Magyar nyelvű történeti korpuszok

Simon Eszter
tudományos főmunkatárs
MTA Nyelvtudományi Intézet, Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály
simon.eszter@nytud.mta.hu

A számítógépes nyelvészeten belül kutatási területeim közé tartozik a tulajdonnév-felismerés, a morfológiai elemzés, a korpuszépítés és -annotáció, a történeti korpuszok fejlesztése, valamint az uráli nyelvek számítógépes nyelvészeti támogatása. Számos hazai és nemzetközi projektben vettem részt, amelyek egy részében a számítógépes nyelvészeti munkálatok koordinátora is én voltam. Több egyetemen tartottam nyelvészeti tematikájú kurzusokat, emellett rendszeresen bírálok szakdolgozatokat, disszertációkat, cikkeket, absztraktokat, pályázatokat hazai és nemzetközi színtéren egyaránt.

1. Bevezetés

A nyelvi kulturális örökség elérhetővé tételében kulcsfontosságú szerep jut a nyelvtechnológiának, melynek módszereivel a kutatók egységes, következetes, nyelvi információval ellátott adatbázisokhoz juthatnak. A nyelvtörténészek és nyelvtechnológusok egyik legfontosabb együttműködési terepe a történeti korpuszok építése, melyek kiváló alapanyagot szolgáltatnak az elméleti és történeti nyelvészeti kutatásoknak. Az elmúlt évtizedekben számos történeti korpuszt fejlesztettek – elsősorban indoeurópai nyelvekre, de a magyarra is készült néhány. Időrendi sorrendben haladva ezek a következők. Az Ómagyar Korpusz (Simon 2014) tartalmazza az összes ómagyar korból fennmaradt szövegemléket és néhány középmagyar kori bibliafordítást is. A Történeti Magánéleti Korpusz (Novák et al. 2018) az ó- és középmagyar kor magánéleti nyelvi regiszteréhez közelebb álló műfajokat tartalmazza: 1772 előtti magánlevelekből és peres eljárások jegyzőkönyveiből épül fel. A Magyar Történeti Szövegtár (Csengery 2006) pedig 1772-től, vagyis az újmagyar kor kezdetétől egészen a 20. század végéig tartalmaz szövegeket.

Jelen tanulmány a Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért – HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban címmel, 2018. október 19-én Szegeden tartott workshopon elhangzott előadásom írott változata. Előadásomban a fent említett korpuszokat és a hozzájuk tartozó lekérdezőfelületeket ismertettem, és néhány példán keresztül azt is illusztráltam,

hogy milyen kutatási kérdésekre hogyan tudunk választ kapni ezeknek az adatbázisoknak a segítségével. A tanulmány felépítése követi az előadás menetét: először a történeti korpuszok jellemzőit ismertetem a 2. fejezetben, majd a 3. fejezetben röviden áttekintem a történeti szövegek feldolgozásának kihívásait, végül a 4. fejezetben ismertetem a magyar nyelvű történeti korpuszokat, és egy vizsgálattal illusztrálom, hogy milyen kutatási kérdések megválaszolásához lehet használni ezeket az adatbázisokat.

2. A történeti korpuszok jellemzői

Egy történeti korpusz elsősorban korpusz, és mint ilyenre vonatkozik rá minden, ami általában egy korpuszról elmondható, vagyis: szövegek vagy szövegrészletek véges elektronikus gyűjteménye, amely jól körülhatárolt és nyelvészetileg releváns kritériumok alapján lett válogatva, valamint legalábbis törekszik a reprezentativitásra (Claridge 2008: 242). A hangsúly a törekvésen van, de a gyakorlatban a reprezentativitás egy mozgó célpont, ami a történeti korpuszok esetében még nehezebben érhető el, mint az általános célú modern korpuszok esetében.

Ami a történeti korpuszokat történetivé teszi, az az, hogy azzal a céllal készülnek, hogy reprezentálják egy nyelv régi állapotait, valamint hogy tanulmányozni lehessen rajtuk a nyelv változásait. Felmerül a kérdés, hogy mit értünk régi állapot alatt? Jellemzően azokat a szövegeket szokták réginek nevezni, amelyek legalább egy generációnyival visszanyúlnak a mai nyelvállapot előttre (Claridge 2008: 242).

A történeti korpuszoknak több típusát lehet elkülöníteni. A tipizálás egyik dimenziója mentén szinkrón és diakrón korpuszokat különböztetünk meg. A szinkrón történeti korpusz a nyelvnek egy múltbeli szeletét mutatja be, arról készít pillanatfelvételt. Ez a „pillanat” a történeti szövegek esetében akár egy évszázados is lehet, mint például a Century of Prose Corpus (Milíc 1990) esetében, amely az angol próza 1680 és 1780 közötti időszakát mutatja be. A diakrón korpusz ezzel szemben egy nagyobb időintervallumot ölel fel, ami a nyelv longitudinális vizsgálatára ad lehetőséget. Erre példa a Helsinki Corpus of English Texts (Rissanen et al. 1991), amely majd egy évezredet fog át (ca. 750–1710).

A korpusztipizálás egy másik dimenziója a korpusz felhasználása lehet, vagyis hogy a korpusz általános célú vagy specifikus. Az általános célú korpusz, mint amilyen a fent említett Helsinki Corpus of English Texts, a nyelvi vizsgáló-

dások széles skáláját teszi lehetővé, míg a specifikus korpuszok egy műfajra, egy szerzőre vagy – szélsőséges esetben — akár csak egy műre koncentrálnak, mint például az Electronic Beowulf[2] esetében.

Az időkeret a történeti korpuszok esetében kiemelt fontosságú. A régebbi korokból jellemzően kevesebb szöveges anyag áll a rendelkezésünkre, ezért fordulhat elő az, hogy még a szinkrón történeti korpuszok is egy évszázadot fognak át, ahogy azt láttuk fentebb. A méret pedig szorosan összefügg az időkerettel: minél nagyobb az időkeret, valószínűleg annál nagyobb lesz a korpusz.

Nem mindegy viszont, hogy melyik korban van a vizsgált időkeret. Ha egy 21. századi évet szemelünk ki időkeretnek, és az abban az évben keletkezett magyar nyelvű szövegekből akarunk korpuszt építeni, nagy valószínűséggel sokkal nagyobb korpuszt kapunk, mint ha ugyanezt egy 16. századi évvel tennénk. Általános szabályként kijelenthető, hogy a történeti korpuszok jellemzően kisebbek, mint a modernek. Ennek egyik oka, hogy a történeti szövegek feldolgozása jellemzően nagyobb kihívást jelent, mint az eleve elektronikusan keletkezett modern szövegek egyszerű letöltése. További ok, hogy a régebbi korokban egyrészt sokkal kevesebb nyelvi anyag keletkezett, másrészt pedig ezek egy része egyszerűen elveszett az idők során. Amik fennmaradtak, azok pedig sokszor még mindig csak kéziratban, papíron érhetők el, vagy ha digitalizálták is őket, akkor is általában csak képként, vagyis a bennük levő szöveges tartalom még mindig nem közvetlenül hozzáférhető.

Itt kanyarodunk vissza a reprezentativitáshoz, amely a korpuszépítés egyik legtöbbet tárgyalt kérdése, lásd például Biber (1993). Hunston (2008) definíciója szerint a reprezentativitás a korpusz és az általa reprezentált nyelv közötti viszonyt jelenti. Ez a definíció problémás, mivel körbenforgó: azt feltételezi, hogy a nyelvről tudunk dolgokat, miközben a korpuszt azért építjük, hogy megtudhassunk dolgokat a nyelvről. Talán érdemesebb úgy megközelíteni a kérdést, hogy mi az, ami biztosan nem reprezentatív. Ha valaki a mai magyar nyelvhasználat általános célú korpuszát óhajtja megépíteni, akkor csak és kizárólag sport tematikájú blogbejegyzéseket gyűjtve ezt nem fogja tudni teljesíteni. Vagy McEnery (2004) példájával élve: képzeljük el, hogy egy kutató egy telefonos dialógusrendszer fejlesztéséhez épít korpuszt. Ha ennek eléréséhez Jane Austen regényeiből szemezget, akkor biztosak lehetünk benne, hogy nem jár jó úton. A reprezentativitás kérdésével tehát érdemes óvatosan bánni, és úgy tekinteni, mint amire törekszünk, de a

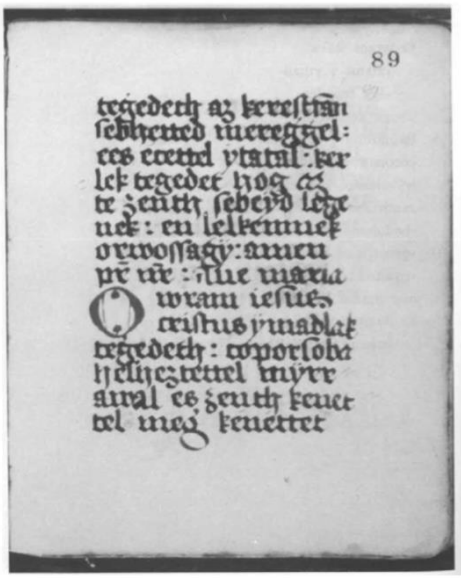
[2] http://ebeowulf.uky.edu

gyakorlatban elképzelhető, hogy sose fogjuk elérni. A történeti korpuszok esetében a reprezentativitást számos nyelven kívüli tényező is befolyásolja. Az egyik ilyen tényezőt már említettük: sok szöveg elveszett, és nem tudjuk, hogy mennyi és mi, vagyis a fennmaradt szövegek köre csak egy véletlen mintát szolgáltat a múltbeli nyelvről, nem reprezentálja a nyelv teljes változatosságát. Egy másik tényező a beszélt nyelvi adatok hiánya. A hangrögzítés technológiájának feltalálása előttről nyilvánvalóan nem lehetnek anyagaink, sőt még a kezdetleges hangrögzítő eszközök korából is csak nagyon kevés maradt fenn, azok is igen rossz minőségűek. Arra is érdemes odafigyelni, hogy a különféle műfajok, regiszterek aránya az írott nyelvváltozaton belül is kiegyensúlyozatlan – sok a vallási irodalom, kevés az informális, a sajtó, a tudományos stílus, aminek hátterében különféle szociopolitikai okok húzódnak. Fontos megemlíteni még a szociolingvisztikai kiegyensúlyozatlanság kérdését is, ugyanis a fennmaradt szövegek a társadalmi elit és az értelmiség nyelvét reprezentálják, mivel régebbi korokban az írástudás csak egy szűk réteg kiváltsága volt.

3. A történeti szövegek feldolgozása

A történeti szövegek különböző forrásokból származhatnak, melyek különböző feldolgozást igényelnek: vannak kézzel írott szövegek a nyomtatás feltalálása előttről és utánról, és vannak korai nyomtatványok. A nyomtatás feltalálása előtti korból származó kézzel írott szövegek egy részének már készült nyomtatott átirata. Erre láthatunk egy példát az 1. ábrán, amelyen egymás mellett szerepel a Margit-legenda 89r oldalának eredeti, kézzel írott változata és az abból nyelvtörténészek által készített nyomtatott átirat (P. Balázs et al. 1990). Itt azt lehet látni, hogy az átirat készítői törekedtek az ortográfiai hűségre, ami egy régi szöveg nyelvészeti szempontú vizsgálatához, illetve további felhasználásához feltétlenül szükséges.

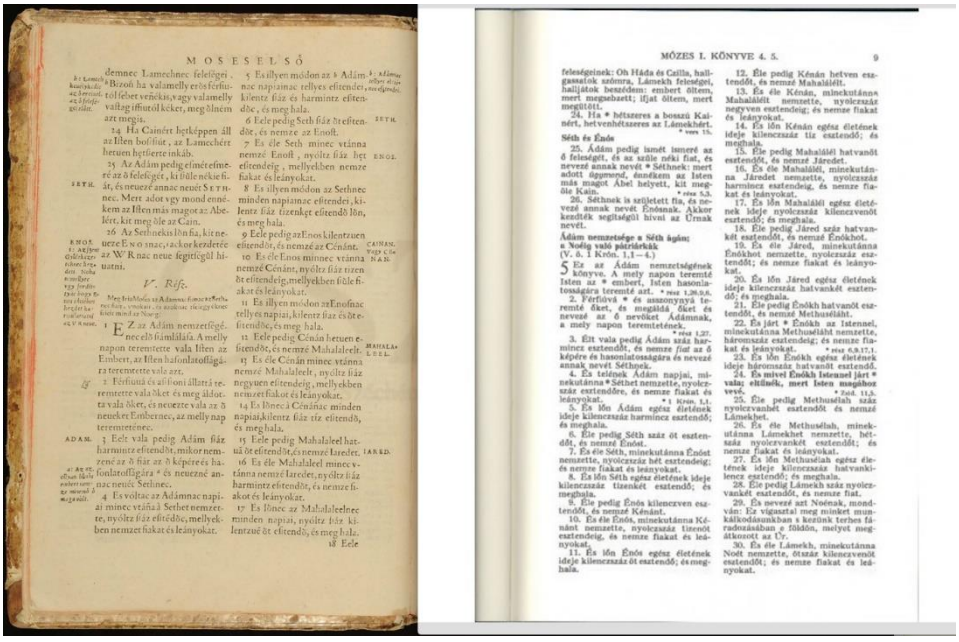
Ehhez hasonlóan léteznek a korai nyomtatványoknak későbbi, átdolgozott kiadásai is. A 2. ábrán a Károli Gáspár-féle bibliafordítás két kiadásának ugyanazon része szerepel. Bal oldalon az eredeti, 1590-es Vizsolyi Biblia (Károli 1590), jobb oldalon pedig az 1908-as revideált kiadás (Károli 1908) egy-egy oldala látható. Itt már jól tetten érhető, hogy az ún. revideált változat készítői nem fektettek nagy hangsúlyt a nyelvi jelenségek megőrzésére, hanem inkább a kor nyelvállapotához próbálták igazítani a régies nyelvezetet.



177
89r

tegedeth az kerestfan
sebhetted mereggel :
ees ecettel ytatal : ker-
-lek tegedet hog az
te zenth sebeÿd legé-
-nek : en lelkemnek
orwossagÿ : amen
pr nr : Aue maria
O wram iesus
cristus ÿmadlak
tegedeth : coporsoba
helheztettel mÿrr-
-awal es zenth kenet-
-tel meg kenéttet

1. ábra: A Margit-legenda 89r oldalának eredeti, kézzel írott változata és az abból készített nyomtatott átirat



MOSES EL SŐ

MÓZES I. KÖNYVE 4. 5. 9

2. ábra: A Károli-féle bibliafordítás egy-egy oldala két kiadásból.
Bal oldalon az 1590-es Vizsolyi Biblia, jobb oldalon az 1908-as revideált változat.

Amikor történeti korpuszt építünk, döntenünk kell, hogy milyen forrásokat használunk. A szerkesztett kiadások használata mellett szól egyrészt az, hogy általában könnyebben elérhetőek, mint az eredetiek. A Károli-féle bibliafordítás esetében, ha az interneten rákeresünk, akkor szinte az összes találat az 1908-as revideált változatra vagy annak valamelyik későbbi kiadására fog mutatni, míg az eredeti Vizsolyi Biblia szöveges változata csak az Ómagyar Korpusz weboldalán[3] érhető el. Szintén a szerkesztett kiadások használata mellett szól az, hogy a nyomtatott könyvek beszkennelése és azon optikai karakterfelismerő szoftver alkalmazása vagy a szöveg begépelése lényegesen könnyebb, mint ha az eredeti kézzel írott verzióval próbálná meg ugyanezt az ember. További fontos tényező, hogy a szerkesztői döntéseket már meghozták, így a történeti szövegek feldolgozásánál elég „csak" a feldolgozás lépéseire koncentrálni.

A szerkesztett kiadások használata ellen is szólnak azonban érvek. Amint már említettem, bizonyos kiadásokban az átírást és a szerkesztést nem nyelvészek végezték, hanem történészek vagy irodalmárok, akik egészen más szempontokat tartottak szem előtt, ám így is számos olyan döntést hoztak meg, amelyek hatással vannak a szövegre, de nincsenek kellően dokumentálva. Ezért ezek a kiadások nem alkalmasak további nyelvészeti vizsgálódásokra. Továbbá szerzői jogi problémákat is felvethetnek, hiszen az átirat készítői vagy élnek, vagy még nem telt el 70 év a haláluk óta, így az általuk szerkesztett kiadvány nem szabadon felhasználható, ellentétben az ómagyar kori kódexekkel, amelyek közkincsnek számítanak, nem köti őket szerzői jog. Kiutat jelenthet a dilemmából, ha szerkesztett kiadásokból indulunk ki, de mindent ellenőrzünk az eredeti verzióban.

A történeti szövegek feldolgozása és a belőlük való korpuszépítés számos kihívást rejt, amelyekkel a mai sztenderd nyelvállapotok feldolgozásakor nem feltétlenül szembesülünk. Az egyik ilyen a kézirat rossz fizikai állapotából következő nehézségek köre. A 3. ábrán a Königsbergi Töredék és Szalagjai elnevezésű korai szövegemlékünkből a Szalagok láthatóak. A nyelvemlék úgy maradt fenn, hogy egy másik, nem magyar nyelvű kódex kötéséhez használták fel: a számukra értelmetlen magyar szöveget tartalmazó lapokat kitépték, majd a Töredék lapját a kötet szennylapjául használták, a Szalagok lapját pedig a kötés megerősítésére csíkokra vágták, és beragasztották. Csak a 19. században fedezték fel a magyar nyelvű részeket, amikor kibontották, hazahozták, lefényképezték, de sajnos nem sikerültek túl jól a fotók. Viszont még a mai napig is csak ezekre a nem túl jól

[3] http://omagyarkorpusz.nytud.hu

sikerült fotókra kénytelen támaszkodni mindenki, aki ezt a nyelvemléket szeretné kutatni, ugyanis az eredeti Szalagok elvesztek (Madas 2009: 230–233).

3. ábra: A Königsbergi Töredék és Szalagjaiból a Szalagok

Az ómagyar kori szövegemlékeket és kódexeket a latin nyelvű és vallásos tárgyú irodalom fordításának igénye hívta életre, de a latin ábécé magyarra alkalmazása számos problémát vetett fel. A legfőbb gond abból fakadt, hogy nyelvünk hangrendszerének több eleme a latinban ismeretlen, így ezek jelölésére új jeleket kellett bevezetni. Kniezsa (1952) az ómagyar kori kódexek kezeinek helyesírását három nagy típusba sorolja. A mellékjel nélküli helyesírás a latinban nem szereplő magyar hangokat több betű kombinációjával írja le, például: *cs* → *ch* ~ *cz* ~ *chy* ~ *chi* ~ *cy*. A mellékjeles helyesírás egy rokonhang betűjének mellékjeles változatával jelöli ezeket, például: *cs* → *č* ~ *ć*. A harmadik típus pedig ezek keveréke, amely egy hang jelölésére karakterkombinációkat és diakritikus jeleket (akár egyszerre is) használ, például: *cs* → *ch* ~ *chy* ~ *cyh* ~ *c* ~ *chi* ~ *č* ~ *ch'*. Az ómagyar kor több mint 6 évszázadot fog át, amelynek során nem volt egységes hangjelölési rendszer, sőt egy kódexet akár több kéz is jegyezhetett, ami további egyenetlenségeket okoz a szövegekben. A különböző helyesírási rendszerekben is ritka az egy hang – egy betű megfelelés (vagyis amikor egy hang jelölésére mindig ugyanaz a betű használatos, és az adott betűnek mindig egy hangértéke van), de

egy alakulóban levő helyesírási rendszerben ilyenfajta következetesség még kevésbé van jelen. Sőt inkább az a tipikus, hogy egy emléken belül is ingadozik egy-egy hang jelölésmódja (pl. *kinec* [*kinek*]), vagy többes hangértéke van egy-egy betűnek (pl. *gimilcíctul* [*gyümölcsöktől*]). Tovább bonyolítja a helyzetet, hogy néhány betű egyaránt utalhat magánhangzóra és mássalhangzóra is, például az *u, v, w* több évszázadon át jelölhette az *u, ú, ü, ű, v, β* hangok bármelyikét (Korompay 2003). Ebből kifolyólag igen magas a speciális karakterek száma: az Ómagyar Korpuszban az 52 latin alapkarakter mellett 42 diakritikus jel, 10 szám, 34 szövegtagoló és egyéb jel, 3 görög betű, valamint 15 egyéb speciális karakter fordul elő, mindösszesen 156 karakter plusz ezek kombinációi. Vagyis már a betűhű szövegváltozat előállítása is sokkal nagyobb kihívást jelent, mint a mai magyar sztenderd nyelvváltozat esetében.

Ez a heterogén helyesírás az oka annak is, hogy a történeti szövegek esetében szükség van egy normalizáló lépésre, amelynek során az eredeti betűhű szóalakokat mai magyar helyesírású szavakra alakítjuk át. A normalizálás nehézsége abban rejlik, hogy egyszerre kell arra törekedni, hogy a helyesírási esetlegességeket kiküszöböljük, és eközben minden, ma már esetleg nem létező nyelvi jelenséget is megőrizzünk.

Ha magasabb szintű nyelvi annotációt szeretnénk a korpuszhoz adni, akkor ahhoz nyilván szükség van az adott szintű elemzőkre. Viszont az elérhető elemzők a modern nyelvállapotra készültek, ezeket alkalmassá kell tenni a régi nyelvállapot elemzésére, ami közelről sem triviális feladat. Akármilyen automatikus elemzőt használunk is, a kimenet mindenképpen kézi ellenőrzést igényel.

4. Magyar nyelvű történeti korpuszok

A magyar nyelvű történeti korpuszok sorában a magyar nyelv történeti szakaszait időrendi sorrendben követve az első az Ómagyar Korpusz. Ez a korpusz az MTA Nyelvtudományi Intézetében készül, a Magyar Generatív Történeti Szintaxis projekt keretében. A korpusz tartalmazza az összes fennmaradt ómagyar kori (896–1526) és néhány középmagyar kori (1526–1772) szövegemléket, valamint számos középmagyar bibliafordítást. A feldolgozott anyag 47 ómagyar kódexet, 24 rövidebb ómagyar szövegemléket, 244 misszilist (elküldött levelet), valamint 5 középmagyar kori bibliafordítást foglal magában, jelenleg mindösszesen 3,2 millió szövegszót. A korpusz egy része normalizálva és morfológiailag elemezve is lett; a normalizált alkorpusz mérete 807.691 token, a morfológiailag elemzett

alkorpusz mérete 285.070 token. A korpusz weboldalán keresztül a teljes anyag betűhű szövege, valamint néhány nyelvemlék normalizált és morfológiailag elemzett változata elérhető, valamint kereshető a grafikus korpuszlekérdező felület segítségével. A korpusz felépítéséről és a korpuszépítés lépéseiről részletes leírást szolgáltat a weboldal, valamint Oravecz et al. (2009, 2010); Simon and Sass (2012); Simon et al. (2011); Simon (2014); Simon and Vincze (2016).

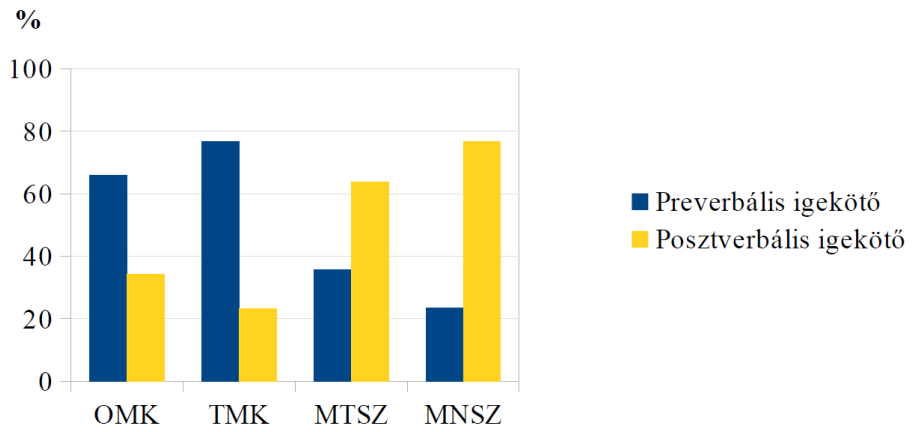
Az időben következő korpusz a Történeti Magánéleti Korpusz (Dömötör et al. 2017; Novák et al. 2018), amely szintén az MTA Nyelvtudományi Intézetében készült. A korpusz az ó- és a középmagyar kor magánéleti nyelvi regiszteréhez legközelebb álló műfajokat tartalmazza: magánlevelekből és peres eljárások jegyzőkönyveiből épül fel. Jelenleg körülbelül 850 ezer normalizált és morfológiailag elemzett szövegszót tartalmaz. A weboldalon[4] találunk leírást a felhasznált forrásokról, az alkalmazott morfoszintaktikai címkékről, valamint egy keresőfelületet és segédletet a használatához.

A Magyar Történeti Szövegtár (Csengery 2006) eredetileg a Magyar nyelv nagyszótárához (Ittzés 2011) készült, szintén az MTA Nyelvtudományi Intézetében. Ez a korpusz 1772-től, vagyis az újmagyar kor kezdetétől tartalmaz szövegeket a 20. század végéig. A régi keresőfelülete mellett készült hozzá egy újabb (Sass 2017), amely mögött a NoSketchEngine (Rychlý 2007) szabadon elérhető korpuszkezelő motor működik, és más magyar nyelvű korpuszok, mint például a Magyar Nemzeti Szövegtár (Oravecz et al. 2014) lekérdezőjéhez hasonlóan a CQL lekérdezőnyelvet használja. Ez a korpusz – a fenti kettővel ellentétben – nem tartalmaz semmilyen nyelvi annotációt, vagyis a keresésnél csak a felszíni szóalakra tudunk támaszkodni.

Mindhárom keresőre igaz, hogy elő tudunk velük állítani konkordancialistát és gyakorisági listát is. Mindhárom alkalmas arra, hogy történeti lexikológiai és szociolingvisztikai kutatásokhoz segítséget nyújtson. A morfológiai elemzést is tartalmazó korpuszok természetesen lehetőséget adnak történeti morfológiai vizsgálatok folytatására is, illetve bizonyos szintaktikai jelenségek is kutathatóak rajtuk. Például Kalivoda (2017) hat prototipikus igekötő (*meg*, *el*, *fel*, *ki*, *be*, *le*) szintaktikai viselkedését vizsgálta az ómagyar kortól napjainkig. A kutatás a felsorolt igekötők és a hozzájuk tartozó finit igék egymáshoz viszonyított helyét számolja, és a tagadó és a tiltó mondatokra jellemző igekötő–ige, illetve ige–igekötő sorrend arányát nézi. É. Kiss (2014) és Gugán (2015, 2017) azt állítja, hogy kétféle tagadó

[4] http://tmk.nytud.hu

szerkezet létezik a magyarban: a régebbi az egyenes szórendű (igekötő–tagadószó–ige, pl. *meg ne fogd*), az újabb pedig a fordított szórendű (tagadószó–ige–igekötő, pl. *ne fogd meg*). Állításuk szerint a fordított szórend is létezett a magyar nyelv korábbi szakaszaiban is, de csak a 19. századtól válik uralkodóvá. Kalivoda (2017) a preverbális (ige előtti) és a posztverbális (ige utáni) igekötők arányát vizsgálta a fent ismertetett három magyar történeti korpuszban, míg a mai nyelvállapot vizsgálatára a szintén említett Magyar Nemzeti Szövegtárt használta. A 4. ábra az egyenes és fordított szórendű tagadó mondatok százalékos arányát mutatja a vizsgált korpuszokban. A diagramon azt látjuk, hogy a posztverbális igekötőt tartalmazó tagadó mondatok arányának növekedése az ómagyar kortól napjainkig egyértelműen kimutatható a történeti korpuszok segítségével.



4. ábra: A preverbális és posztverbális igekötők arányaa Kalivoda (2017) által vizsgált korpuszokban

Irodalom

Biber, D. 1993. Representativeness of Corpus Design. *Literary and Linguistic Computing* 8/4.

Claridge, C. 2008. Historical corpora. In: Lüdeling, A., Kytö, M. (szerk.). *Corpus Linguistics. An International Handbook*. Berlin: Walter de Gruyter. 242–259.

Csengery K. 2006. Az elektronikus korpusz. In: Ittzés N. (szerk.). *A magyar nyelv nagyszótára 1. Segédletek.* Budapest: MTA Nyelvtudományi Intézet.

Dömötör A., Gugán K., Novák A., Varga M. 2017. Kiútkeresés a morfológiai labirintusból – korpuszépítés ó- és középmagyar kori magánéleti szövegekből. *Nyelvtudományi Közlemények* 113. 85–110.

É. Kiss K. 2014. A tagadó és a kérdő mondatok változásai. In: É. Kiss K. (szerk.). *Magyar generatív történeti mondattan*. Budapest: Akadémiai Kiadó. 34–49.

Gugán K. 2015. És mégis: mozog? Tagadás és igemódosítók az ómagyarban és a középmagyarban. *Általános Nyelvészeti Tanulmányok* 27. 153–178.

Gugán K. 2017. A magyar tagadó mondatok szórendje és a konstansrátahipotézis. In: *Nyelvelmélet és diakrónia 3*. Budapest; Piliscsaba: Pázmány Péter Katolikus Egyetem BTK; Szt. István Társulat. 91–110.

Hunston, S. 2008. Collection strategies and design decisions. In: Lüdeling, A. Kytö, M. (szerk.). *Corpus Linguistics. An International Handbook*. Berlin: Walter de Gruyter. 154–167.

Ittzés N. 2011. *A magyar nyelv nagyszótárának lexikográfiai koncepciója, különös tekintettel a szemantika és a grammatika összefüggésére a szótárírásban.* PhD-értekezés. Szeged: Szegedi Tudományegyetem.

Kalivoda Á. 2017. *Prototipikus igekötők mondatbeli helye az ómagyar kortól napjainkig*. Előadás a PPKE BTK Nyelvtudományi Doktori Iskolájának házi doktoranduszkonferenciáján.

Károli G. 1590. *SZENT Biblia, az az Istennec O es Wy testamentvmanac prophétác es apostoloc által meg iratott szent könyuei*. Vizsoly.

Károli G. 1908. *Szent Biblia, azaz: Istennek Ó és Új Testamentomában foglaltatott egész Szent Írás*. Magyar nyelvre fordította Károli Gáspár. Az eredetivel egybevetett és átdolgozott kiadás. Budapest: Brit és Külföldi Biblia-Társulat.

Kniezsa I. 1952. *Helyesírásunk története a könyvnyomtatás koráig*. Budapest: Akadémiai Kiadó.

Korompay K. 2003. Helyesírás-történet (az ómagyar korban). In: Kiss J., Pusztai F. (szerk.). *Magyar nyelvtörténet*. Budapest: Osiris Kiadó.

Madas E. (szerk.). 2009. *„Látjátok feleim...” Magyar nyelvemlékek a kezdetektől a 16. század elejéig*. Budapest: Országos Széchényi Könyvtár.

McEnery, T. 2004. Corpus Linguistics. In: Mitkov, R. (szerk.). *The Oxford Handbook of Computational Linguistics*. New York: Oxford University Press. 448–463.

Milić, L. 1990. The Century of Prose Corpus. *Literary and Linguistic Computing* 5/3. 203–208.

Novák A., Gugán K., Varga M., Dömötör A. 2018. Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. *Language Resources and Evaluation* 52/1. 1–28.

Oravecz Cs., Sass B., Simon E. 2009. Gépi tanulási módszerek ómagyar kori szövegek normalizálására. In: Tanács A., Szauter D., Vincze V. (szerk.). *VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2009)*. Szeged: Szegedi Tudományegyetem. 317–324.

Oravecz Cs., Sass B., Simon E. 2010. Semi-automatic Normalization of Old Hungarian Codices. In: *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*. Lisbon: Faculty of Science, University of Lisbon. 55–60.

Oravecz Cs., Váradi T., Sass B. 2014. The Hungarian Gigaword Corpus. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014)*. Reykjavík: European Language Resources Association.

P. Balázs J., Dömötör A., Pólya K. (szerk.). 1990. *Szent Margit élete, 1510. A nyelvemlék hasonmása és betűhű átirata bevezetéssel és jegyzetekkel*, volume 10 of *Régi magyar kódexek*. Budapest: Magyar Nyelvtudományi Társaság.

Rissanen, M., Kytö, M., Kahlas-Tarkka, L., Kilpiö, M., Nevanlinna, S., Taavitsainen, I., Nevalainen, T., Raumolin-Brunberg, H. (szerk.). 1991. *The Helsinki Corpus of English Texts.* Helsinki: University of Helsinki.

Rychlý, P. 2007. Manatee/Bonito – A modular corpus manager. In: *Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University. 65–70.

Sass B. 2017. Keresés korpuszban: a kibővített Magyar történeti szövegtár új keresőfelülete. In: Forgács T., Németh M., and Sinkovics B. (szerk.). *A nyelvtörténeti kutatások újabb eredményei IX*. Szeged: SZTE Magyar Nyelvészeti Tanszék. 267–277.

Simon E. 2014. Corpus building from Old Hungarian codices. In: É. Kiss K., (szerk.). *The Evolution of Functional Left Perpiheries in Hungarian Syntax*. Oxford: Oxford University Press. 224–236.

Simon E. Sass B. 2012. Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből. In: Prószéky G., Váradi T. (szerk.). *Általános Nyelvészeti Tanulmányok XXIV*. *Nyelvtechnológiai kutatások*. Budapest: Akadémiai Kiadó. 243–264.

Simon E., Sass B., Mittelholcz I. 2011. Korpuszépítés ómagyar kódexekből. In: Tanács A. Vincze V. (szerk.). *VIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem. 81–89.

Simon E., Vincze V. 2016. Universal Morphology for Old Hungarian. In: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Berlin: Association for Computational Linguistics. 118–127.