gyűjtemények szerint is (utóbbinak például Kazinczynál van jelentősége, a *Glottomchusok* című nyelvújítási levelezésgyűjtemény formájában). A bevezetőben említett szerzőlista alapján is látható, hogy ha a tervek megvalósulnak, az páratlan mennyiségű, sőt reprezentatív levelezésanyagot jelent majd a klasszikus magyar irodalom korszakát illetően, melyben a toposzok, olvasmányok áramlását, a kirajzolódó kapcsolatok mintázatát, egy-egy korabeli jelentős esemény levélíróktól függő alternatív leírásait vagy a térkép által megjelenített vizualizált összefüggéseket ugyanúgy könnyedén analizálhatjuk.

A közeljövőben a konzorciumi együttműködés révén már nem csupán egy-egy életmű tárulhat fel előttünk úgy, ahogy eddig még soha, hanem egy teljes korszak világába nyerhetünk majd mélyreható szakmai betekintést, ha úgy tetszik, egy okostelefonnal is.

A vállalkozás technikai maradandóságát a nemzetközi szabványnak számító TEI-XML-alapú jelölőnyelvi kódolás, illetve a Debreceni Egyetem Egyetemi és Nemzeti Könyvtár szervere biztosítja. A kezelőprogram megalkotása és folyamatos fejlesztése a kutatócsoporti informatikus Nyitrai Nóra érdeme, miként az is, hogy a szövegkiadói munkát számos innovatív megoldással – a kódolási fázist például általánosan használt grafikus felületek adaptálásával – optimalizálta.

Olyan szakembergárda és olyan keretfeltételek alakultak ki az évek során, melyek okot adnak bízni a nagyívű tervek tényleges megvalósításában. Voltak idők, amikor egy-egy munkamegbeszélés során informatikusunk és projektvezetőnk, Debreczeni Attila sokszor vetettek borús tekinteteket egymásra. Azóta derűsebbek az arckifejezések, a folyton előkerülő újabb és újabb problémák ellenére is. Ezek a nagyon is emberi, hús-vér „kódok" pedig még mindig sokkal inkább szavatolnak bármiért is, mint az egyébként igencsak üdvözlendő jelölőnyelviek.

A cikk az MTA–DE Klasszikus Magyar Irodalmi Textológiai Kutatócsoport programja keretében készült (vezető Debreczeni Attila), és az OTKA (K K108831) támogatásával jött létre.

Kulcsszavak: *elektronikus, kritikai kiadás, genetikus szöveg, navigátor, belépési profil, szövegforrás, hálózat*

IRODALOM

Debreczeni Attila (2012): [Módszertani bevezető]. In: Debreczeni Attila: *Csokonai költői életművének kronológiai rendje (Csokonai Vitéz Mihály összes művei. Pótkötet)*. Akadémiai–Debreceni Egyetemi, Budapest–Debrecen, 13–34.

Debreczeni Attila (2014): Kritikai kiadás papíron és képernyőn. In: Czifra M. – Szilágyi M. (szerk.): *Textológia–filológia–értelmezés. Klasszikus magyar irodalom.* Debreceni Egyetemi, Debrecen, 26-39.

Az illető elektronikus kiadások gyűjtőhelye • http://irodalom.arts.unideb.hu/kutatas/textologia/ekiadasok.php

FŐURAK LEVELEI, BOSZORKÁNYOK PEREI ÉS EGY ÚJONNAN ÉPÜLT ADATBÁZIS: A TÖRTÉNETI MAGÁNÉLETI KORPUSZ

Dömötör Adrienne
MTA Nyelvtudományi Intézet,
tudományos főmunkatárs
domad13@gmail.com

Novák Attila
tudományos munkatárs,
MTA–PPKE Nyelvtechnológiai Kutatócsoport
novak.attila@itk.ppke.hu

Bevezetés: adat, adatgyűjtés és adatbázis

A nyelvtörténeti kutatás adatigényes tudományterület: a nyelvtörténész rá van utalva a vizsgált korszakból fellelhető adatállományra (már amennyiben nyelvemlékekkel rendelkező időszakról van szó). Az utóbbi időben a gyűjtött adatok presztízse az elméleti indíttatású nyelvészeti kutatásokban is megnövekedett; gyakran elhangzik: nem helyes ellentétesnek tekinteni az elméleti és az empirikus nézőpontot, hiszen az elméletek éppen az adatokon állnak vagy buknak meg. Az adatgyűjtés népszerűségét nagyban fokozza a lehetőség, ha anélkül is elvégezhető, hogy a munkafolyamatra hosszú hónapokat kellene rászánni. Az elektronikus korpuszok segítségével a felhasználók néhány kattintással gazdag adatgyűjteményre tehetnek szert. A számítógépes technikák kínálta lehetőségek közepette megkerülhetetlen feladat tehát annak az összetett, hosszadalmas (és nem minden fázisában hálás) munkának az elvégzése, amelynek eredményeként előáll egy-egy adatbázis.

Az új adatbázis bemutatkozik

A Történeti magánéleti korpusz (URL1) a magánéleti nyelvhasználathoz legközelebb álló szövegtípusok anyagát dolgozza fel: magánleveleket és bírósági jegyzőkönyveket tartalmaz. Az élőnyelv kiemelkedően fontos terepe a nyelvtörténeti kutatásoknak, hiszen ez az a nyelvi regiszter, amelyben a nyelvi változások megindulnak. Korábbi korokat illetően azonban nyilvánvalóan korlátozott az élőnyelvi anyagok elérhetősége; ezért beszélhetünk esetünkben is csak a magánéleti nyelvhasználathoz legközelebb álló forrásokról. Anyagaink az ilyen jellegű fennmaradt szövegek legkezdetétől a felvilágosodás szimbolikus indulásáig terjedő időszakból származnak. A 15. század végétől – kisszámú szöveggel – a kései ómagyar kort képviselik, a 16. század első harmadától a 18. század második harmadáig pedig bőséges anyaggal reprezentálják a középmagyar kort. A minél szélesebb körű szociolingvisztikai kutathatóság érdekében a források kiválasztásakor a változatosságra törekedtünk: tekintetbe vettük az időbeli, földrajzi, társadalmi és nemek szerinti megoszlást. Ezeket fel is tüntetjük; a pereken az első kettőt, a leveleken pedig a továbbiakat is: a küldő és a címzett társadalmi státuszát, nemét, kettejük viszonyát és a levél keletkezési módját (saját kezű/nem saját kezű).

A jelenleg 6,5 millió karakter terjedelmű adatbázis – cikkünk címének is megfelelően – főúri leveleket és boszorkányperek szövegét tartalmazza, az anyag azonban folyamatosan bővül, s már készen állnak a feldolgozásra más típusú levelezések is (jobbágylevelek, szépírók magánlevelei, peregrinuslevelek), illetve további perek jegyzőkönyvei (úriszéki iratok). Az adatbázis tartalmazza a feldolgozott anyagok eredetijét, egyfajta olvasatát – pontosabban: az eredeti szövegeknek a mai sztenderdhez közelített változatát – és minden szavának szófaji–morfológiai elemzését. Erről a hármasságról alább még szót ejtünk. Keresés mindhárom szinten végezhető; különösen a grammatikai annotációnak köszönhetően kínálkoznak szerteágazó lehetőségek a nyelvész felhasználó számára.

A *Történeti magánéleti korpusz* jelentőségét kiemeli, hogy ez az első teljes egészében elemzett magyar nyelvtörténeti adatbázis. A korpusz – amely a cikk végén látható linken bárki számára szabadon hozzáférhető – az MTA Nyelvtudományi Intézetének Finnugor és nyelvtörténeti osztályán készült. Alapvető céljai szerint elsősorban a történeti morfológia és szintaxis, a történeti szociolingvisztika, pragmatika, a grammatikalizáció stb. kutatóinak munkáját könnyítheti meg, de haszna a felsőfokú nyelvészeti vagy akár a középfokú nyelvtani oktatásban is megmutatkozhat. A felhasználó tájékozódását a honlapon elhelyezett segédletek próbálják megkönnyíteni (eljárásaink ismertetése; a keresőfelület használati lehetőségeinek bemutatása; a morfológiai címkék rövidítéseinek feloldása).

Amit a korpusz anyagával kapcsolatban mindezen felül fontos szem előtt tartani: szövegkiadásokból dolgoztunk (az idő és a lehetőségek kényszeréből fakadóan), vagyis mindenben ki voltunk szolgáltatva a szövegközlők eljárásainak, döntéseinek. Emiatt anyagaink bizonyos típusú vizsgálatokra nem ajánlhatók (hangjelölés–helyesírás, fonológia és határterületei). Mindig hasznos viszont magukkal a kiadásokkal is megismerkedni; az átfogó anyagismeret szakmai követelménye mellett azért is, hogy szövegközlési módszerüket, apparátusukat is figyelembe lehessen venni. (A könyvek adatai a honlapon megtalálhatók.)

Hogyan készült?

A korpuszépítés folyamatának bemutatására itt csak vázlatszerűen van lehetőség. Az előkészítő folyamatokat – amelyek eredményeképpen elkészül a források digitális szövegváltozata – ezért éppen csak érintjük (anyagkijelölés és -beszerzés, adatrögzítés: gépeléssel vagy szkenneléssel és karakterfelismerő program segítségével, ellenőrzés, korrektúra). Az adatbázis-építés lényegi munkálatai akkor kezdődhetnek, amikor rendelkezésünkre állnak a digitalizált szövegek. Ekkor a következő kézi, illetve számítógépes munkafolyamatok állnak előttünk: tagmondatokra osztás (gépi + kézi); a szövegek normalizálása (kézi); ellenőrzés, javítás (kézi); szófajtani–morfológiai elemzés (gépi); egyértelműsítés (gépi + kézi); utóellenőrzés, javítás (kézi). Ennek eredményeként áll elő a korpusz összes szavának háromféle megjelenítése: az eredeti, a „normalizált” és az elemzett alak. Ahogy már utaltunk rá, ezek mindegyikére rákereshet a felhasználó.

Az alábbiakban a digitalizálásról, az elemzésről és az egyértelműsítésről lesz szó nagyon röviden. (A munkafolyamatok részletesebb bemutatását lásd pl. Dömötör, 2013, 2014; Novák et al., 2015.)

A digitalizálás

A korpuszokat alkotó szövegek eredetileg kéziratos formában maradtak fenn, azonban – ahogy fentebb már említettük – a projektnek nem képezte részét kéziratok feldolgozása, azaz minden esetben nyomtatott szövegkiadásokból dolgoztunk. A szövegek digitalizálását többnyire számítógépes OCR-programok *(Optical Character Recognition)* alkalmazásával automatikusan végeztünk el. Egyes szövegkiadások esetében nehéz feladatot jelentett a szokatlan karakterek és mellékjelkombinációk feldolgozása. Ezek konvertálásához újra be kellett tanítani az alkalmazott OCR-programot, hiszen más-más különleges karakterek szerepeltek az egyes anyagokban. Az automatikusan felismertetett szövegben azonban így is számos hiba maradt, ezért minden szöveg eredeti és digitalizált változatát össze kellett hasonlítani, és a beviteli hibákat kézzel javítani.

A normalizálás

Mivel a rendelkezésünkre álló morfológiai elemzőprogram a mai magyar nyelvi sztenderdre lett kidolgozva, és mert általában a gépi elemezhetőség feltétele az egységes íráskép, először azt kellett megoldanunk, hogy a rendkívül nagy változatosságot mutató szövegek megfeleljenek az adott feltételeknek. Vagyis létre kellett hoznunk egy szövegváltozatot, amely mentes azoktól a jellemzőktől, amelyek egyfelől a feldolgozott források helyesírási-hangjelölési sokszínűségéből fakadnak, másfelől dialektális jegyekként, illetve a nyelvtörténeti (elsősorban fonológiai) változások következményeként adódnak. A munkafolyamat során a szövegeket tagmondatokra is bontottuk.

Ha ránézünk például a következő két mondatrészletre – az első egy boszorkányperből, a második egy Telegdy-levélből származik –, egyértelművé válik, miről is van szó: *„az Fatens kapvan edgy Lapocka Zaradnokot Beke Istvannak a hatara csapta ugyan pesget az szüri”* (normalizálva: „a fatens, kapván egy lapocska zsarátnokot, Beke Istvánnak a hátára csapta; ugyan pezsgett a szűre”); *„edig niluan uagjon Knel, hogj tülünk egj mily földön Törökök gjülekeßtek óßue”* (normalizálva: „eddig nyilván vagyon kegyelmednél, hogy tőlünk egy mérföldön törökök gyülekeztek össze”). A korpuszépítés folyamatának ez az a szakasza, amely egyrészt alapos nyelvtörténeti felkészültséget igényel (emellett szoros barátságot a nyelvtörténeti, etimológiai, táj- és egyéb szótárakkal, valamint a történeti nyelvtanokkal), másrészt nagy kitartást követel meg, hiszen – a fentebb mondottaknak megfelelően – az összes szöveg összes tagmondatát „le kell fordítanunk” mai magyarra, mégpedig a lehető leghűségesebb módon. A *Történeti magánéleti korpusz* az első adatbázis, amely magyar nyelvi anyagon ezt a módszert viszi végig. (Hasonló eljárásmódokra nemzetközi szinten lásd Novák et al., 2015, 7–8.) A normalizálást és az ezt követő többszörös javításokat többen végezzük, ezért nemcsak állandóan bővített, pontosított szabályzatra, hanem rendszeres egyeztetésekre is szükség van.

A normalizálás legfőbb elve a „morfémamegmaradás” törvénye: az, hogy a szavakat felépítő, jelentést hordozó egységek, azaz morfémák a normalizálás folyamán ne tűnjenek el, vagy alakuljanak át más morfémákká. A morfémahűség helyes megvalósításához általában alaposan mérlegelnünk kellett az adott korszak nyelvi sajátosságainak és helyesírásának jellegzetességeit. Törekedtünk arra is, hogy a korabeli helyesírás bizonytalanságaiból adódó inherens és ténylegesen feloldhatatlan többértelműségeket lehetőleg ne tüntessük el a normalizálás során. A tőmorfémákat illetően a *zuhaj*-tól, *tyúkmonysüttő*-től a *szerencsít*-en, *frajcimmer*-en át *a restáns*-ig

és a *skrupulizál*-ig eddig több mint négyezer olyan lexémával találkoztunk, amelyek a mai magyarban nem használatosak; zömükben elavult (képzésformájú) és/vagy nyelvjárási, rétegnyelvi, idegen nyelvből átemelt szavak. De a toldalékmorfémák szintjén is számos, a mai sztenderdben nem élő elemet kell megtartanunk annak érdekében, hogy az elemzőprogram fel tudja őket dolgozni, illetve a felhasználó majd keresni is tudja őket. Ennek értelmében – hogy csak egy nagyon kézenfekvő esetet említsünk – a *Vác**cá** men**e*** nem alakítható át *Vácra ment* alakúra, hiszen a régi és a mai szóelemek (bár funkcióikban hasonlók) különböző nyelvi egységeket tartalmaznak. Az ilyen és ehhez hasonló esetekben az elemzőprogramot kell betanítani, hogy kezelni tudja a maitól eltérő formákat.

Az elemzés

A számítógépes nyelvészet egyik alapvető feladata a szóalakok automatikus alaktani elemzése. Ennek során az elemzőprogram a szavakhoz meghatározza azok szótövét, annak szerkezetét, szófaját és a szóban szereplő toldalékokat. A lentebbi példákban látható morfológiai címkék ezeket az információkat kódolják. A [N.Pl.Acc] címke jelentése például, hogy az adott szó főnév (N), többes számú (Pl) és tárgyesetben van (Acc). Az ilyen elemzőprogramok számára szükség van egy tőtárra és egy toldaléktárra, amelyekben az adott nyelvben előforduló lehetséges szótövek, illetve toldalékok vannak eltárolva. Ezen kívül a programban megvalósított alaktani leírás tartalmazza a szavak toldalékolását meghatározó paradigmákat, illetve a szavak felépítését leíró szabályrendszert.

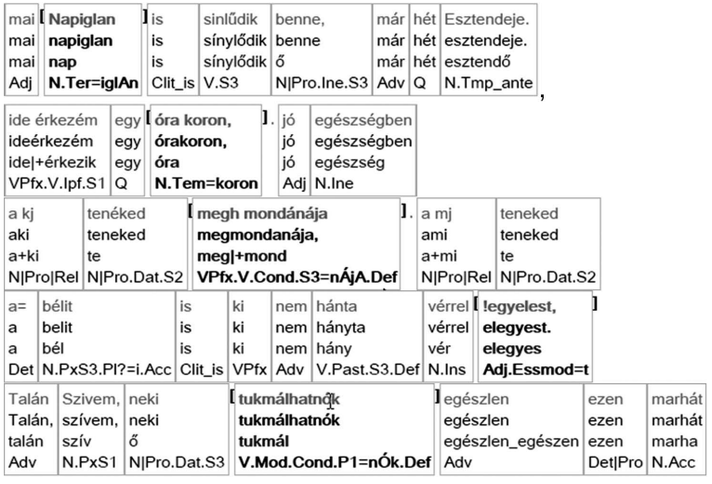
A digitalizált és normalizált szövegek automatikus elemzésére a Humor magyar morfológiai elemző (Novák, 2003) egy erre a célra kibővített változatát alkalmaztuk. Ehhez ki kellett bővíteni a program tőtárát és toldaléktárát az időközben kihalt paradigmákkal, szótövekkel és toldalékokkal, illetve a toldalékok alakváltozataival. Az alábbiakban az utóbbiakra láthatunk példákat (félkövérrel kiemelve). A példákban az egyes szavakat négy jellemzőjük írja le: az eredeti alak, a normalizált alak, a szótő és a morfológiai címke. Az utóbbi kettő együtt adja a morfológiai elemzést (illusztrációink egy részében egymás alatt, egy részében egymás mellett láthatók) *(1. ábra)*.

A elemző toldaléktárába ötven új toldalékot vettünk fel (ezek alakváltozatait, allomorfjait nem számolva). Az alábbiakban olyan toldalékmorfémákra láthatunk példákat az igei alaktan köréből (félkövérrel kiemelve), amelyek a mai magyarban már nem használatosak *(2. ábra)*.

A toldalékok és paradigmák leírásánál nagyságrendileg több munkát jelentett azoknak a töveknek a felvétele, amelyek a mai magyar elemző lexikonából hiányoztak. Sok esetben a tő ugyan megvolt, de a régi szövegekben más szófajú (is) lehetett, mint ma, illetve bizonyos konstrukciókban másképp kell elemezni őket, mint a mai megfelelőjüket. Ilyen például a régi névutós szerkezetek egy része, amelyben a névutó a *-nak /-nek* ragos birtokos szerkezethez hasonló formában egyeztetve van a főnévvel. Ebben a ragos névutó elemzése más, mint az azonos alakú, a mai magyarban is létező névmást tartalmazó *(ő)miatta* alaké *(3. ábra)*.

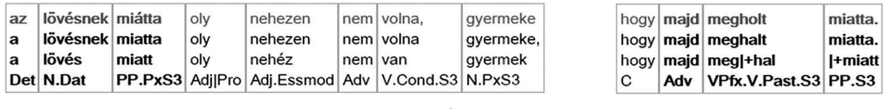
Az egyértelműsítés

A szövegek elemzését egyértelműsíteni is kellett, mivel maga az elemző az adott szóalak minden lehetséges elemzését megadja. Ezek általában mind helytálló elemzések a szót



1. ábra

2. ábra



3. ábra

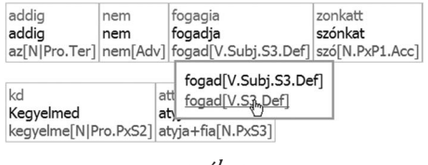
önmagában vizsgálva, a szövegkörnyezet alapján viszont egyértelműen ki lehet választani, hogy az adott kontextusban melyik elemzés a helyes. Ráadásul a történeti szövegekben a többértelműségek aránya nagyobb, mint a mai szövegek sztenderd elemzővel való elemzése esetében. Ez egyrészt amiatt van, mert az elemző lazább, megengedőbb (ez a mai sztenderdben elő nem forduló szerkezetek elemzéséhez szükséges), amely a korpusz ritkább szerkezeteit olyan helyeken is felismerni véli, ahol nem azok szerepelnek, másrészt pedig az eldönthetetlen többértelműségek ilyenként való címkézéséből fakad.

A morfológiai annotáció egyértelműsítésében a munka oroszlánrészét géppel végeztük. Sztenderd szövegek esetén ugyanis erre a feladatra is létezik számítógépes megoldás,

így csupán adaptálni kellett egy ilyen meglévő programot erre a nyelvváltozatra. Az ó- és középmagyar morfológiai elemző elemzéseit felhasználva a PurePos nevű statisztikai egyértelműsítő eszközt (Orosz – Novák, 2013) használtuk erre a célra. A program már elemzett és egyértelműsített szövegekből megtanulja, hogy milyen szövegkörnyezetben melyik elemzés a legvalószínűbb a morfológiai elemző által előállított lehetséges elemzések közül. Természetesen minél több tanítóanyagból tanul a rendszer, annál jobban működik, ezért a programot inkrementális módon egyre több egyértelműsített és ellenőrzött szöveggel újratanítottuk.

Az így egyértelműsített szövegek kézi ellenőrzéséhez (illetve az első szövegek még teljesen manuális egyértelműsítéséhez) olyan webes felületet hoztunk létre, amelyen a téves egyértelműsítések, illetve normalizálási hibák nagyon hatékonyan javíthatók. Az automatikusan kapott elemzés helyett úgy lehet másikat választani, hogy az egérmutatót a szó fölé húzzuk, és a megjelenő listából másikat jelölünk ki. Kézzel javítható az eredeti és a normalizált szóalak (és akár az elemzés is). Az eredeti vagy a normalizált szóalak javítása után a szó a programmal azonnal újraelemeztethető. Az automatikusan megjelenő lista olykor csak két-három elemből áll (ilyenkor gyorsan lehet haladni a kézi egyértelműsítéssel) *(4. ábra)*.

Egyes igei alakok esetében viszont – főként a sok elvileg lehetséges igenévi, szenvedő és műveltető szerkezet miatt – meglehetősen



4. ábra

hosszas lista áll előttünk (ilyenkor tovább tart, amíg sikerül kiválasztani a szövegkörnyezetnek megfelelő alakot; a lentebbi szövegdobozból például a 15. sort, azaz a tárgyas ragozású, múlt idejű, egyes szám második személyű formát) *(5. ábra)*.

A keresőfelület; keresési lehetőségek

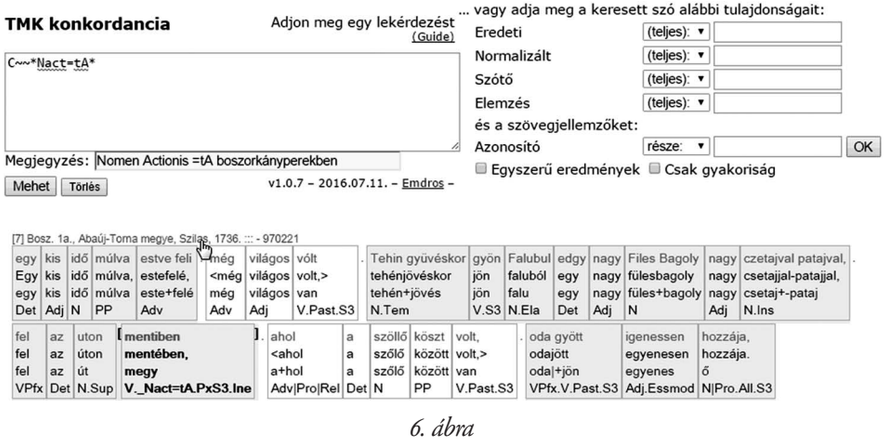
Végül a normalizált és elemzett szövegek fölé egy keresőfelületet hoztunk létre. A szövegekben való keresést támogató korpuszkezelő (Petersen, 2004) nemcsak azt teszi lehetővé, hogy a felhasználó különböző nyelvtani szerkezetekre keressen a szövegekben példákat (amit a hozzáadott morfológiai elemzés tesz lehetővé), hanem azt is, hogy a munkatársak a kereső találataiban is azonnal kijavíthassák az annotációban vagy a szövegben fellelt hibákat. A hibakeresés és -javítás egyik hatékony módja, amikor a korpuszban kifejezetten olyan szerkezeteket keresünk, amelyek valószínűleg hibásak, és a valóban hibás találatokat azonnal javítjuk. A javított korpuszt ezután exportálni lehet, és az automatikus elemzőprogramot a javított korpusszal újratanítani. Bár a korpusz nem tartalmaz kifejezett mondattani elemzést, a morfológiai annotáció alapján a mondattani szerkezetek nagy része is hatékonyan megkereshető. A keresőben kellő szaktudással jól megfogalmazhatók olyan lekérdezések, amelyek segítségével az ó- és középmagyar időszak mondattana iránt érdeklődő kutatók is eredményesen használhatják a korpuszt. Ehhez bonyolultabb lekérdezéseket kell összeállítani, amelyek megfogalmazásához érdemes megtanulni a keresőhöz kifejlesztett speciális lekérdezőnyelvet.

A kereső lehetővé teszi, hogy mondaton, tagmondaton vagy adott metaadatokkal megjelölt tulajdonságú szövegen belül keressünk, illetve akár többmondatos egységek is

5. ábra

lekérdezhetők. A kereső által megjelenített találati egység a normalizált szövegekben mondatként kijelölt szövegszakasz. A tagmondatok lehetnek nem folytonosak (ez az alárendelő szerkezetek esetén gyakran előfordul, de olykor a főmondat vagy egy mellérendelő szerkezet valamelyik eleme ékelődik be). Az alábbi példa olyan találati mondatot mutat be, amelyben több megszakított tagmondat is szerepel *(1. ábra)*.

Az egyes találatok fejlécére kattintva új böngészőablakban megjeleníttethető az adott mondatot tartalmazó teljes dokumentum, amelyen belül a keresőkifejezés által illesztett



6. ábra

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,