A digitális Mikes-szótár: nyelvtörténet és számítógépes lexikográfia találkozása

Kiss Margit – T. Somogyi Magda

Kulcsszavak: Mikes, digitális írói szótár, nyelvtörténet, filológia

1. A készülő Mikes-szótár bemutatása, jellemzői

A Mikes-szótár a második valóban teljes írói szótárunk lesz, ugyanakkor első az elektronikus feldolgozás és megjelenítés tekintetében. Alapjául a teljes életművet felölelő, mintegy 6000 oldal terjedelmű kritikai kiadás szolgál (Mikes 1966–88), amelynek digitalizálása a MEK-kel kötött együttműködési szerződés keretében jött létre. E munka eredményeképpen az író összes művének betűhív átirata elektronikus formában is olvasható a http://mek.oszk.hu/09000/09000/index.phtml oldalon.

A szkennelés, szövegfelismertetés, korrektúra mellett kialakítottuk a mikesi szövegeknek megfelelő XML-struktúrát. Olyan tageket, más néven címkéket illesztettünk a szövegbe, amelyek a szócikkírás során nélkülözhetetlenek: például címek, címfokozatok, versbetétek, idegen nyelvű szövegek, rövidítések (nyelvek szerint is), margináliák, széljegyzetek, idézetek, fordításbetétek. Létrehoztuk a művek címén alapuló rövidítésjegyzéket is, amely a locusok feloldásához nélkülözhetetlen, például É: *Épistolák*, IK/B: *Az Iffiaknak kalauzza*, KKU: *A Keresztnek királyi uttya*, ML: *Misszilis levelek*, MN: *Mulattságos napok*, TL: *Törökországi levelek*, VKT: *A Valoságos Keresztényeknek Tüköre*.

A szövegek ily módon való rögzítése lehetővé tette számunkra azt is, hogy (Mártonfi Attila segítségével) különböző szempontok szerinti gyakorisági listákat készítsünk, beleértve a konkordancialistát is. A gyakorisági listák közül elkészült: a rövidítések gyakorisági listája, a szóalakok művenkénti gyakorisági listája, a szóalakok betűrendes mutatója a gyakorisági értékekkel, a szóalakok gyakorisági listája, amelyekből újabb szempontokat és adatokat ismerhetünk meg Mikes nyelvhasználatáról. Néhány informatív számadat a gyakorisági listákból: a szövegszavak száma (Mikes minden egyes szava) 1,5 millió, a szótár várható címszavainak száma kb. 20 ezer, a különféle szóalakok száma 162 ezer, a teljes konkordancialista mennyisége kb. 60 ezer oldal; az 50 feletti szóelőfordulással rendelkező szóalakok száma is igen magas értéket mutat, 1352.

A következőkben röviden összefoglaljuk az eddigi eredményeket és a várható feladatokat. Maga az adatbázis a középmagyar kor hiánypótló korpusza is egyben. A szótári feldolgozással lehetőség nyílik olyan grammatikai elemzésekre, rejtett szövegösszefüggések vizsgálatára, az adatok több szempontú csoportosítására, különféle statisztikák létrehozására is, amelyekre eddig nem volt mód. (Bővebben: Kiss–Tüskés 2011, Kiss 2012b.)

2. A szócikkek felépítése

A szócikkekben a címszó után felsoroljuk a szóalakokat is, valamint a hozzájuk tartozó szóelőfordulásokat a kötetek, illetve a kötetekben szereplő művek meghatározott sorrendjében közölve a szövegkörnyezetet a pontos előfordulási hely megadásával. A *gangréna* (egyelőre jelentésmegadás nélküli) címszó esetében ez a következőképpen áll össze:

gangréna
cángréna
cángrénától – 1 (szóalak, szóelőfordulások száma)
és féltik attol a veszet *cángrénától,* a ki is az árviz (TL 113)
gángréna
gángréna – 1 (szóalak, szóelőfordulások száma)
hogy olyan tagok. akikben a *gángréna* el hatot, de ök még (VKT 826)

Az eljárás alapvetően egyszerű, viszont a 60 ezer nyomtatott oldalnyi anyag minden egyes példamondatát át kell nézni, s a megfelelő helyre tenni a kérdéses címszóknak, szóalakoknak megfelelően. Nehézséget jelentenek a maitól eltérő formák legjellegzetesebben a különírás és egybeírás vonatkozásában, mint például a ***beli***: *a császárnak, a belsö udvara* ***beli*** *fö tisztei mind inkáb heréltekböl* (TL 265); **leg**: *és énnekem gyakran kel irni,* ***leg*** *aláb minden héten. hét levelet* (TL 9); ***ajutalomrol***: *Isten rendel választatinak, és elmélkedni* ***ajutalomrol****, melyet igér azoknak. kik végig* (KKU 424); ***felejtesékel***: *meg untanak más új gyönyörüségel.* ***felejtesékel****. hasonlo rendet kel tartani amulattságban* (VKT 900). Ezeket a rendszeres vagy egyszeri, esetleg írásmódbeli esetlegességeket mind egységes rendszerbe kell illeszteni. (A munkafolyamat részletesebb leírását l. Kiss 2012a.)

3. Címszavasítási kérdések

A jelenlegi munkafázisban a konkordancialista felhasználásával rendezzük címszavakba az anyagot. A konkordancialista a számítógép által meghatározott szóalakból és a hozzá tartozó szóelőfordulásból áll a gép által generált kb. tízszavas szövegkörnyezettel együtt. A címszólista készítésének fő szempontja a mai címszóhoz sorolandó szavak rendszerszerű csoportosítása, a szótári szavak és az alak-, valamint a helyesírási változatok meghatározása a mai nyelvből ki nem mutathatók esetében is. Bár az anyag kezelésében nagy segítségünkre van a számítógép, de a sok alak- és írásváltozat szegmentálására a gépi elemzési módszerek alkalmatlanok, ezért minden példát egyenként meg kell vizsgálni a besorolási helyét illetően. A sokváltozatú, régies írásmódú szavakhoz kialakítottunk egy olyan struktúrát, amelyben nyelvrendszerbeli helyük szerint egységes módon tudjuk kezelni őket. Rendszerünk a szószintű szóelőfordulások meghatározásából építkezve az alakváltozatok bemutatásán át alapesetben a mai címszó meghatározásáig terjed; a teljes eredeti szöveget is megtartva eljutunk a mai szókészletig.

3.1. A címszó kiválasztása, meghatározása

Az irányadóak a mai köznyelvi szótárak, illetőleg a legfontosabb történeti és tájszótárak, esetleg írói szótárak. Az alak- és írásváltozatok a címszótól csak kiejtésben, illetőleg helyesírásban térhetnek el. A szótárakban fel nem lelhető szavak címszavasítása egyedenként történik etimológiai összefüggések alapján. Minden egyes címszónál jelöljük azt is, hogy az adott címszót Mikes használta-e mai formájában vagy sem, illetve a címszavakat összevetjük a mai köznyelvi szótárak címszóállományával, valamint a nagyszótári korpusz és a rendelkezésre álló történeti-etimológiai, táj- és írói szótárak szókészletével is, s megjelöljük Mikes semmilyen más szótárban nem szereplő szavait. Ilyenek például: *drágafű, csákósüveg, kontraktuscsinálás, dervisség, dézsmabor, facsésze, szőlőfa, rozmaringfa, fácánnyelv, félberúg, gyapotköntös, háborúságszerető.*

A köznyelvi alak segíti a címszóválasztást, de ha a Mikesnél előforduló alakváltozat kizárólagos, és a mai köznyelvben vagylagosan előfordulhat, akkor választhatjuk ezt is címszónak. Például Mikesnél a *csoda, csodál* változat sohasem szerepel, csak a zártabb forma adatolható, tehát *csuda, csudál, csudaállat, csudálatos, csudálkozik, csudálkozás, csudatétel*, de a digitális keresési lehetőségek miatt a címszóválasztásnak csak elvi jelentősége van.

3.2. A tulajdonnevek mint címszók

Az, hogy a Mikes-szótár készítői a teljesség kedvéért vállalták a tulajdonnevek szótárazását, olyan problémakört hoz magával, amelynek elemzésére itt most nem vállalkozhatunk. A tulajdonnevek – személy- és helynevek – esetében is meg kell határozni a ma leginkább használatos alakot, ami a címszó lehet, azonban sokkal körültekintőbben kell meghatározni az alak- és írásváltozat fogalmát. Külön kell választani a magyar névhasználatot és az idegen nyelvi alakokat, például ha az adat *Casimirus,* a címszó nem lehet ***Kazimír.***

3.3. A török (eredetű) szavak

Sajátos problémát jelentenek a viszonylag gyakran előforduló török szavak, amelyek közül nem egy jövevényszóként ma is megtalálható nyelvünkben, de jó néhány már teljesen ismeretlen a mai olvasó számára. Helyesírásuk, lejegyzésük jelentősen eltér az egyéb történeti munkákban találhatóktól, esetenként a felismerésük is gondot okoz, a címszó meghatározásához nemcsak jelentős hangtörténeti ismeretekre, hanem turkológusok segítségére is szükség van. Több esetben a TESz., valamint egyes írói szótárak adatai is iránymutatóak lehetnek. Ezek után már egyértelmű, hogy a Mikesnél előforduló *csámé* a ***dzsámi*** alakváltozata, a *Caszáp basa* szókapcsolatban a régi *kaszab* 'mészáros' szót azonosíthatjuk, és a *hancsár*, *hansár* változatok a ***handzsár*** címszó alá tartoznak.

3.4. Magyar szavak ~ latin szavak

A latin szavak és kifejezések esetében is hasonló a helyzet. Összhangot kell teremteni a Mikes korában használatos latin és a ma szótározott alakok között. Itt is az alak- és írásváltozat összefüggéseit kell feltárni. A latinos szóhasználatot nem szabad magyar címszóra cserélni. A *decembris* nem szerepelhet a ***december*** változataként, tehát kell lennie külön ***decembris*** címszónak is. A magyar toldalékolás is eligazít a címszóválasztásban. A *competensek* a ***kompetens*** címszó alá kerül mint írásváltozat, de nem vehető ide a latin többes számban álló *competantes*, hiába azonos a jelentésük. Ugyanakkor a latin írásmódú *hymnus* egyszerűen a mai ***himnusz*** írásváltozata. Latin írásmóddal érdemes címszavasítani a *clinicus* és *colligatio* szavak adatait, ezt jelentésük és az IdSzSz. gyakorlata igazolja.

3.5. Az ikesség kérdése a címszóválasztásban

Az, hogy Mikes mely igéket használt ikesen, egyrészt a mai nyelvhasználattal vethető össze, másrészt pedig a korban (a 17. század végétől, 18. század elejétől) már erőteljesen jelentkező keveredés jeleit vizsgálhatjuk. Az ikes igék, illetőleg az ikes paradigma használata szerint több csoportot különíthetünk el. Nemcsak a nyelvleírás számára okozhat gondot a különféle csoportok egyértelmű elhatárolása, lexikográfiai szempontból is számos meggondolandó kérdést vet fel az ikes igék megfelelő címszavasítása. Sok esetben problémát jelenthet az is, hogy – mivel a szótár korpusz alapú – valóban ikes címszót kell-e felvenni akkor is, ha a szóelőfordulások között nincs olyan igealak, amelyből biztosan megállapítható az ikesség, és az adott szó Mikes korában, nyelvjárásában nem ikes ige volt, ma viszont ikesként szoktuk szótárazni. Kérdéses esetekben csak akkor minősíthető ikesnek az ige, és ez alapján a címszó, ha a kitüntetett paradigmahelyekre van biztos adat.

Találhatunk olyan igéket is, amelyek ma iktelennek minősülnek, de Mikes ikesen ragozta őket (pl. *megcsökken**ik**, kihat**ik***). A mai normatív nyelvhasználat az ikes igék ható képzős alakjait iktelennek minősíti, viszont Mikesnél gyakorlatilag nincs példa arra, hogy a szóban forgó *-hAtik* képzős igéket ne ikesen toldalékolná. A *megházasodhatik* 12 adattal szerepel ellenpélda nélkül, ezen kívül sokatmondóak a következő példák is: *dohányozhatom* (TL 32), *meg nem gyarapodhatik* (É 575), *megcsalatkozhatol* (IK/B 585), *gyanakodhatnám* (MN 147).

Jellegzetes csoportot alkotnak azok az igék, amelyek a mai nyelvhasználatban ikesnek számítanak, legalábbis a szótári alakjuk *-ik* végződésű, de Mikes műveiben nyomát se találjuk ennek az *-ik*-nek. Ezeknél a címszóban jelölni kell az ige iktelen voltát. A mai *havazik* a Mikes-szótár adatai alapján egyértelműen iktelennek minősül, tehát a felveendő címszó: ***havaz.*** Ugyanígy kell minősíteni a mai *hazudik* igét is, amelynek teljes mikesi paradigmája a ***hazud*** címszó felvételét indokolja.

3.6. Egyéb címszó-meghatározási problémák

A fentebb tárgyalt kérdéseken túl még sok más probléma is felmerül a címszók meghatározása kapcsán. Ezek közül csak néhányat említünk. Nem gyakran, de előfordulnak népetimologikus alakok is Mikesnél. Ilyen a *kaptán, kapitány,* amelyek a török *kapudán* 'tengernagy' szócikkébe tartoznak, vagyis hiába nem tüntetjük még fel a jelentéseket, a címszavasításnál szinte minden esetben tisztázni kell a szövegkörnyezet és egyéb ismereteink birtokában. A homonimák is csak a jelentés tisztázásával különíthetők el.

4. A Mikes szótár helye a magyar írói szótárak között

Az írói szótárak a vizsgált korpusz mennyiségétől, a vizsgálat céljától és – nemegyszer – a lehetőségektől függően különböző típusokat képviselnek. A magyar írók műveit, életművét felölelő szótárak, szótárjellegű adattárak, konkordanciák az utóbbi évtizedekben gyarapodnak, de számuk még mindig kevésnek mondható. Az általános kérdésekkel még Benkő László foglalkozott 1979-ben, azóta hasonló átfogó mű nem született a kérdéskörrel kapcsolatban; legutóbb Büky László (2010) tekintette át a magyar írói szótárak sorát. Megállapíthatjuk, hogy eddig egyetlen, a szó valódi értelmében teljes írói szótár született, amely a szerző minden egyes leírt és ránk maradt szavát feldolgozta, az idegen szavakkal és a tulajdonnevekkel együtt: ez a Petőfi-szótár.

1. ábra
A címszók száma az egyes írói szótárakban

A fenti grafikon (1. ábra) felsorolja az eddig megjelent írói szótárakat nem téve különbséget a különböző típusok között. Az oszlopok magassága a szótárakban található szócikkek mennyiségét mutatja. Leolvasható, hogy a Mikes-szótár várhatóan 20 ezer szócikket fog tartalmazni, ez elsősorban a Petőfi-szótárral való összevetésben érdekes, ahol a szócikkek száma jelentősen meghaladja a 25 ezret. Jelen összefoglalásban nincs mód további elemzésre, az egyes szótárak sajátosságaiból következő tanulságok levonására, de ha az itt látható adatokat összevetjük az egy-egy reprezentáns írói szótár elkészítéséhez feldolgozott szóelőfordulások (szóadatok) számával (2. ábra), egyértelmű, hogy a Mikes-szótár meglepően sok magas adatszámú szócikkből fog állni. (Egyelőre informális adatként a készülő József Attila-szótár – Mártonfi Attila közlése – is szerepel a sorban.)

2. ábra
A feldolgozott szövegszók száma egyes írói szótárakban

5. A szótár felhasználásának lehetőségei, további feladatok, tervek

Elkészült a Folio-adatbázis is, amely nemcsak a szövegek gépen történő olvasására alkalmas, hanem a különféle logikai kapcsolatokon alapuló keresési lehetőségeknek köszönhetően szavakra, illetve szavak együttállására is kereshetünk.

A szótár nemcsak egy-egy szó keresésére, hanem különféle típusú szövegvizsgálatokra is alkalmas már jelenlegi formájában is: így elemezhetjük a szövegvándorlásokat, vagyis azokat az eseteket, amikor különböző művekben szóról szóra azonos szövegrészletekkel találkozunk; másrészt azokat az eltéréseket is megvizsgálhatjuk, amelyek ugyanazon idegen nyelvű művek különböző mikesi fordításai között fordulnak elő. Megfigyelhetünk grammatikai sajátosságokat is, többek között a maitól eltérő egyeztetéssel, pl. *a' két Hugaimnak;* a főnévi igenév használatával, pl. *innod ne kérj;* a bővíthetőséggel, pl. *idönek jol töltése;* valamint az igeidők kifejezésével kapcsolatban is, pl. *a te ötséd meg holt volt, és fel támadot.*

Legközelebbi terveink között szerepel a szótár eddig feldolgozott anyagának internetes közzététele. A szótári weboldalon bemutatjuk a címszavakat, az alak- és írásváltozatokat, a szóalakokat és rövid szövegkörnyezettel együtt a szóelőfordulásokat is. Mindezekre külön is lehet majd keresni. Ezen kívül a teljes szövegkorpusz is elérhetővé válik. A szóelőfordulások számát is feltüntetjük, valamint azt is, ha más köznyelvi szótárban nem fordul elő a keresett szó. A még távolabbi terveink között szerepel minden szó jelentésstruktúrájának meghatározása a megfelelő szócikkekben, ahol a paradigmatikus változatok már grammatikailag meghatározott sorrendben szerepelnének. Végső célunk tehát Mikes teljes szókincsét a legmagasabb szintű értelmező-minősítő típusú írói szótár keretében hozzáférhetővé tenni a szakemberek és az érdeklődők számára. Végezetül szeretnénk kiegészíteni a szótárt az író munkásságához kapcsolódó tartalmi elemekkel, dokumentumokkal, szakmai kommentárokkal is. Meggyőződésünk, hogy Mikes Kelemen munkásságának alaposabb feltárásával nemcsak a nyelvészek, nyelvtörténészek, hanem más

tudományterületek (irodalomtörténet, történelem, művelődéstörténet stb.) művelői számára is számos eddig még nyitott kérdés válik megválaszolhatóvá.

Irodalom

Benkő L. 1979. *Az írói szótár.* Budapest: Akadémiai Kiadó.

Büky L. 2010. Beke József szerk., Radnóti-szótár. Radnóti Miklós költői nyelvének szókészlete. *Magyar Nyelv* 106. évf. 3. szám. 372–380.

J. Soltész K., Szabó D., Wacha I. 1973–1987. *Petőfi-szótár. Petőfi Sándor életművének szókészlete 1–4.* Budapest: Akadémiai Kiadó.

Kiss M. 2012a. A digitális Mikes-szótár. *Magyar Tudomány* 173. évf. 3. szám. 279–284.

Kiss, M. 2012b. The Digital Mikes-Dictionary, In: Tüskés, G. et al. (Hrsg.) *Literaturtransfer und Interkultralität im Exil Das Werk von Kelemen Mikes im Kontext der europäischen Aufklärung*. Bern: Peter Lang Verlag. 288–297.

Kiss M., Tüskés G. 2011. Mikes-szótár. Kutatási beszámoló. *Magyar Nyelvőr* 135. évf. 3. szám. 313–323.

Forrás

Mikes K. 1966--88. *Mikes Kelemen összes művei 1–6*. Sajtó alá rendezte: Hopp Lajos. Budapest: Akadémiai Kiadó.