szövegrész ugyanúgy kiemelt, mint a mondatokat tartalmazó találati listában. A kiemelt részre rá lehet keresni a teljes dokumentum szövegében is, így a találat szélesebb kontextusa könnyen áttekinthető.

A keresőrendszer lehetővé teszi a találatok egyszerűsített, csak az eredeti változatot tartalmazó megjelenítését is. Ennek az egyszerűsített megjelenítési formának a bevezetését az tette szükségessé, hogy a rendszer által alapesetben visszaadott, annotációt is tartalmazó reprezentáció nem jelenik meg megfelelően az általánosan használt szövegszerkesztő programokba átmásolva. Az egyszerűsített kimenet ezzel szemben szövegszerkesztőbe másolható, így a korpuszt használó kutatók könnyen idézni tudják a találatokat a kutatásaikat bemutató tanulmányokban. További lehetőségként a szövegkörnyezet teljes mellőzésével megjeleníthető a keresőkifejezésre illeszkedő szavak, kifejezések gyakoriság szerint rendezett listája is.

Összefoglalás

Cikkünkben bemutattuk egy ó- és középmagyar történeti korpusz létrehozásának lépéseit, melyek során a nyelvészeti feladatok egy részét kézzel, más részét pedig a nyelvtechnológia eszközeit felhasználva automatikusan végeztünk. Az adatgyűjtés és a nyersanyagok digitalizálása után elkészült a szövegek mai magyar helyesírásnak megfelelő átirata és morfológiai elemzése. Az adatbázist egy webes felületen keresztül tettük elérhetővé és kereshetővé, lehetővé téve a kutatók és a nagyközönség számára is a feldolgozott korszakok nyelvi kincsei között való kutakodást.

A munkálatot az OTKA K 81189 és 116217 sz. pályázata támogatta, illetve támogatja.

Kulcsszavak: *elektronikus adatbázis, nyelvtörténet, ó- és középmagyar kor, magánéleti regiszter, morfológia, elemzőprogram, keresőfelület*

IRODALOM

Dömötör Adrienne (2013): Nyelvtani elemzésekkel ellátott online szöveggyűjtemény. Nádasdy-levelektől a boszorkányperekig. *Élet és tudomány.* 43, 1363–1365.

Dömötör Adrienne (2014): Az ó- és középmagyar kori magánéleti nyelvhasználat morfológiailag elemzett adatbázisa. In: Fazakas Emese – Juhász D. – T. Szabó Cs. –Terbe E. – Zsemlyei B. (szerk.): *Tér, idő, társadalom és kultúra metszéspontjai a magyar nyelvben.* ELTE Magyar Nyelvtörténeti, Szociolingvisztikai, Dialektológiai Tanszék– Nemzetközi Magyarságtudományi Társaság, Budapest–Kolozsvár, 11–21.

Novák Attila – Gugán K. – Varga M. – Dömötör A. (2015): *Creation of an Annotated Corpus of Old and Middle Hungarian Court Records and Private Correspondence.* Kézirat. • http://tinyurl.com/jg2mxkc

Novák Attila (2003): Milyen a jó humor? In: Alexin Zoltán – Csendes Dóra (szerk.): *Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, 138–145. • http://tinyurl.com/juc5979

Orosz György – Novák Attila (2013): PurePos 2.0: A Hybrid Tool For Morphological Disambiguation. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, 2013.* Incoma, Shoumen. 539–45. • http://tinyurl.com/hfgf8z2

Petersen, Ulrik (2004): Emdros – A Text Database Engine for Analyzed or Annotated Text. In: *Proceedings of the 20th International Conference on Computational Linguistics,* Volume II. 1190–1193. DOI:10.3115/1220355.1220527 • http://tinyurl.com/j3deohy

URL1: www.tmk.nytud.hu

EGY XVIII. SZÁZADI ÍRÓI KORPUSZ MODERN FELDOLGOZÁSA

Kiss Margit

PhD, tudományos munkatárs,
MTA Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézete
kiss.margit@btk.mta.hu

A digitális Mikes-szótár

A digitális Mikes-szótár projekt (OTKA K 81 337, témavezetője Tüskés Gábor) indulásakor két alapvető célt tűzött ki maga elé: egyfelől a hazai írói szótárirodalom hiánypótló művének elkészítését, másfelől a digitális szótárkészítés módszerének kialakítását ebben a speciális műfajban. Ezzel a vállalkozással az első teljes életművet feldolgozó magyar elektronikus írói szótár készítése vette kezdetét, eredményeiben azonban jóval tovább mutat a vállaltaknál.

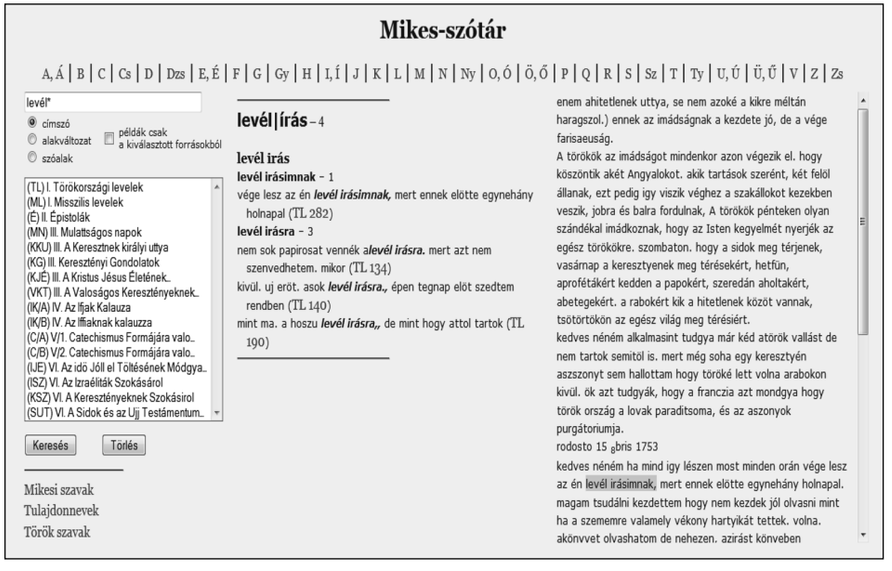
Az írói szótár fő célja, hogy egy-egy szerző életművének meghatározó részét vagy akár a teljes szókincs alapján az egészét bemutatva láttassa a nyelvhasználat, a stílus, a szó- és kifejezéskészlet elemeit. Mivel a papíralapú szótárakban lényegében egy-egy szócikkre fókuszálunk, a kereshetőség meglehetősen redukált, ami többnyire egy-egy szó, egy-egy adott szöveghely értelmezésére koncentrálódik. A digitális szótár segítségével az adatok rendszerezése és megtalálhatósága több szempontból is kedvezőbbé válik a papíralapú kötetekkel szemben. A kötetbeni megjelenés számos korlátozó funkció alkalmazását teszi szükségessé, így például a példaanyag szűkítését, bonyolult utalórendszer, terjedelmi megszorítások kialakítását. Ezzel szemben a digitális szótárakban nem indokoltak a használhatóságot megnehezítő, sokszor bonyolult korlátozások; a terjedelmi korlátok megszűnnek, sőt az adatok közti kapcsolatrendszer kialakításának az informatikai környezet kedvez igazán. A keresések gyorssá, pontossá és hatékonnyá válnak. A kutatásokban a digitális szótár egyfajta módszertani szemléletváltást is eredményez: itt azoknak az adatoknak a lekérdezésére is lehetőség nyílik, amelyekre ezidáig a papíralapú szótárak használatával egyáltalán nem vagy csak igen jelentős energiaráfordítással lehetett esély. Ezáltal új összefüggések megfogalmazására is mód nyílik. E módszertani váltás eredményeképpen a szótár hozzáadott értéke jóval több lesz, mint a hagyományos értelemben vett szótáré. Segítségével nemcsak korábbi megállapításokat vizsgálhatunk, hanem új kutatási irányokat is kijelölhetünk, amelyek még feltáratlanok, vagy nem volt mód, eszköz, lehetőség a vizsgálatukra. Ha nemcsak a nagy terjedelmű szövegek tárolási kapacitását látjuk egyedüli lehetőségként az informatika nyújtotta előnyök között, hanem olyan számítógép támogatta módszereket is alkalmazunk, amelyekkel más perspektívából vizsgálhatjuk, elemezhetjük és értelmezhetjüük a szövegeket,

a digitális írói szótár új eredmények elérésében nyújt segítséget irodalomtörténeti, nyelvészeti, filológiai szempontból egyaránt.

Mikes életműve hatezer oldalnyi terjedelmével, összesen másfél millió szavával önmagában is kiemelkedik a hazai írói életművek sorából: ez a mennyiség háromszor több Petőfi és ötször több József Attila írásainál, de *A magyar irodalmi és köznyelv szótárának* – közismertebb nevén a *Nagyszótárnak* – elektronikus szövegkorpusza, amely 1772-től napjainkig tartalmaz szövegeket, jelenleg mindössze tizennyolcszorosa a Mikes-korpusznak. Az életmű szótári feldolgozása (URL1) a terjedelem miatt aligha képzelhető el informatikai támogatás nélkül, de nem nélkülözhető a manuális munka sem. (A szótári munkafolyamatokról részletesebben lásd Kiss, 2012). Az elmúlt századok szövegei nem szabályozottak, sokszor még ugyanannak a szerzőnek különböző műveiben is ingadozik az írásmód, így a központozás hiánya, a szavak, szerkezetek szegmentálási anomáliái jelentős mértékben nehezítették és lassították a gyors és automatikus elemzést. Az előzetes kézi feldolgozás a későbbi gépi eljárásokat támogatja, hiszen a mai címszavakhoz történő rendezés a gépi kereshetőség alapfeltétele. A szótárszerkesztésben kulcsfontosságú, hogy a lehető legoptimálisabban segítsük a számtalan eltérő változatú és régies írásmódú szó keresését. Nemcsak az a Mikes-szótár előnye, hogy a teljes kritikai kiadás szöveganyagát dolgozza fel, hanem az is, hogy jelentősen megkönnyíti a szavak kereshetőségét: akár a mai alak alapján, akár bármely régies alakváltozat alapján, akár ragozott alakokkal végzünk lekérdezéseket, minden esetben eljuthatunk a vonatkozó szócikkhez. A szavak megtalálhatóságát optimalizáltuk a különböző címszóváriánsok felvételével és a más címszavakkal fennálló kapcsolódások láttatásával. A keresési spektrum szélesítésével a szótárhasználó megtalálja a keresett szót akkor is, ha több variánsa van, s az általa keresett variáns eltér a szövegből meghatározható címszóváltozattól. Ezenkívül egyéb – különböző szavak közti – összefüggésekre is rávilágítunk. Minderre azért fektettünk hangsúlyt, mert ha a tökéletlen keresések miatt pontatlan eredményeket kapunk, az erre támaszkodó kutatások is tévútra vihetnek. Választ kaphatunk azokra a kérdésekre is, hogy Mikes mely szavakat használt és melyeket nem, így például megvan-e nála a *csoda, kanál* szó – függetlenül attól, hogy leírta-e ebben a mai formájában vagy sem. Megtudhatjuk azt is, hogy melyek azok a szavak, amelyek az író önálló szóalkotásai, és más szótárban nem találjuk meg őket, mint például a *halálszolgálat, varrótűeső* stb. Az utalások segítségével eljuthatunk más, vonatkozó címszavakhoz akár szótáron belül, akár más szótárak állományához, de listázhatjuk a török eredetű szavakat és a tulajdonneveket is (Kiss, 2014). A kialakított eljárások más hasonló munkálatokban is eredményesen alkalmazhatók. A szűkebb értelemben vett szótári, de általában véve az informatikai támogatású feldolgozási módszerek lehetővé teszik, hogy a korábbiaktól eltérő módon vizsgálhassunk terjedelmes szövegeket *(1. ábra).*

Új módszerek a filológiai vizsgálatokhoz

A Mikes-filológiában az új módszerek alkalmazási területeinek feltérképezéseinél járunk, de már ebben a korai fázisban is figyelemre méltó eredmények születtek irodalomtörténeti, stilisztikai, grammatikatörténeti és az írói alkotófolyamatok alaposabb megismerése szempontjából. A digitális szótár szerepe az írói életmű vizsgálatában túllép a kizárólagos szótári funkciókon. A számadatok, statisztikák



1. ábra

és a tetszőlegesen szűrt, különféle szempontok szerint rendezett szólisták új összefüggések láttatására alkalmasak. A *Törökországi levelek* mellett a tizenhétszer nagyobb mennyiségű, franciából átültetett Mikes-fordítások kevésbé kerültek ezidáig a figyelem középpontjába: a saját szerzőségű művel szemben az irodalmi kánonban nem töltöttek be jelentős szerepet. A mintegy tízezer elkészült szócikk alapján végzett vizsgálat cáfolni látszik a korábbi megállapításokat. A *Leveleskönyv* és a fordítások címszavainak összevetéséből arra következtethetünk, hogy a szókészlet markáns elkülönülése miatt nemcsak a Mikes-kutatásokból, de a kortörténeti szintézisekből sem mellőzhetők a fordítások szavai. Az előremutató nyelvi, stilisztikai változásokat valójában a fordítások szókészletállományában fedezhetjük fel. Eszerint – szemben az eddigi vélekedésekkel – fordításaiban Mikes nemcsak a koránál, hanem a saját szerzőségű műveinél is modernebb szemléletet képviselt. A XVIII. században a szóösszetétel mint szóalkotási forma meghatározó volt a még csak kialakulóban lévő szaknyelvi műszavak létrehozásában. A fordításokban találni jelentősebb mennyiségű összetételt, és itt látszanak a XIX. századot megelőlegező, újító jegyek.

Ahogy a maga idejében a Petőfi-szótár a *-ság, -ség* képző speciális használatára irányította a figyelmet, úgy a Mikes-szótárban a konkordancialista feldolgozása során is kirajzolódtak azok a grammatikai kérdések, amelyek részint a szerzői nyelvhasználatot (például a *való* sajátos szerepe, kötőszói funkciók, igeidők, egyeztetés, bővítmények tanulmányozása), részint a korszak nyelvi állapotát (például az ikes paradigma fellazulásának a folyamata, a *mondván* idéző szerepben és grammatikalizációs jelenségek) láttatják új perspektívából (Kiss – T. Somogyi, 2014). A különféle szólisták előállítása teremti meg az alapot ezekhez az elemzésekhez. A *mondván* esetében a határozói igenévből a mára már a kötőszóvá

válás határára jutott folyamatnak a középmagyar kor az egyik jelentős állomása (Dömötör, 2013). Ennek a vizsgálatnak a Mikes-korpusz is része volt, amelyet tovább elemeztünk művek szerinti eloszlásban is. Azt tapasztaltuk, hogy Mikes funkcionálisan, vélhetően tudatosan használta a kétféle idézési mód (egyenes és függő) kifejezésére és elkülönítésére a *mondván*-t: a folyamatos, gördülékeny, az élőbeszéd jegyeit mutató függő beszéd beépítésére, valamint a biblikus, archaikus hangnemet sugalló egyenes idézés bevezetésére. Mindez részletkérdésnek tűnhet, ám távolabbi perspektívából nézve az írói szövegalkotási folyamatokra is tudunk az ehhez hasonló adatokból következtetni. A vizsgálat ráirányította a figyelmet arra is, hogy a saját szerzőségű művein kívül Mikes fordítói módszere szintén elismerésre méltó stilisztikai teljesítmény. A szókincsvizsgálat mellett a grammatikai elemzések szintén azt támasztották alá, hogy az életmű eddig kevéssé feldolgozott és értékelt részének, a fordításoknak a mellőzése vagy háttérbe szorítása revideálandó.

A kritikai kiadásban az életmű papíralapú feldolgozása korábban nem tett lehetővé ilyen típusú szókeresésekre, statisztikai vizsgálatokra is támaszkodó kutatásokat, de a szónál nagyobb egységeket érintő vizsgálatokra is korszerűbb lehetőségünk van már. Az informatikai támogatású szövegelemzések új eredményekkel szolgáltak. A *Törökországi levelek* kollokviális előadásmódja, közvetlen, társalkodó hangvétele sokáig központi eleme volt a Mikes-elemzéseknek, és kevesebb figyelem jutott arra, hogy a levélírói alkotófolyamatban meghatározó szerephez jutott az azzal párhuzamosan végzett fordítói munka. Hopp Lajos a kritikai kiadás elkészítését követően megállapította, hogy a *Leveleskönyv* szövegének mintegy 23%-a a levelekbe szőtt fordításbetét (Hopp, 2002). A szótári munka során a konkordancialista feldolgozásakor kiderült, hogy az életmű különböző, olykor távoli pontjai között szövegszerű összefüggések rajzolódnak ki; vannak súlyozott és kevésbé érintett területei e tekintetben. A részletes összefüggések feltárására ezidáig a nagy szövegterjedelem és a technológia hiánya miatt nem volt mód. Olyan szöveghasonlóságokról és -párhuzamokról van szó, amelyeknél ugyan nem áll fenn a teljes szövegegyezés, de kisebb variánsokat találni a szövegrészletek között, és manuális eszközökkel nincs lehetőség kigyűjteni ezeket. Emiatt automatizáltuk a szövegvándorlások számítógépes támogatással történő listázását a történeti szövegkorpuszból. Nemcsak a szóról szóra meglévő egyezést tudjuk már lajstromozni a korpuszban, hanem azokat a variánsokat is meg tudjuk találni, amelyeknél valamiféle eltérés (például írásmódbeli különbség, kisebb alaki variációk, közbetoldások) mutatkozik a szövegben, de fennáll a tágabb értelemben vett szövegazonosság vagy szöveghasonlóság. E módszer jelentősége abban áll, hogy mindezt célzott szókeresés nélkül érhetjük el XVIII. századi szövegben. A szövegkritikai vizsgálatok következő lépése a számítógép által készített gyűjtés módszeres feldolgozása. Már a részletes elemzés és alapos feldolgozás előtt is látszik, hogy jelentős mennyiségű szöveg vándorol a művekben; esetenként akár hét különböző helyen is előfordul ugyanaz a (tag)mondatnyi részlet. Ezzel a módszerrel összehasonlíthatatlanul többet fogunk tudni az írói alkotófolyamatokról, szerzői módszerekről az egész életmű alapján.

Összegzés, jövőbeni feladatok

Az alkalmazott módszereket, néhány részterületet és kutatási irányt felvázolva arra szeretnek volna felhívni a figyelmet, hogy a digitális szótár teremtette új módszerek alkalmazásával jóval közelebb kerültünk a Mikes-filológia nyitott kérdéseinek megválaszolásához. Tisztázásra vár, hogy a különböző fordításaiban a szerző miért kezelte eltérő módon a bibliai idézeteket; a Káldi-féle Biblia hatása mellett Károli fordításának a nyomai hol, milyen mértékben érhetők tetten. Az új módszerekkel könnyebbé válhat mindezek pontosítása, a további rokonszövegek feltárása és a szövegek továbbélésének felderítése. További feladat, hogy kideríts ük, bizonyos fordítások esetében az adott műnek mely kiadása szolgált forrásul, és elvégezzük az alkotófolyamatok mélyrehatóbb tanulmányozását. A teljes életművet aprólékosabban, hatékonyabban és új perspektívából tudjuk vizsgálni.

Terveink között szerepel a szóértelmezések rendszerének kialakítása. A kontextuális jelentések definícióinak megadása a hagyományos módszerekkel számos szubjektív megoldást rejt, és az egyes szöveghelyhez történő kötődésük miatt izoláltan nem lennének alkalmasak összefüggéseiben és egészében láttatni a szerzői világképet. Ezért a jelentések megadását meghatározott fogalomkörökhöz történő besorolással végeznénk el egyfajta ontológiát létrehozva, amely lehetőséget teremtene informatikai feldolgozások számára is.

A digitális Mikes-szótár készítése korszerű alapokra helyezte a lexikográfiai munkákat. Új kutatási területek körvonalazódnak, amelyek túlmutatnak a szűkebb értelemben vett szótári tevékenységen. Az informatika alkalmazásának és a megváltozott eszköztárnak köszönhetően Mikes életművének feldolgozásával a kutatási módszertant is szeretnénk továbbfejleszteni és kiegészíteni a szövegek összehasonlításának különféle módozataival. Az MTA Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézete és a Budapesti Műszaki és Gazdaságtudományi Egyetem Villamosmérnöki és Informatikai Kar Méréstechnika és Információs Rendszerek Tanszékének együttműködése megteremtette a megfelelő alapot ehhez: első lépésben megkezdtük a kritikai jegyzetanyag korszerű feldolgozását egy kibővített szótár létrehozásának érdekében. Csak úgy érhetők el jelentős eredmények, ha a bölcsészet és az informatika közös célok megfogalmazására válik képessé.

Kulcsszavak: *digitális írói szótár, Mikes, korpusz*

IRODALOM

Dömötör Adrienne (2013): Idéző szerkezetből diskurzusjelölő elem: a mondván szerepei és története. In: Csepregi Márta – Kubínyi K. – Jari Sivonen (szerk.): *Grammatika és kontextus. Új szempontok az uráli nyelvek kutatásában* III. ELTE, Budapest, 20–30. • https://edit.elte.hu/xmlui/handle/10831/9785

Hopp Lajos (2002): *A fordító Mikes Kelemen.* (Tüskés Gábor szerk.) (*Historia Litteraria* 12) Universitas, Budapest

Kiss Margit (2012): A digitális Mikes-szótár. *Magyar Tudomány* 173, 3, 279–284. • http://www.matud.iif.hu/2012/03/04.htm

Kiss Margit (2014): Mit tészen ez a szó? – Az elektronikus Mikes-szótár. In: Korompay Klára – Stemler Á. – Terbe E. – C. Vladár Zs. (szerk): *Fornáskutatás, fornáskiadás, tudománytörténet II.* Magyar Nyelvtudományi Társaság, Budapest, 48–56. • http://real.mtak.hu/30965/1/KissM.pdf

Kiss Margit – T. Somogyi Magda (2014): A digitális Mikes-szótár: nyelvtörténet és számítógépes lexikográfia találkozása. In: Ladányi Mária – Vladár Zs. – Hrenek É. (szerk.): *Nyelv – társadalom – kultúra. Interkulturális és multikulturális perspektívák I–II.* (*MANYE* XXIII) Tinta, Budapest, 651–657.

URL1: http://mikesszotar.iti.mta.hu/