokánsok hasonlóan a *magyar–nyelv* kollokációkhoz
– közvetlenül követik egymást, nagyon ritkán ékelődik be a két szó közé másik lexikai elem; ha igen, akkor jellemzően minősítőjelző (*magyar felelős kormány*, *magyar demokratikus kormány*, *magyar királyi kormány*);
– határozott névelő előzi meg őket (*a magyar kormány, a magyar kormányt, a magyar kormánynak*);
– a *kormány* két leggyakrabban használt ragozott formája a tárgyragos *kormányt* (*a felelősség a magyar kormányt terheli, magyar kormányt nevezett ki, természetvédők kérik a magyar kormányt*) és a genitívuszi-datívuszi toldalékkal ellátott *kormánynak* (*a magyar kormánynak nincs semmi jogi lehetősége, a magyar kormánynak hatalmas felelőssége van, a magyar kormánynak van főleg mozgástere*);
– mondatkezdő pozícióban előfordulnak (*A magyar kormány, hasonlóan az Európai Unió tagországaihoz, támogatná egy új, környezetvédelmi világszervezet létrehozását az ENSZ égiszén belül. A magyar kormány pedig asszisztál ehhez! A magyar kormány az*

egészségügyi szektor átalakítását kezdte meg az elmúlt évben.), mondatvégiben viszont csak ritkán (*Tehát innentől kezdődően nem tudni pontosan, hogy mit képviselt a magyar kormány. Ez a feltétele annak is, hogy világos helyzetelemzéssel és kérésekkel keressük meg ezúttal a magyar kormányt. Igen nehéz helyzetben volt 1991 februárjában az Antall József vezette magyar kormány.*)

A *MagyarOK* pedagógiai korpusz esetében a *magyar* nemzetiséget jelölő melléknév két leggyakoribb főnévi kollokánsa a *nyelv* (5. ábra) és az *étel* (6. ábra***ábra***); mindkét esetben 15–15 találattal.

5. ábra A MagyarOK pedagógiai korpuszban a 'magyar' nemzetiséget jelölő melléknév leggyakoribb főnévi kollokánsa

– A *magyar* és *nyelv* kollokánsok egy kivétellel (itt a magyar nemzetiséget jelölő melléknév nem a kollokánssal áll jelzői szerkezetben, hanem egy másik főnévvel: *magyar nyelvzseni*), mindig közvetlenül egymás mellett állnak;
– a kollokációt gyakran előzi meg határozott névelő (*a magyar nyelv*, *a magyar nyelvet*, *a magyar nyelvben*);
– a nyelv két leggyakoribb ragozott formája a tárgyragos *nyelvet* (*magyar nyelvet tanítok, szeretem a magyar nyelvet, megújította a magyar nyelvet*) és a határozóragos *nyelvben* (*a magyar nyelvben általános volt, a magyar nyelvben több igeidő volt*);
– mondatkezdő (*A magyar nyelvet tanulom.*) és mondatvégi pozícióban (*Szeretem a magyar nyelvet. Szereti a magyar nyelvet? Kazinczy Ferenc gazdagította és megújította a magyar nyelvet.*) is előfordulnak.

6. ábra A MagyarOK pedagógiai korpuszban a 'magyar' nemzetiséget jelölő melléknév második leggyakoribb főnévi kollokánsa

– A korpuszban a *magyar* és *étel* kollokánsok mindig egymás mellett állnak, nem ékelődik be közéjük másik lexikai elem;
– a kollokációt 6 alkalommal határozott névelő, 5 alkalommal a *tipikus*, 2 alkalommal az *ismert*, illetve egy alkalommal a *tradicionális* jelzők előzik meg (*John tipikus magyar ételeket szeretne enni. Milyen ismert magyar ételek vannak? Éttermünkben tradicionális magyar ételeket, különlegességeket és vegetáriánus ételeket is kínálunk.*);
– az *étel* kollokáns érdekessége, hogy ritka kivétellel mindig többes számú (*Sajnos, a magyar ételek nem mindig egészségesek. Miből készülnek a tipikus magyar ételek? Tehát a magyar ételek fűszeresek, és nagyon zsírosak szoktak lenni, de ellenben nagyon finomak.*), és gyakran áll tárgyraggal (*Soroljon fel három tipikus magyar ételt! Ismer tipikus magyar ételeket?*);
– a kollokációk mondatkezdő és mondatvégi pozícióban is szerepelnek (*A magyar ételek finomak, de sajnos túl zsírosak és sok a hús bennük, ezért általában olasz ételeket eszem. Írországban hiányoznak neki a szülei és a barátai, hiányzik, hogy nem magyarul beszél, hiányoznak neki a magyar ételek.*).

A *KorSzak* tanulói korpuszban a *magyar* nemzetiséget jelölő melléknév két leggyakoribb főnévi kollokánsa a *nyelv* és az *ember*. A *nyelv* 100, az *ember* 15 találattal szerepelt az adatbázisban.

– A tanulói korpusz példáiban a *magyar* és a *nyelv* kollokánsok mindig egymás mellett állnak, nem ékelődik be közéjük más lexikai egység;
– a *magyar nyelvű* kollokáció kivételével mindig határozott névelő szerepel a kollokáns párok előtt (*a magyar nyelv*, *a magyar nyelvet*, *a magyar nyelvben*);
– a nyelv lexikai elem legjellemzőbb ragozott formái a tárgyragos *nyelvet* (*gyakorlom a magyar nyelvet*, *szeretem a magyar nyelvet*, *tanulom a magyar nyelvet*), a helyhatározóragos *nyelvben* (*a magyar nyelvben van sok, a magyar nyelvben egy ábécé van, a magyar nyelvben hosszú kiejtés és rövid kiejtés van*), illetve a melléknévi szerkezet részeként álló *nyelvű* (*magyar nyelvű előadásokra is járok, néha kapok magyar nyelvű e-mailt, magyar nyelvű oldalak segítségével tanulok*);
– mondatkezdő (*A magyar nyelvben egy ábécé van. A magyar nyelv szép, de nem olyan könnyű. A magyar nyelvben van sok nagyon hosszú szó, például "megszentségteleníthetetlenségeskedéseitekért"*) és mondatvégi pozícióban is előfordulnak (*Szerintem nehéz a magyar nyelv. Sok házi feladatot ad a tanár, de ez nagyon jó, mert gyakorolom a magyar nyelvet. Most tanulom a magyar nyelvet*).

- *Soha nem kapok **magyar nyelvű** e-mailt.*
- *Szerintem nehéz a **magyar nyelv**.*
- *Sok házi feladatot ad a tanár, de ez nagyon jó, mert gyakorolom a **magyar nyelvet**.*
- *A **magyar nyelvben** egy ábécé van.*
- *Szeretem a **magyar nyelvet**, mert sok magyar dolog érdekel, de írni elég nehéz.*
- *Most tanulom a **magyar nyelvet**.*
- *A **magyar nyelvben** van sok nagyon hosszú szó, például "megszentségteleníthetetlenségeskedéseitekért".*
- *A **magyar nyelvben** hosszú kiejtés és rövid kiejtés van.*
- *Angol és **magyar nyelvű** előadásokra is járok.*
- *A **magyar nyelv** szép, de nem olyan könnyű.*
- ***Magyar nyelvű** szemináriumokra és előadásokra is járok.*

– A korpuszban a *magyar* és az *ember* kollokánsok közé nem ékelődik be másik lexikai elem;
– jellemzően határozott névelő áll a kollokációk előtt (*a magyar emberek, a magyar embereket, a magyar emberekkel*);
– az *ember* kollokáns leggyakoribb ragozott alakjai a többes számú *emberek*, illetve ennek tárgyragos alakja: az *embereket*, valamint a többes számú instrumentalisi *emberekkel*;
– a kollokációk mondatkezdő (*A magyar emberek nyelvében egy ábécé van.*) és mondatvégi pozícióban is előfordulnak (*Szeretnék találni és kommunikálni a magyar emberek is. Nagyon szeretem Magyarországot, és szeretem a magyar embereket. Én kell gyakorlat magyarul nyelvet a magyar emberekkel*).

A listázott mondatokat elolvasva az általam megkérdezett anyanyelvi beszélők nyelvi intuíciója szerint az alábbi példák több mondata furcsának, helytelennek tűnik. Az „ezt így biztosan nem használjuk" kategóriába sorolták például a *magyar emberek nyelvében* kollokációt. Ha az anyanyelvi óriáskorpuszban és a pedagógiai korpuszban rákeresünk a *magyar ember* kollokációra, akkor azt láthatjuk, hogy a *huTenTen12*-ben 42 574 alkalommal, míg a *MagyarOK*-ban egyetlen alkalommal sem szerepel. Az óriáskorpusz konkordanciasorait megfigyelve azonban arra az eredményre jutunk, hogy a fenti példákban kiemelt kollokációk valóban nem szerepelnek az anyanyelvi óriáskorpuszban. Így például *a magyar emberek nyelvében* helyett a szűrt gyakorisági adatok szerint valószínűsíthetően *a magyar nyelvben* vagy *a magyarban* kollokációk szerepelnének, hiszen *a magyar emberek nyelvében* egyetlen egyszer sem szerepelt a gyűjteményben. További vizsgálat tárgyát képezheti az is, hogy annak ellenére, hogy a pedagógiai korpuszban, mely tartalmazza a *MagyarOK* tankönyvek teljes szövegét, a *magyar emberek* kollokáció nem szerepel, a diákok vajon miért használják ilyen gyakorisággal ezt a szerkezetet, mennyire befolyásolja őket például az első nyelvük grammatikai struktúrája.

- *Sajnos nem ismerek sok* ***magyar embert****, mert nem sok időm.*
- *Szeretem a* ***magyar embereket****, az építészetet, a természetet.*
- *Szeretnék találni és kommunikálni a* ***magyar emberek*** *is.*
- *A* ***magyar emberek*** *nyelvében egy ábécé van.*
- *Nagyon szeretem Magyarországot, és szeretem a* ***magyar embereket****.*
- *Én kell gyakorlat magyarul nyelvet a* ***magyar emberekkel****.*
- *Szerintem* ***magyar emberek*** *nem szeretik idegen nyelveket tanulnak.*
- *Magyarul is tanulok, de szerintem írni és a nyelvtan elég nehéz, viszont már sok mindent értek a* ***magyar emberek*** *nyelvében.*
- *Jövő nyáron szeretnék Magyarországra menni, mert szeretnék tudni, hogy jó tudok* ***magyar emberekkel*** *beszélgetni.*
- *Jó lenne, ha* ***magyar emberekkel*** *beszélgetni tudnék, mert most sok magyart tanulok.*

Megfigyelhetjük tehát, hogy mindhárom korpuszban a *magyar* nemzetiséget jelölő melléknév leggyakoribb főnévi kollokánsa a *nyelv* szó volt. Míg azonban az anyanyelvi óriáskorpuszban a *magyar nyelv* komplex kontextusban jelenik meg, addig a pedagógiai és tanulói korpuszban mint a nyelvtanítás és a nyelvtanulás tárgya, a *szeret, tanít, tanul, jár* igék ragozott alakjaival (jellemzően E/1 és E/3) kollokálva (*Kijevben magyar nyelvet tanítok egy egyetemen. Szereti a magyar nyelvet? Magyar nyelvű szemináriumokra és előadásokra is járok. Most tanulom a magyar nyelvet*). A második leggyakoribb kollokánsnál mindhárom esetben eltérő eredményt kaptunk, de a pedagógiai és tanulói korpusz esetében itt is hasonlóság áll fent, hiszen az *étel* és *ember* kollokánsok és kontextusba helyezésük a magyar nyelvvel és kultúrával kapcsolatos ismeretszerzéssel függnek össze, míg az anyanyelvi korpusz *kormány* kollokánsa történelmi

és politikai szövegkörnyezetben fordul elő, melyek a sajtónyelv regiszterét képviselik. Nagyon szépen kirajzolódnak tehát az anyanyelvi óriáskorpusz, a pedagógiai korpusz és a tanulói korpusz közötti átfedések, de azt is láthatjuk, hogy az óriáskorpusz tartalmilag nem minden esetben egyezik a pedagógiai és tanulói korpusz gyűjteményével, hiszen míg az anyanyelvi korpuszok az anyanyelvi beszélők természetes nyelvhasználatát hivatottak reprezentálni, addig a pedagógiai korpuszok célja az, hogy a természetes nyelvhasználat egyes elemeit úgy szemléltessék, hogy azok érthetőek legyenek a nyelvtanulók számára a nyelvi szintjüknek megfelelően (mely feltétel nem zárja ki a korpusz reprezentativitását sem – l. a vizsgálatban a *magyar nyelv* leggyakoribb kollokációt mindhárom korpuszban). A tanulói korpuszokban megjelenő adatok rávilágítanak arra, hogy a nyelvtanulóknak mennyire sikerült adaptálniuk és transzformálniuk az adott tananyagot.

5. Zárszó

A folyamatosan bővülő *KorSzak* tanulói korpusz elemzésével elméleti, gyakorlati és pedagógiai kérdésekre kaphatunk választ, s ezáltal közelebb kerülhetünk a magyar mint idegen nyelv elsajátításának folyamataihoz, módjaihoz. A korpusz empirikus alapot biztosít azoknak a nyelvi sajátosságoknak a feltárására is, amelyek a magyart mint idegen nyelvet tanulók internyelvét jellemzik a nyelvtudás különböző szakaszaiban és nyelvi szituációkban. A korpusz vizsgálatainak eredményei rámutathatnak a nyelvtanulók azon hiányosságaira, melyek kiküszöbölésével még hatékonyabbak lehetünk a nyelvtanítás és a nyelvtanulás során. A kutatási eredményekből nemcsak az elméleti kutatók és gyakorló nyelvtanárok, hanem az önálló nyelvtanulók is profitálhatnak. A korpuszalapú szemlélet szélesebb körű elterjedése pedig minőségi változást eredményezhet a MID oktatásban.

Irodalom

Baumann Tímea – Majoros Judit – Pelcz Katalin – Schmidt Ildikó – Szita Szilva – Vermeki Boglárka 2020. Bemutatkozik a Korpusznyelvészeti és Szakmódszertani Munkacsoport. *Hungarológiai Évkönyv* 21/1–2: 23–31.

Dickinson, Markus – Ledbetter, Scott 2012. Annotating Errors in a Hungarian Learner Corpus. In: *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey. 1659–1664.

Durst Péter – Szabó Martina Katalin – Vincze Veronika – Zsibrita János 2013. A „HunLearner" magyar tanulói korpusz fejlesztése és várható hozadékai. *THL 2: A magyar nyelv és kultúra tanításának szakfolyóirata*. 28–41.

Durst, Péter – Szabó, Martina Katalin – Vincze, Veronika – Zsibrita, János 2014. Using Automatic Morphological Tools to Process Data from a Learner Corpus of Hungarian. *APPLES: Journal of Applied Language Studies* Vol. 8/3: 39–54.

Gablasova, Dana – Brezina, Vaclav – McEnery, Tony 2017. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning 67 (S1)*. 130–154.

Gablasova, Dana – Brezina, Vaclav – McEnery, Tony 2019. The Trinity Lancaster Corpus. Development, description and application. *International Journal of Learner Corpus Research*, Volume 5, Issue 2. 126–158.

Gardner, Robert C. – Lambert, Wallace E. 1959. Motivational variables in second language acquisition. *Canadian Journal of Psychology* 13. 191–200.

Granger, Sylviane 1993. International Corpus of Learner English. In: Jan Aarts – Pim de Haan – Nelleke Oostdijk (eds.) *English language corpora: design, analysis and exploitation*. Rodopi, Amsterdam. 57–69.

Granger, Sylviane 2004. Computer learner corpus research: current status and future prospects. In: Ulla Connor – Thomas A. Upton (eds.) *Applied Corpus Linguistics: A Multidimensional Perspective*. Rodopi, Amsterdam – Atlanta, GA. 123–145.

Granger, Sylviane – Dupont, Maïté – Meunier, Fanny – Naets, Hubert – Paquot, Magali 2020. *International Corpus of Learner English. Version 3*. Presses universitaires de Louvain.

Hana, Jirka – Rosen, Alexandr – Škodová, Svatava – Štindlová, Barbora 2010. Error-tagged learner corpus of Czech. In: *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Sweden. Association for Computational Linguistics. 11–19.

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

Jantunen, Jarmo Harri – Brunni, Sisko 2013. Morphology, lexical priming and second language acquisition: a corpus-study on learner Finnish. In: Sylviane Granger – Gaetanelle Gilquin – Fanny Meunier: *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead*. *Corpora and Language in Use* 1. 235–246.

Nesselhauf, Nadja 2005. *Collocations in a Learner Corpus*. John Benjamins, Amsterdam.

Reznicek, Marc – Lüdeling, Anke – Krummes, Cedric – Schwantuschke, Franziska – Walter, Maik – Schmidt, Karin – Hirschmann, Hagen – Andreas, Torsten 2012. *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01*. Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik – Korpuslinguistik, Berlin.

Rosen, Aleksandr 2016. Building and using corpora of non-native Czech. In Brejová, Brona: *ITAT. 2016. Information Technologies – Applications and Theory (Proceedings)*. Bratislava, CreateSpace Independent Publishing Platform – CEUR Workshop Proceedings. 80–87.

Selinker, Larry 1972. Interlanguage. *IRAL10*. 209–230.

Sinclair, John 1996. *EAGLES. Preliminary recommendations on Corpus Typology*. http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html (letöltve: 2021.08.15.)

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

Szita Szilvia 2020. Korpuszépítés és korpuszhasználat alacsonyabb nyelvtudási szinteken. *Hungarológiai Évkönyv* 21/1–2: 173–179.

Antal, Zsófia

The *KorSzak* learner corpus, and a corpus analysis of the adjective ‘magyar’ denoting Hungarian nationality

Learner corpora can play an important role in second-language learning research. The available corpora and text analysis software allow us to systematically analyze large language samples of students to see how they learn a new language. The article presents the dynamic learner corpus of the Work Group for Corpus Linguistics and Didactics (*KorSzak*) and draws attention to some potential research areas based on the corpus. It presents a possible study, namely the most frequent noun collocational partners related to the adjective ‘magyar’ (*Hungarian*) denoting Hungarian nationality.