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

MAGYAR MINT IDEGEN NYELV TANKÖNYVEK NYELVI ANYAGÁNAK SZÁMÍTÓGÉPES ELEMZÉSE[5]

Abstract

This paper presents the results of an analysis carried out on six coursebooks of Hungarian as a foreign language with the help of *magyarlanc*, a sentence splitter, morphological analyzer, POS-tagger and dependency parser. The same analysis was performed on data from two corpora (HunLearner – a learner corpus of Hungarian as a foreign language and a subcorpus of the Szeged Treebank – the largest fully manually annotated treebank of Hungarian), which was then compared to data from the coursebooks. Our results include the proportions of different conjugated verb forms according to personal endings, the different types of definite objects marked on verbs and frequency lists of nouns and verbs.

Keywords: *learner corpus, computational linguistics, coursebook of Hungarian as a foreign language*
Kulcsszavak: *tanulói korpusz, számítógépes nyelvészet, MID nyelvkönyv*

1. Bevezetés

A THL2 előző számában megjelent tanulmány (Durst–Szabó–Vincze–Zsibrita 2013) bemutatta a *HunLearner* tanulói korpuszt, és röviden összefoglalta a korpuszon végzett elemzések eredményeit, köztük a Károli Gáspár Református Egyetemen 2013. december 14-én *„A magyar mint idegen nyelv napja"* című rendezvényen tartott előadását is. Miután többen jelezték, hogy szeretnék felhasználni munkájukhoz az előadáson bemutatott adatokat, úgy döntöttünk, hogy azokat a fent említett tanulmányban található rövid összefoglalónál részletesebb formában is közreadjuk.

[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.

A jelen tanulmány tehát hat MID-tankönyv szövegét elemzi és veti össze a HunLearner tanulói korpusz anyagával valamint a Szeged Treebank egy alkorpuszával. Az elemzésben szereplő tankönyvek (megjelenésük sorrendje szerint és a jelen elemzésben használt rövidítésekkel): *Halló, itt Magyarország! I (HALLÓ); Hungarolingua 1 (HL1).; Lépésenként magyarul 1 (L1); Új színes magyar nyelvkönyv 1. (SZÍNES); Hungarian the Easy Way 1-2. (HEW), MagyarOK 1. (MOK).* A *Hungarian the Easy Way* a többi tankönyvtől eltérő módon kettő helyett 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 elemzésben csak a tankönyvek olvasmányai szerepelnek, a feladatok anyaga nem. A hallott és az olvasott szöveg értését gyakran fejlesztik kiegészítendő feladatokkal, de az ilyen „hiányos" szövegek nem szerepelnek az elemzésben, még abban az esetben sem, ha a teljes szöveg a könyv függelékében megtalálható. Nagy előfordulási arányuk miatt több ilyen jellegű szöveggel kivételt tettünk a MagyarOK esetében.

A jelenlegi elemzésben a teljes tanulói korpusznak csak egy része szerepel, amelyben többféle szöveg található: horvát anyanyelvű diákok által írt esszék *(Egy szimpatikus ember, Nehézségek a magyar nyelv tanulásában,* illetve *Magyar bevándorlók Angliában* címmel), valamint különböző anyanyelvű nyelvtanulók fogalmazása (*Egy szimpatikus ember* címmel). Az elemzésben szereplő korpusz 1427 mondatból és 22 000 tokenből áll.

2. Az elemzés módszere és eszközei

A *magyarlanc* nevű programcsomag (Zsibrita–Vincze–Farkas 2013) magyar nyelvű szövegek automatikus nyelvi elemzését hajtja végre 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, mint a függőségi elemzést tekintve, így vizsgálatainkban is ezt az eszközt alkalmaztuk.

Az elemző első lépésben a nyers szövegeket mondatokra, majd szavakra (tokenekre) bontja. A következőkben a szófaji egyértelműsítés során minden egyes szóhoz hozzárendeli annak az adott környezetben érvényes morfológiai elemzését, illetve a hozzá tartozó szótövet. Ezáltal rendelkezésünkre áll a szövegnek egy morfológiailag elemzett és lemmatizált verziója, mely lehetővé teszi, hogy elemzésünkben egységesen tudjuk kezelni egy adott szó (szótő) előfordulásait toldalékolástól függetlenül, továbbá a homonim szavakat is képesek vagyunk szófaj szerint elkülöníteni (tehát például a *nő* igei és főnévi előfordulásait külön tudjuk figyelembe venni).

Az alábbiakban bemutatunk egy példát a morfológiailag egyértelműsített és lemmatizált szövegre. A felső sorban látható az eredeti mondat, alatta az egyes tokenekhez

tartozó szótövek, majd a morfológiai elemzések láthatók. A morfológiai elemzések első karaktere határozza meg a szófajt (pl. V – ige, N – főnév), a további pozíciók a részletes morfológiai jellemzést adják (eset, szám, személy stb.).

Tudod	,	hogy	feleségül	akartalak	venni	?
tud	,	hogy	feleség	akar	vesz	?
Vmip2s---y	,	Cssp	Nn-sw	Vmis1s---2	Vmn	?

A magyarlanccal elemzett tankönyvi szövegekben automatikusan megszámoltuk az igei, illetve főnévi elemzéssel rendelkező szótöveket, a leggyakrabban előforduló szótövek részletesen a 3-6. táblázatokban láthatók. A részletes morfológiai elemzés segítségével pedig az igealakokat szám-személy szerint is csoportosítani tudtuk, a számszerű adatokat az 1. táblázat mutatja.

A magyarlanc egy további modulja segítségével a szövegeket szintaktikai elemzésnek is alávetettük, így minden egyes mondathoz hozzátársítottuk annak függőségi elemzését. Jelenleg részletesebben a határozott és határozatlan ragozás kérdéseit kutatjuk (lásd még pl. Vincze–Zsibrita–Durst–Szabó 2014), így elsődlegesen az ige-határozott tárgy kapcsolatokra fókuszáltunk vizsgálataink során. A szavak közti szintaktikai kapcsolatok felhasználásával automatikusan összegyűjtöttük az ige-tárgy párokat a szövegekből, majd a morfológiai és szintaktikai elemzés segítségével megállapítottuk a tárgy típusát is. A részletes eredmények a 2. táblázatban láthatók. A táblázatokban látható adatokban a jobb áttekinthetőség érdekében mindenhol egy tizedesjegyig kerekített adatok szerepelnek.

3. Az adatok és rövid értelmezésük

3.1. Az igealakok gyakorisága az egyes tankönyvekben és a korpuszokban

Az igealakok gyakoriságának vizsgálatakor szembetűnik a tananyagok közti viszonylag nagy eltérés: míg a HEW összesen 500 igét tartalmaz, addig a SZÍNES 2518-at. Ez a tankönyvek közti alapvető különbséget is tükrözi, hiszen az utóbbi jóval hosszabb szövegekkel és nagyobb terjedelemmel dolgozik.

A ragozott igealakok eloszlása azonban – néhány kiugró adattól eltekintve – meglehetősen hasonló, és az arányok mögött valószínűleg könnyen azonosítható okok húzódnak meg, így például a kommunikációban betöltött szerepe és gyakorisága miatt érthető az E/1 igealakok nagy aránya. A MOK esetében ez még hangsúlyosabb, ami vélhetően a könyv tudatosan kialakított célrendszerének köszönhető. Az E/2 igealakok használatának súlyát a tipikus tankönyvi beszédhelyzeteken és szövegtípusokon túl a tegezés/magázás közti választás is befolyásolja, így jól érthető a HL1-ben a többi tan-

anyagnál érezhetően magasabb arányuk is, ugyanis a könyvben túlnyomórészt párbeszédeket folytatnak az egymás között tegeződő állandó szereplők. Szintén a tegeződés használatára utal a többi tananyagnál gyakoribb (ugyanakkor még így is igen ritkán előforduló) T/2 igealak is. Minden tananyagra igaz, hogy a narratív szövegeken túl a magázó formát használó párbeszédek is az E/3 igealakok arányát növelik. Meg kell jegyezni, hogy ugyan a magyarázatok között említettük egyes igealakok (pl. E/1) gyakoriságát a mindennapi kommunikációban, beszélt nyelvi korpuszban ezt megfelelő nyelvi adatok hiányában nem vizsgáltuk, tehát hivatkozható formában ezt nem lehet alátámasztani. A legnagyobb, kézzel annotált magyar nyelvű szintaktikai adatbázisban, a Szeged Dependencia Treebankben (Vincze et al. 2010) azonban megvizsgáltuk az iskolai fogalmazások alkorpuszban található igei eloszlásokat, hiszen ezeket a szövegeket magyar anyanyelvű diákok írták, így a fogalmazás mint műfaj sajátosságai megjelennek itt is és a HunLearnerben is. Az igealakok hasonló eloszlást mutatnak a magyar és nem magyar anyanyelvűek által írt fogalmazásokban, az E/3 és T/1 alakok kivételével. A T/1 alakok nagyarányú használatának az lehet a magyarázata, hogy a Szeged Treebankben a diákok egy érdekes napjukról írtak, ahol az események több, a mesélő csoportjába tartozó szereplőt is érinthetnek, míg a HunLearner esetében a fogalmazások témája inkább az egyénhez kapcsolódott, és kevesebb csoport szintű eseményt szerepeltettek a szövegben.

1. táblázat: Az igealakok megoszlása az egyes tananyagokban és a korpuszokban

Tankönyv neve (összes ige)	E/1	E/2	E/3	T/1	T/2	T/3
HALLÓ (695)	25,5%	4,2%	43,5%	10,9%	0,7%	15,3%
HEW (500)	16,8%	8,8%	57,2%	5,4%	0,8%	11%
HL1 (1667)	32,3%	12,5%	36,1%	9,7%	3,4%	6%
L1 (1114)	20,3%	6,1%	51,2%	7,8%	0,5%	14,2%
MOK (844)	44,8%	3,8%	34,1%	7,5%	0,1%	9,7%
SZÍNES (2518)	19,4%	3,9%	54,8%	6,1%	0,8%	15%
Összesen (7338)	25,8%	6,5%	46,7%	7,8%	1,2%	11,9%
Teljes tanulói korpusz (2423)	29,1%	1,2%	51%	7,7%	0,1%	10,9%
Szeged Dependencia Treebank (iskolai fogalmazások) (50218)	28,2%	0,5%	39,2%	21,8%	0,1%	9,7%

3. 2. A határozott tárgyak megoszlása az egyes tankönyvekben és a korpuszokban

Tudatos nyelvhasználók, nyelvtanárok, nyelvészek és tankönyvszerzők minden bizonnyal intuitív módon is a táblázat adataihoz hasonló megoszlást feltételeznének a

határozott tárgyak gyakorisága között. Igazi érdekességnek inkább a pontos arányok számíthatnak, illetve a tankönyvek szövegében és a tanulói korpuszban megfigyelhető arányok összevetése. Látható, hogy a határozott tárgyak közül mindkét vizsgált korpuszban a tulajdonnév, a határozott névelős köznév, a birtokos szerkezetek, valamint a mutató névmás fordul elő számottevő rendszerességgel. Ez a hasonlóság valószínűleg nem a tananyagoknak a későbbi nyelvhasználatban játszott közvetlen szerepét mutatja, sokkal inkább a tankönyvek szerzőinek a helyes választását. Ez a választás pedig minden bizonnyal a hétköznapi beszélt nyelvben megfigyelhető gyakoriságon (ami statisztikai adatokkal ismét nem támasztható alá), valamint az általános nyelvtanári tapasztalaton alapszik. Összehasonlításképpen ismét közöljük a Szeged Dependencia Treebank iskolai fogalmazások alkorpuszából származó adatokat, melyek megerősítik, hogy a HunLearner korpuszban és a Szeged Treebankben hasonló arányokat mutat a határozott tárgy különböző típusainak előfordulási aránya, azaz a nyelvtanulók a valós magyar nyelvhasználatnak megfelelően használják e tárgytípusokat.

2. táblázat: A határozott tárgyak egyes típusainak előfordulása a tananyagokban és a tanulói korpuszban

Elemzett anyag (elemzett igék száma) / Határozott tárgy típusa	Halló (94)	HEW (61)	HL1 (214)	L1 (174)	MOK1 (64)	SZÍN (511)	Össz (1118)	Tanulói korpusz	SZDT
Tulajdonnév	35,1%	42,6%	23,8%	44,8%	43,8%	42,7%	38,8%	63%	51,2%
Határozott névelős köznév	34%	36,1%	44,9%	31,6%	31,3%	27%	32,5%	16%	18,8%
Birtok	18,1%	8,2%	7,9%	8,6%	7,8%	17,2%	13,1%	7%	12,2%
3. személyű névmás	4,3%	0%	8,4%	2,9%	4,7%	1,8%	3,5%	2%	1,5%
Visszaható névmás	2,1%	1,6%	0%	0%	0%	2,7%	1,5%	2%	2,7%
Kölcsönös névmás	3,2%	1,6%	0%	0%	0%	0,6%	0,6%	0,3%	0,4%
-ik végű kérdő névmás	0%	0%	2,3%	0%	0%	0%	0,4%	0,4%	0%
Mutató névmás	3,2%	9,8%	12,6%	12,1%	12,5%	7,8%	9%	9,4%	8,5%

3.3. A tankönyvekben és a tanulói korpuszban gyakran előforduló igék

A MID tankönyvek szókincsét – legalábbis a kezdő tananyagokban – általában nagymértékben meghatározza a grammatikai alapon szerveződő tanmenet, valamint az alap szintű kommunikációban természetesen előforduló szituációk. Jelentésük és a

hozzájuk kapcsolható grammatikai ismeretek miatt is érthető a listákon szereplő igék nagy részének magas gyakorisága, valamint a tananyagok között megfigyelhető nagy egybeesés. Az előfordulásban nagyságrendi különbséget a SZÍNES és a többi könyv között figyelhetünk meg, de ennek oka ismét a tananyagok terjedelme közti eltérés.

A statisztika valószínűleg ismét nem kínál sok meglepetést a tankönyveket használó és ismerő tanároknak, de a számadatokat érdemes értelmezni és persze néhány kiemelkedőbb adaton is hasznos lehet elgondolkodni. Mindenképpen meg kell jegyezni, hogy általában még a gyakorinak számító igék sem fordulnak elő tíznél többször az olvasmányokban, így mondhatjuk, hogy az olvasmányok funkciója leginkább az új nyelvtan bemutatása, a gyakorláshoz, a tanultak megerősítéséhez azonban nem adnak elég teret. Ennek a vizsgálatához érdemes lenne szemügyre venni a feladatok szóanyagát is.

Egy-egy ige kiemelkedően sokszor szerepel egy adott tankönyvben, így például a HL1-ben és a Hallóban a *kíván* és a *parancsol* igék. Ennek igen kézenfekvő magyarázata a *Jó napot kívánok* köszönésforma, valamint az éttermi szituációkban a *Mit parancsol?* kifejezés használata, ami talán az adott tananyagok arculatára, stílusára, jellemző témaválasztásaira utal. Ugyanakkor azt is meg kell jegyezni, hogy például a HEW esetében már olyan ige is szerepel az első 40 igét tartalmazó gyakorisági listán, amelyik mindössze egy olvasmányban szerepel *(kertészkedik)*.

Ha a tanulói korpusszal hasonlítjuk össze a könyvek anyagát, akkor jónéhány olyan igét találunk, amely a könyvekben is gyakori (pl. *van, tud, tanul, szeret, beszél, megy*), ami a nyelvtanulók valós nyelvhasználata és az autentikus nyelvhasználatra alapozni szándékozó könyvek sikeres törekvése közti kapcsolatot is mutatja. Meg kell jegyezni azt is, hogy a fogalmazások témája valamennyire leszűkítette a felhasználható igék körét (pl. vásárlási szituációkban előforduló *parancsol* vagy *kíván* igék ezért sem szerepelnek itt).

3. táblázat: A leggyakrabban előforduló igék az egyes tananyagokban (az előfordulás számával)

HALLÓ				**HEW**			
1. van	136	21. ismer	7	1. van	84	21. mond	7
2. megy	34	22. táncol	7	2. tud	22	22. segít	7
3. kér	31	23. vár	7	3. szeret	21	23. áll	7
4. köszön	21	24. iszik	6	4. megy	16	24. kap	6
5. nincs	20	25. jön	6	5. dolgozik	15	25. lát	6
6. szeret	17	26. tölt	6	6. jön	15	26. megnéz	6
7. kíván	15	27. ad	5	7. beszél	13	27. ismer	5
8. lesz	14	28. akar	5	8. lesz	12	28. késik	5
9. tetszik	14	29. bemegy	5	9. tanul	11	29. utazik	5
10. lakik	13	30. elmegy	5	10. beszélget	10	30. figyel	4
11. tud	13	31. eszik	5	11. csinál	9	31. kertészkedik	4
12. dolgozik	11	32. játszik	5	12. vesz	9	32. kiabál	4
13. jár	11	33. lehet	5	13. ül	9	33. lakik	4
14. parancsol	11	34. megnéz	5	14. dolgoz	8	34. lehet	4
15. beszél	9	35. mond	5	15. eszik	8	35. olvas	4
16. találkozik	9	36. olvas	5	16. kell	8	36. ráér	4
17. keres	8	37. él	5	17. köszön	8	37. találkozik	4
18. lát	8	38. beszélget	4	18. nincs	8	38. él	4
19. örül	8	39. csókol	4	19. akar	7	39. örül	4
20. csinál	7	40. felmegy	4	20. kér	7	40. ebédel	3

HL

1. van	319	21. tanít	15
2. megy	91	22. érkezik	15
3. kér	77	23. akar	14
4. jön	76	24. dolgozik	13
5. tud	59	25. olvas	13
6. köszön	56	26. él	13
7. lesz	55	27. beszélget	12
8. szeret	50	28. ismer	12
9. tanul	32	29. mond	12
10. vesz	31	30. találkozik	12
11. kíván	29	31. ajánl	11
12. parancsol	29	32. ül	11
13. vár	23	33. iszik	10
14. bemegy	22	34. siet	10
15. nincs	21	35. tesz	10
16. csinál	19	36. fest	9
17. beszél	17	37. fáj	9
18. néz	17	38. lakik	9
19. hoz	15	39. ráér	9
20. lát	15	40. tetszik	9

Lépésenként

1. van	235	21. kell	12
2. megy	49	22. dolgozik	11
3. tud	44	23. ad	10
4. szeret	31	24. bemegy	10
5. jön	29	25. beszélget	10
6. köszön	25	26. elmegy	10
7. csinál	23	27. hisz	10
8. lesz	22	28. nincs	10
9. kér	18	29. utazik	10
10. lehet	18	30. fizet	9
11. találkozik	18	31. segít	9
12. tanul	18	32. hoz	8
13. lát	17	33. iszik	8
14. vesz	16	34. kimegy	8
15. akar	15	35. lakik	8
16. visz	15	36. indul	7
17. mond	14	37. jár	7
18. olvas	14	38. pihen	7
19. ül	13	39. tetszik	7
20. beszél	12	40. vár	7

MAGYAROK

1. van	207	21. játszik	10
2. szeret	39	22. kell	10
3. tud	38	23. főz	9
4. köszön	34	24. lehet	9
5. nincs	28	25. olvas	8
6. tanul	26	26. tölt	8
7. él	21	27. alszik	7
8. beszél	20	28. dolgozik	7
9. megy	20	29. eszik	7
10. kér	17	30. utazik	7
11. lesz	17	31. hallgat	6
12. kíván	15	32. hoz	6
13. tetszik	15	33. működik	6
14. vesz	14	34. néz	6
15. lakik	13	35. pihen	6
16. ír	13	36. sportol	6
17. csinál	12	37. beszélget	5
18. örül	12	38. gyakorol	5
19. jön	11	39. vár	5
20. segít	11	40. vásárol	5

SZÍNES

1. van	447	21. indul	21
2. tud	78	22. segít	21
3. megy	70	23. él	21
4. dolgozik	60	24. beszél	20
5. nincs	52	25. keres	20
6. lesz	50	26. elmegy	19
7. kér	44	27. érkezik	19
8. szeret	42	28. fog	18
9. jár	39	29. gondol	18
10. tanul	35	30. beszélget	17
11. akar	33	31. dolgozik	17
12. köszön	31	32. vesz	17
13. lát	31	33. eszik	16
14. jön	30	34. ismer	16
15. lakik	28	35. ül	16
16. kell	25	36. néz	15
17. mond	25	37. tölt	15
18. kíván	24	38. vár	15
19. áll	24	39. iszik	14
20. csinál	21	40. utazik	14

4. táblázat: A leggyakrabban előforduló igék a tanulói korpuszban (az előfordulás számával)

Tanulói korpusz			
1. van	491	21. csinál	27
2. tud	119	22. jön	27
3. kell	86	23. olvas	27
4. tanul	73	24. talál	25
5. lesz	68	25. dolgozik	24
6. mond	61	26. okoz	24
7. szeret	59	27. lát	22
8. beszél	48	28. létezik	21
9. lehet	48	29. tesz	20
10. megy	47	30. hisz	19
11. használ	45	31. ír	19
12. akar	38	32. néz	16
13. gondol	38	33. hall	15
14. nincs	38	34. keres	15
15. megtanul	36	35. marad	15
16. dolgozik	32	36. tűnik	15
17. kezd	32	37. válik	15
18. él	29	38. ért	15
19. fog	28	39. ismer	14
20. kap	28	40. találkozik	14

3.4. A tankönyvekben és a tanulói korpuszban gyakran előforduló főnevek

A MID területén jártas kollégákban intuitív módon megfogalmazódó gyakorisági listákra, a szavak grammatikai sajátosságainak jelentőségére és a tananyagok terjedelmének fontosságára vonatkozó eddigi megállapítások természetesen a főnevek statisztikájára is érvényesek. Az alábbi eredményeket megfigyelve persze ismét kiemelhetünk néhány olyan sajátosságot, amely kifejezetten a főnevek gyakoriságára jellemző. Ilyen például a tulajdonnevek magas előfordulása. Felmerült annak a lehetősége, hogy a tulajdonneveket kizárjuk a statisztikából, hogy nagyobb rálátásunk legyen a köznevekre, azonban végül úgy döntöttünk, hogy az adatok a tulajdonnevekkel együtt mutatnak igazán teljes képet, és ezeknek a számoknak is van információértékük. A tulajdonnevek gyakorisága például rámutat arra, hogy egy állandó szereplőkkel dolgozó tankönyvben elkerülhetetlen ugyan a nevek gyakori említése, más könyveknél azonban a váltogatott, csupán egy-egy szituációban használt tulajdonnevek helyett hasznos lehet hivatalos megszólítások, foglalkozásnevek alkalmazása, mert ezek tizenöt-húsz körüli előfordulása már segíthet a szó rögzülésében, továbbá a pragmatikai kompetenciát is fejlesztheti.

Természetesen a gyakori főnevek is «árulkodnak» az adott tananyagban jellemző helyekről és beszédhelyzetekről. Az ilyen jellegű következtetésekkel kapcsolatban ugyanakkor óvatosságra intenek olyan példák, mint a L1 egyik leggyakoribb tulajdonneve *(Hófehérke)*, amelyik a könyv 55 leckéjéből mindössze kettőben szerepel.

Összességében a számokat megfigyelve azt látjuk, hogy a főnevek esetében még a leggyakoribb szavak is kevesebb alkalommal fordulnak elő, mint a gyakoribb igék, és akár a lista első felében is találunk olyan szavakat, amelyek csupán egy vagy két leckében szerepelnek. Az egyes tankönyvek között jelentősebb különbségeket fedezhetünk fel a főnevek tekintetében, és lényegesen kevesebb a hasonlóság, mint az igék gyakoriságában. Alapvetően jóval több főnév jelenik meg a szövegekben, de kisebb az ismétlődés, ami nyelvpedagógia szempontból azt is jelenti, hogy az olvasmányok az egyes főnevek gyakorlására, rögzítésére még annyira sem lesznek alkalmasak, mint az igék esetében.

A tanulói korpusz adatai szintén meglehetősen nagy eltérést mutatnak a tananyagoktól, amit a korpusz gyűjtésekor leginkább a témaválasztás által beszűkített lehetőségek magyarázhatnak.

5. táblázat: A leggyakrabban előforduló főnevek az egyes tananyagokban (az előfordulás számával)

HALLÓ				**HEW**			
1. lecke	20	21. egyetem	7	1. Lóri	49	21. úr	7
2. szálloda	16	22. István	7	2. igazgató	31	22. asztal	6
3. úr	16	23. Marika	7	3. Erzsike	19	23. Einstein	6
4. étterem	14	24. újságíró	7	4. Zoli	19	24. irodalomóra	6
5. jegy	13	25. autó	6	5. polgármester	17	25. könyv	6
6. óra	13	26. Barta	6	6. gimnázium	16	26. nagy	6
7. Laci	12	27. barát	6	7. óra	16	27. szék	6
8. Miklós	12	28. bérlet	6	8. diák	15	28. év	6
9. lakás	11	29. ház	6	9. iskola	15	29. Éva	6
10. asztal	10	30. Kati	6	10. Anikó	12	30. ajtó	5
11. feleség	10	31. kocsi	6	11. iroda	11	31. gyerek	5
12. Géza	10	32. Péter	6	12. tanár	10	32. ló	5
13. nap	10	33. szoba	6	13. Csaba	9	33. perc	5
14. Paul	10	34. utca	6	14. Jenny	9	34. tanya	5
15. Braun	9	35. város	6	15. ember	8	35. város	5
16. forint	9	36. bor	5	16. idő	7	36. épület	5
17. gyerek	9	37. család	5	17. kert	7	37. Albert	4
18. hely	9	38. előadás	5	18. köpeny	7	38. Anna	4
19. villamos	8	39. emelet	5	19. lány	7	39. anyuka	4
20. Budapest	7	40. este	5	20. pénz	7	40. baj	4

HL1

1. nap	46	21. diák	15
2. Márta	44	22. fiú	15
3. Gábor	43	23. Mustafa	15
4. Jean	41	24. baj	14
5. egyetem	37	25. Carla	14
6. apa	32	26. lakás	14
7. John	30	27. szoba	14
8. úr	30	28. bank	13
9. lány	29	29. sör	13
10. Debrecen	27	30. feleség	12
11. család	24	31. anya	11
12. Eszter	22	32. lecke	11
13. tévé	22	33. férfi	10
14. Mike	21	34. gyerek	10
15. Kurt	20	35. magyar	10
16. óra	20	36. mozi	10
17. forint	18	37. állatkert	10
18. Mary	17	38. állomás	10
19. Baker	16	39. újság	10
20. bocsánat	15	40. buli	9

Lépésenként

1. ház	25	21. János	10
2. Hófehérke	19	22. könyv	10
3. lány	19	23. pulóver	10
4. szekrény	15	24. törpe	10
5. ember	14	25. vonat	10
6. óra	14	26. étterem	10
7. asztal	13	27. buli	9
8. Szeged	13	28. busz	9
9. gyerek	12	29. bácsi	9
10. baj	11	30. egyetem	9
11. hétvége	11	31. Gyula	9
12. kert	11	32. gyémánt	9
13. Kovács	11	33. Gábor	9
14. Móni	11	34. mozi	9
15. nap	11	35. Silvia	9
16. Péter	11	36. Tamás	9
17. Sándor	11	37. utca	9
18. épület	11	38. ágy	9
19. újság	11	39. állat	9
20. diák	10	40. ajtó	8

MOK

1. Magyarország	23	21. rend	8
2. nap	22	22. utca	8
3. nyelv	22	23. édesapa	8
4. ember	20	24. év	8
5. barát	14	25. forint	7
6. gyerek	13	26. Gábor	7
7. óra	13	27. magyar	7
8. elnézés	11	28. Nóra	7
9. hét	11	29. nő	7
10. kiló	11	30. ország	7
11. asztal	9	31. autó	6
12. család	9	32. baj	6
13. idő	9	33. főnök	6
14. iroda	9	34. ház	6
15. munka	9	35. japán	6
16. édesanya	9	36. kenyér	6
17. barátnő	8	37. mozi	6
18. egyetem	8	38. szerda	6
19. konyha	8	39. Zalán	6
20. Kovács	8	40. étterem	6

SZÍNES

1. Budapest	48	21. ország	26
2. lakás	47	22. ház	24
3. óra	44	23. nő	24
4. autó	40	24. hely	23
5. utca	40	25. kocsi	23
6. nap	39	26. János	22
7. munka	38	27. szálloda	21
8. év	38	28. barát	20
9. egyetem	37	29. férfi	20
10. ember	36	30. diák	19
11. Kati	36	31. szoba	19
12. Magyarország	34	32. vendég	19
13. autóbusz	32	33. pincér	18
14. nyelv	32	34. család	17
15. város	32	35. perc	17
16. Béla	30	36. busz	16
17. idő	29	37. feleség	16
18. úr	29	38. híd	16
19. asztal	28	39. konyha	16
20. gyerek	26	40. kulcs	16

6. táblázat: A leggyakrabban előforduló főnevek a tanulói korpuszban (az előfordulás számával)

Tanulói korpusz			
1. nyelv	268	21. jelentés	28
2. ember	91	22. London	25
3. szó	88	23. baj	23
4. magyar	77	24. egyetem	23
5. Magyarország	77	25. barát	22
6. igekötő	69	26. helyzet	22
7. ige	62	27. nap	22
8. nehézség	59	28. Anglia	19
9. idő	54	29. gyerek	19
10. munka	51	30. film	18
11. év	48	31. pénz	18
12. élet	44	32. eleje	17
13. probléma	43	33. eset	17
14. tanulás	42	34. hely	16
15. mondat	39	35. kiejtés	16
16. ragozás	37	36. magánhangzó	16
17. szórend	35	37. ország	16
18. horvát	32	38. Szeged	15
19. dolog	31	39. világ	15
20. család	28	40. hang	14

4. Összefoglalás

A fentiekben olyan adatokat mutattunk be, amelyeket MID tankönyvek és két nyelvi korpusz számítógépes feldolgozásával nyertünk. Az igealakok és az egyes tárgytípusok megoszlását, továbbá az igék és a főnevek gyakorisági listáit szemügyre véve egyrészt igazolhatjuk a gyakorlati nyelvtanári munkánk során kialakult feltételezéseinket, másrészt pontos adatokra támaszkodva kapunk képet a tananyagok néhány jellemzőjéről. Mindenképpen megállapíthatjuk, hogy a tananyagok elemzéséhez is érdemes segítségül hívni a számítógépet, és így megfogalmazhatunk további célokat is (például az olvasmányokon kívül a gyakorlatok anyagának elemzését vagy az újabb tananyagok szókincsének az itt látható adatok figyelembe vételével történő kiválasztását).

Irodalom

Durst Péter – Szabó Martina Katalin – Vincze Veronika – Zsibrita János 2013. A „HunLearner" magyar tanulói korpusz fejlesztése és várható hozadékai. *THL2 1–2*: 28–41.

Vincze Veronika – Szauter Dóra – Almási Attila – Móra György – Alexin Zoltán – Csirik János 2010. Hungarian Dependency Treebank. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta.

Vincze Veronika – Zsibrita János – Durst Péter – Szabó Martina Katalin 2014. Automatic Error Detection concerning the Definite and Indefinite Conjugation in the HunLearner Corpus. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, Reykjavik, Izland. 3958–3962.

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.

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