Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből

Simon Eszter, Sass Bálint
MTA Nyelvtudományi Intézet, Budapest
{simon.eszter,sass.balint}@nytud.mta.hu

Kivonat

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 kutatásoknak. Cikkünkben egy ómagyar nyelvtörténeti adatbázis létrehozásáról számolunk be, bemutatjuk a teljes korpuszépítési munkafolyamatot a szkenneléstől a korpuszlekérdező eszközig.

Kulcsszavak: kulturális örökség, nyelvtechnológia, történeti korpusz, szövegnormalizálás, korpuszépítés

1. Bevezetés: nyelvtechnológia és kulturális örökség

A társadalom- és bölcsészettudományok területén tényked ő kutatók korábban elsősorban papíralapú forrásokból: kéziratokból, könyvekből dolgoztak. Az elmúlt évtizedek során azonban az információhoz való hozzáférés módja a számítógépek és az internet használatának elterjedésével merőben megváltozott. Ma már a könyvtárban sem kell katalóguscédulákat átbogarászni, ha meg akarjuk tudni egy könyv elérhetőségét, hanem viszonylag könnyen és egyszerűen tudunk az interneten keresztül keresni a könyvtári adatbázisokban a könyvekhez tartozó metaadatok (szerző, kiadó, kiadás ideje és helye stb.) alapján. A humán tudományok és az információs technológiák találkozásával egyre több adat válik digitálisan is elérhetővé, akár a nagy mértékű digitalizációs törekvéseknek köszönhetően, akár amiatt, hogy az adat eleve digitális formában jön létre.

A nyelvi kulturális örökség elérhetővé és feldolgozhatóvá tételében kulcsfontosságú szerep jut a nyelvtechnológiának. Az egyszerű digitalizálás, ami általában kimerül a primér adat képként való beszkennelésében, nem nyújt elég széleskörű és szofisztikált keresési lehetőséget. Az olyan szöveges adatbázisok, melyekben az elemek különféle nyelvészeti (és/vagy történeti, paleográfiai stb.) információval vannak ellátva, sokkal kifinomultabb kutatási alternatívákat kínálnak.

A humán tudományok és a nyelvtechnológia ötvözése mindkét tudományterületnek nagy hasznot hozhat. A kutatók az egyik oldalon időt nyernek a hatékonyabb adateléréssel. A számítógépes feldolgozás támogatja a következetességet, az egységességet és a metaadatok könnyebb kezelését. A digitalizált adat nem helyhez kötött, vagyis a kutatók bárhonnan hozzáférhetnek – akár egy időben párhuzamosan is.

Ami a nyelvtechnológiai oldalát illeti: a nyelvtechnológusok az elmúlt évtizedekben jellemzően relatíve kicsi, szűk domainre specializált és szűrt adathalmazokkal dolgoztak. A nyelvi kulturális örökség területén viszont elsősorban a sztenderdtől eltérő, illetve archaikus nyelvváltozatokkal találkozunk, amelyek számos kihívást állítanak a nyelvtechnológusok elé. A korpuszépítési munkálatok során elsősorban már digitalizált szövegekből indulnak ki – de nem ez a helyzet a történeti dokumentumokkal. Az elektronikus formátumok (sőt az elektromosság) előtti korból származó szövegekből való korpuszépítés sokkal idő- és munkaigényesebb folyamat, és bizonyos esetekben más módszereket is igényel, mint a mai szövegek esetében. Már az alapszintű szövegfeldolgozó lépések (szavakra és mondatokra bontás, morfológiai elemzés és egyértelműsítés) során az eddigieknél robusztusabb vagy teljesen új módszerekre van szükség. Az ezen a területen kifejlesztett eszközök a nyelvtechnológia más területein is valószínűleg sikerrel alkalmazhatóak. Vagyis a kulturális örökség digitalizálása során nem csak a már bevált módszerek új területeken való alkalmazása történik, hanem az új módszerek új kutatási kérdéseket is felvetnek. Ezek megoldásához a különböző tudományterületek képviselői közötti szoros együttműködésre van szükség.

A nyelvtörténészek és nyelvtechnológusok egyik legfontosabb együttműködési terepe a történeti korpuszok építése. A kilencvenes, de legfőképp a kétezres években sorra indultak olyan projektek, melyek egy adott nyelv valamely régebbi változatának digitalizálását és feldolgozását célozzák (Kroch and Taylor, 2000; de Sousa and Trippel, 2006; Kunstmann and Stein, 2007; Thomas et al., 2007). Ezek a korpuszok természetesen sok paraméterükben különböznek: teljes szövegeket vagy csak részleteket tartalmaznak; egy korszak teljes lefedésére törekszenek, vagy egy nagyobb kor szövegeiből kívánnak reprezentatív válogatást adni; morfológiai és szintaktikai annotációt is tartalmaznak, vagy a puszta szöveget adják szövegegységekre tagolva stb. Annyiban azonban megegyeznek, hogy valamilyen szintű nyelvi információt mindenképpen tartalmaznak, és szofisztikált kereséseket tesznek lehetővé, hogy minél inkább megkönnyítsék a nyelvészeti, irodalmi vagy történelmi célú kutatásokat.

Cikkünkben egy, a fenti trendbe illeszkedő projektet mutatunk be, melynek célja, hogy diakrón szintaktikai vizsgálatokat végezzen magyar nyelvű szövegeken, melyhez elsődleges fontosságú egy elektronikus nyelvtörténeti adatbázis létrehozása. A *Magyar Generatív Történeti Szintaxis* című projekt keretein belül felépítünk egy olyan korpuszt, amely tartalmazza az összes fennmaradt ómagyar kori (896–1526) szövegemléket, és amely nyelvészeti információkat tartalmaz elektronikusan előhívható és interpretálható módon.

A cikkben a teljes korpuszépítési munkafolyamatot bemutatjuk. A 2. fejezetben a korpusz anyagának összegyűjtését írjuk le, majd a 3. fejezetben a feldolgozási lépéseket a szkenneléstől a betűhű szöveg előállításáig. A 4. és az 5. fejezetek a kézi és a gépi norma-

lizálást mutatják be. A 6. fejezet a morfológiai elemzés és egyértelműsítés feladatkörét tárgyalja. A 7. fejezetben azt vizsgáljuk, hogy hol kaphatnak helyet az automatikus, félautomatikus és manuális nyelvfeldolgozó eljárások a korpuszépítési munkálatokban. A 8. fejezet a korpusz felépítését, a 9. fejezet pedig a hozzá készült lekérdező eszközt mutatja be. Ugyanitt néhány példán keresztül azt illusztráljuk, hogy a korpusz segítségével milyen típusú nyelvészeti kérdéseket tudunk megválaszolni. Végül az összegzés előtt a korpuszépítéssel kapcsolatos további feladatainkról esik szó.

2. A korpusz anyagának összegyűjtése

A reprezentativitás, de legalábbis a kiegyensúlyozott szövegválogatás a korpuszépítés fontos elve. Ez azonban háttérbe szorul, ha eleve korlátozott az elérhető nyelvi anyag mennyisége (például ha egy holt nyelv vagy egy nagyon speciális nyelvi réteg adja a korpusz anyagát). Ez a helyzet az ómagyar korpusz esetében is, amely – célkitűzésének megfelelően – az összes ómagyar korból fennmaradt szövegemléket tartalmazza. Szövegemlék alatt az összefüggő mondatokat tartalmazó nyelvemlékeket értjük; az ún. szórványemlékekkel, amelyekben csak sporadikusan fordulnak elő magyar szavak vagy nevek, jelen projektben nem foglalkozunk. Nem szerepelnek továbbá a korpuszban azok a szövegek sem, amelyeket még soha nem adtak ki nyomtatásban, vagyis a nyelvtörténeti átírási munkát nekünk kellene elvégezni.

A fenti megszorításokat figyelembe véve a feldolgozandó ómagyar anyag 48 kódexet, 27 rövidebb szövegemléket és 244 misszilist (elküldött levelet) foglal magában, vagyis mindösszesen körülbelül 2 millió szövegszót.

A korpuszépítés első lépése a valamilyen elektronikus szöveges formátumban már meglévő nyelvtörténeti anyagok összegyűjtése. A különböző forrásokból (kiadóktól, nyelvtörténészektől) származó, változatos fontkészleteket használó dokumentumokat egységes, UTF-8 kódolású, sztenderd Unicode-karaktereket tartalmazó sima szövegfájlokká alakítjuk (ld. 3.3. fejezet).

Másik forrásunk a *Számítógépes Nyelvtörténeti Adattár*, amelyben több ómagyar kódex ábécérendes adattára elérhető (Jakab and Kiss, 1994, 1997, 2001; Jakab, 2002). A kódexfeldolgozási munkálatok még a hetvenes években kezdődtek a Debreceni Egyetemen Jakab László vezetésével. Az adattárban a kódex címszavai (a szövegszavak tövei mai magyar átírásban) ábécérendbe rendezve szerepelnek. A hozzájuk tartozó betűhű szövegszavakat a lelőhely (lapszám, sorszám) megjelölésével közlik, mellettük számokkal rögzítették az adatra vonatkozó helyesírás-történeti, szótörténeti, hangtani, szófajtani, jelentéstani és alaktani tudnivalókat. A szövegben sokszor előforduló szavakat egy függelékben különítették el, melyeket a lelőhely alapján visszahelyezünk az eredeti kódexbeli helyükre. Az egyes szövegszavak soron belüli sorrendjét nem közlik, ezért a sorbarendezést is elvégezzük. Ezután a többféle fontkészletet alkalmazó táblázatot UTF-8 kódolású sima szöveggé alakítjuk, majd ebből állítjuk vissza a kódexek eredeti betűhű szövegét. Az egyes szövegszavakhoz tartozó morfológiai elemzést az általunk használt morfológiai elemző kimeneti formátumára alakítjuk, továbbá a mai magyar tövek és az elemzés alapján rekonstruáljuk a normalizált szóalakot (ld. 4. fejezet). Ennek a kon-

vertálási munkafolyamatnak a végén megkapjuk az adott kódex szavainak betűhű és normalizált alakját, valamint a hozzájuk tartozó egyértelmű morfológiai elemzést (a feldolgozási szintekről részletesen ld. a 8. fejezetet).

Az ómagyar szövegek nagy részének azonban nincsen elektronikusan elérhető szöveges változata, így ezeket a számítógép által olvasható és feldolgozható formára kell hoznunk. Ez a rövidebb szövegek esetében általában begépeléssel, a hosszabbak esetében szkenneléssel, optikai karakterfelismerő (OCR) program alkalmazásával és kézi ellenőrzéssel történik.

3. A korpusz anyagának feldolgozása

3.1. Szkennelés

Néhány kódex beszkennelt verziója megtalálható a Magyar Elektronikus Könyvtárban, sőt ezek egy része ún. ,,szendvics" PDF, vagyis a kép mögött megtalálható az OCR-ezett szöveg is. Ennek ellenére ezeket nem tudtuk használni: a mögöttes szöveg nem esett át kézi ellenőrzésen, vagyis meglehetősen sok benne a hiba, a képek felbontása pedig nem elég jó az OCR-ezéshez.

Így minden kódexet, amelyet nem tudtunk szöveges formában megszerezni, minimum 300 dpi felbontásban beszkenneltünk.

3.2. Optikai karakterfelismerés

Az ómagyar kódexekben található nagyszámú különleges karakter kezelése miatt az OCR programmal szemben alapvető elvárásunk volt a *taníthatóság*. Ez utóbbi azt jelenti, hogy a program nem zárt karakterkészlettel dolgozik, hanem meg lehet neki adni bármilyen új karaktert. A szóba jöhető nyílt forráskódú szoftverek közül a *Tesseract*ot próbáltuk ki, amelynek az a hátránya, hogy az összes felismerendő dokumentum alapján egy egész karakterkészletet (nyelvet) kell megtanítani neki. Ezért végül az *Abbyy FineReader 9.0 Professional edition* mellett döntöttünk. Ez ugyan nem nyílt forráskódú, de karakterről karakterre, interaktív módon tanítható, és elég jó minőségű kimenetet ad.

Az OCR program teljesítményét szópontossággal (*word accuracy, WAcc*) mértük, amely egy dokumentumban a helyesen felismert szavak és az összes szó számának az aránya. Az előzetes elvárásoknak megfelelően az eredmények azt mutatják, hogy a pontosság nagyban függ a kódexekben alkalmazott helyesírástól. Kniezsa (1952) az ómagyar kori kódexek kezeinek helyesírását három nagy típusba sorolja; a kiértékelésnél ezt a kategorizálást követtük. 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'*. A kiértékeléshez három kódexet választottunk a három különböző típusból, továbbá összehasonlítási alapként egy rövidebb mai magyar szövegen is kiértékeltük a szoftver teljesítményét.

Az 1. táblázatból kiolvasható, hogy legjobban a mellékjel nélküli helyesírással boldogult a program: ez nagyjából megegyezik a mai magyar szövegek felismerésében nyújtott pontossággal. A mellékjeles és keverék helyesírású kódexekben használt speciális karakterek nagy száma a tanítás ellenére is közel 30%-kal rontotta a pontosságot. A mellékjel nélküli kódexek esetében a latin ábécé betűit kell felismerni, ezért itt az OCR program jó teljesítményt nyújt. A bonyolult, akár többszörös, illetve egymáshoz hasonló ékezetek elkülönítése viszont problémát okoz. A jelentős teljesítménycsökkenés hátterében tehát ezeknek a diakritikus jeleknek a nem kielégítő kezelése állhat, ahogy erről például Volk et al. (2010) is beszámol.

kódex	helyesírás	tokenszám	felismert	WAcc (%)
Kulcsár	mellékjel nélküli	36.321	35.258	97,07
Müncheni	mellékjeles	74.657	50.790	68,03
Czech	keverék	11.478	7.910	68,91
–	mai magyar	5.121	5.068	98,97

1. táblázat. Az OCR szópontossága helyesírási típusok szerint.

3.3. A betűhű szöveg előállítása

A betűhű szöveg elkészítésekor nem a kódexek kézzel írott változatát, hanem az általunk használt átirat szerkesztőjének konvencióit követjük, vagyis nem feltétlenül törekszünk tökéletes paleográfiai pontosságra. Például a Jókai-kódex esetében a Jakab-féle adattárból (Jakab, 2002) indultunk ki, amely nem jelöli külön a korban gyakran használt, ám a nyelvtörténészek nagy része szerint jelentésmegkülönböztető szereppel nem rendelkező hosszú *s*-t. Így ebben a kódexben mi sem jelöljük ezt a karaktert, annak ellenére, hogy a kódexek jelentős hányadában jelölve van. Ahol egyedi indokkal mégis eltérünk a szerkesztő közlésétől, azt mindig külön jelezzük.

A szabványosság előnyei miatt a teljes korpuszt UTF-8 kódolású sztenderd *Unicode-karakterek*kel tároljuk, és jelenítjük meg. A nemzetközi Unicode szabvány (http://unicode.org) éppen azért jött létre, hogy a világ összes nyelvének összes karakterét egy kódolási rendszerbe foglalja, lehetővé téve minden ma használatos karakter egységes megjelenítését. Mivel minden platformon elérhető, széles körben elterjedt és elfogadott szabvány, érdemes volt az ómagyar karakterek tárolására és reprezentálására is az UTF-8 kódolású Unicode-ot választani. A Unicode nagy előnye, hogy az alapkaraktereket és a diakritikus jeleket külön egységekként (külön kóddal) tárolja, és lehetőséget nyújt ezek szabad összeépítésére. Így nemcsak az *a*-ból és a vesszőből (´) gyárthatunk *á*-t, hanem például az *y*-ból és az umlautból (¨) is előállíthatjuk az ómagyar kódexekben nagyon gyakori *ÿ* karaktert. A hozzáadott ékezetek halmozhatók is, így ezen a módon a kódexek különleges karaktereinek jelentős részét szabványos kódolással tudjuk reprezentálni.

Mindenképpen szükséges egy az egész korpuszra kiterjedő szigorúan *egységes* formátum, ez teszi lehetővé, hogy a lekérdezéseket az egész anyagra vonatkoztathassuk. A

korpuszok egyik haszna, hogy nem csak példákat szolgáltatnak bizonyos jelenségekre, hanem adott lekérdezésre az *összes* találatot megadják, ezáltal lehetővé teszik a jelenségek statisztikai vizsgálatát is. A korpusz ezen fontos tulajdonságát csak úgy biztosíthatjuk, ha következetesen betartjuk azt az alapelvet, hogy azonos dolgokat mindig ugyanúgy, különbözőeket pedig mindig eltérően jelölünk. Ugyanakkor viszonylag nagy erőfeszítést kíván ennek az egységességnek a megvalósítása, mert előfordulnak olyan régi magyar karakterek is, melyek a sztenderd kódtáblában nincsenek reprezentálva. Ezeket a karaktereket egy kiválasztott Unicode-karakterrel helyettesítjük, mégpedig úgy, hogy az adott helyettesítő karaktert kizárólag az adott hiányzó eredeti karakter helyett használjuk a korpuszban. Jó példa erre az ún. *huszita cs*, amely megjelenésében leginkább egy kiskapitális L-hez hasonlítható, és amelyet Volf (1874)-et követve rendre *č*-vel helyettesítünk.

Éppen a Unicode-táblában nem szereplő különleges karakterek teszik szükségessé, hogy a háttérben egy másik fajta kódolást is alkalmazzunk. Az ún. *Prószéky-kód*ban a különböző diakritikus jelekkel ellátott és speciális történeti karaktereket betűk és számok kombinációjával jelöljük: például az *á*-t a1, az *ö*-t o2, az *ű*-t u3 jelöli. A Magyar Történeti Korpusz számítógépes adatbázisának előállításakor használt kódtáblából (Kiss and Pajzs, 2001) indultunk ki, amelyet az ómagyar kori speciális karakterek nagy száma miatt folyamatosan bővítünk. Minden szöveget a Unicode-változat mellett Prószéky-kódokkal is rögzítünk, amivel a Unicode hiányosságai ellenére is rögzíteni tudunk minden információt. A betű-szám kombinációk alkalmazása a szövegbevitel és -javítás során is hasznos, mivel így a begépelők és a nyers OCR-kimenet javítását végzők operációs rendszertől és szövegszerkesztőtől függetlenül, egyszerűen be tudják vinni a speciális karaktereket is.

A betűhű szövegváltozat előállításakor a korabeli írásjeleket, elválasztásokat (illetve azok hiányát), egybe- és különírást, a mondat- és tulajdonnévkezdő kis- és nagybetűket megtartjuk úgy, ahogy a kódexkiadásban szerepelnek. Az eredeti kódexbeli színezéseket, betűvastagításokat és kiemeléseket nem őrizzük meg, és a nyomtatott kiadás során belekerült sor- és oldaltörést jelölő virgulákat is elhagyjuk.

4. Normalizálás

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. 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. HB: *k̲ine c̲* [k̲ine k̲]), vagy többes hangértéke van egy-egy betűnek (pl. HB: *gimil c̲ íc̲tul* [gyümölcsök̲től]). 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).

E probléma megoldása céljából szükség van egy ún. *normalizálás*i lépésre, amelynek során az eredeti betűhű szóalakokat mai magyar helyesírású szavakra alakítjuk át. A többféle, különböző nyelvtörténeti szakmai érvekkel alátámasztható lehetséges feldolgozási forgatókönyvek egyik gyakori közös átalakító lépése ez a fajta normalizálás (pl. McEnery and Hardie (2003)). A szövegfeldolgozásnak ez a lépése kritikus fontosságú, enélkül ugyanis a (fél)automatikus annotáció hatékonysága a következő lépésekben drámaian visszaesik (Rayson et al., 2007).

A normalizálás során két alapelvet tartunk szem előtt. Első elvünk, hogy az összes ma nem létező szót, toldalékot, morfológiai konstrukciót megtartjuk, vagyis morfémát nem toldunk be, és nem hagyunk el. A 2. táblázat utolsó sora kiváló példa erre a jelenségre: a *-va/-ve* végű határozói igenév személyragozható volt, sőt a teljes paradigmája megvolt ebben a korban (Jászó, 1992). Ha a normalizálás során ezt az alakot a ma használatos *-va/-ve* végű alakra írnánk át, nyilvánvalóan elvesztenénk a morfológiai információt.

betűhű	**normalizált**	**értelmezés**
villamik	villamik	villámlik/villanik
isa	isa	bizony
iesek	jeszek	jövök
ymaduam	imádvám	imádva E/1.

2. táblázat. A normalizálás első alapelve.

A normalizálás második alapelve, hogy elhagyunk minden fonológiai és helyesírási esetlegességet, vagyis egységes, amennyire lehet, a mainak megfelelő helyesírásra törekszünk. Ez utóbbi azt is jelenti, hogy egy adott szót mindig ugyanúgy írunk le – ez is az egységesség elvének egy megnyilvánulása (vö. 3.3. fejezet).

betűhű	**normalizált**
mēden	minden
menden	minden
minden	minden
algyu	ágyú
agyu	ágyú
strumlast	ostromlást

3. táblázat. A normalizálás második alapelve.

A normalizálási lépés során történik meg a szöveg tokenekre és mondatokra való bontása is – mindkettő manuális munkával. Az ómagyar szövegekben a szavak egybeírása és elválasztása nem a mai szabályokat követi. Ezért a tokenizálás, vagyis a szöveg

szavakra szegmentálása során az ómagyar szövegben a szavakat a mai helyesírásnak megfelelően összevonjuk, illetve szétválasztjuk, természetesen jelölve a változtatásokat.

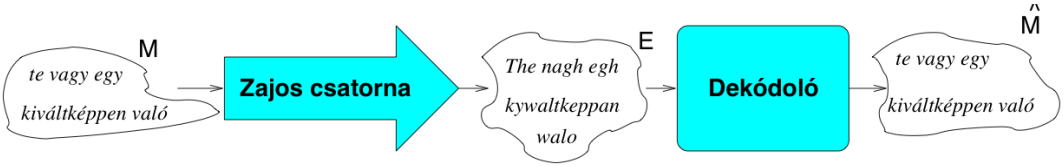
A ma használatos logikai-grammatikai írásjelezés kibontakozása csak a 17. században kezdődik, vagyis a korabeli központozásra nem támaszkodhatunk a mondatra bontásnál. Ezért a mai értelemben vett automatikus mondatra bontás lehetetlen vállalkozásnak tűnik, így ezt a szövegfeldolgozási lépést is manuálisan végezzük el. Természetesen a kézi mondatra bontás sem mindig egyértelmű – kétséges esetben inkább nem teszünk mondathatárt, vagyis azt az elvet követjük, hogy a mondat legyen inkább hosszabb, mint rövidebb. Alapesetben az alárendelő tagmondatot nem választjuk el a főmondattól, míg a mellérendelő tagmondatot igen. A feladat végrehajtása során a mai központozási alapelvekhez igazodunk.

Mivel a korabeli szövegek jó része vallási tárgyú, nagyon sok bibliai nevet találunk bennük. Az egységesség jegyében a különböző bibliafordításokban és bibliai históriákban említett tulajdonneveket is normalizáljuk, vagyis az adott nevek különbözőképpen használt alakjait egységesítjük. Ehhez a Szent István Társulat bibliafordítását használjuk: minden tulajdonnevet abban az alakban normalizálunk, ahogy ebben a kiadásban szerepel. Természetesen ez sem mentes a következetlenségektől: bizonyos neveket ebben a kiadásban sem közölnek mindig egységesen. Ilyen esetekben a kétféle névhasználat közül a gyakoribbat választjuk.

5. Gépi normalizálás

Mivel a normalizálás rendkívül időigényes manuális munka, megpróbáltuk kiváltani automatikus eljárással. A folyamat számítógépes modellezésének célja az volt, hogy választ kapjunk arra a nagyon fontos gyakorlati kérdésre, hogy a szükséges emberi erőforrás alkalmazása leszűkíthető-e a teljes anyagnál nagyságrendekkel kisebb méretű kézzel normalizált részkorpusz előállításának feladatára, mely az automatikus módszerhez *tanítókorpuszként* szükséges. Mivel ez a szövegnormalizáló konverzió analóg több klasszikus nyelvfeldolgozási probléma során jelentkező feladattal, így érdemesnek tűnt az azokban sikerrel alkalmazott módszerek adaptálása és eredményességének vizsgálata.

Fő kérdésünk az volt, hogy az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe, és melyek azok a jegyek, amelyek felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását eredményezi. Ennek érdekében szükség volt az adott modellben használt jegyeket tartalmazó specifikusan annotált tanító szövegekre, melyekből korlátozott mennyiség áll rendelkezésünkre – éppen a normalizálás szakértelmet kívánó, időigényes volta miatt. A fentebb leírt szövegbeli egyenetlenségek miatt nehéz egyértelmű konverziós szabályokat meghatározni, valamint emiatt kritikus kérdés az is, hogy a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvemlékekre. Mindezek miatt célszerű a problémát valamilyen valószínűségi alapú paradigma keretei között vizsgálni. Az átírás (transzliteráció) nyelvtechnológiai szempontú kutatásának igen gazdag eszköztára van, a különféle módszerek közül mi Shannon zajos csatorna modelljét (Shannon, 1948) választottuk. (A feladat lehetséges megközelítéseiről bővebben ld. Oravecz et al. (2009, 2010).)



1. ábra. Szövegnormalizálás zajos csatorna modellben.

Az 1. ábrán látható modellben az eredeti szöveget úgy tekintjük, mint a normalizált változat egy zajos kommunikációs csatornán átment „eltorzított" változatát. M jelöli a normalizált szövegváltozat egy részét (a példában egy részmondatot), E pedig ennek betűhű átiratát. A dekódoló feladata annak az M karaktersorozatnak a megtalálása, amelyre a $P(M|E)$ feltételes valószínűség maximális, vagyis a Bayes-tételbe behelyettesítve:

$$\hat{M} = \underset{M}{\operatorname{argmax}}\, P(M|E) = \underset{M}{\operatorname{argmax}}\, P(E|M)P(M) \tag{1}$$

A feladat egyrészt a $P(E|M)$ csatornamodell, másrészt a $P(M)$ forrásmodell meghatározása.

A *csatornamodell* az „eredeti betűhű szöveg → normalizált változat" leképezésekből áll elő. Ehhez szükségünk volt egy tanítókorpuszra, amely két ómagyar kori szövegemlék (Müncheni emlék, Szabács viadala) nyelvész szakértők által kézzel normalizált változatából állt elő. A két nyelvemlék tokenszáma (a nem magyar nyelvű részek elhagyásával) összesen 1525. Gépi eszközökkel és kézi ellenőrzéssel karakterszinten párhuzamosítottuk a betűhű és a normalizált szövegváltozatokat, így a tanítókorpusz körülbelül 17.000 megfeleltetést tartalmaz. Ebből már kiszámítható az egyes megfeleltetések valószínűsége.

A *forrásmodell* azt modellezi, hogy a normalizált szövegben milyen valószínűséggel szerepelnek bizonyos karakterszekvenciák. Mivel a normalizált szöveg a mai magyarhoz nagyon hasonló, a forrásmodell előállításához a rendelkezésünkre álló mai magyar szövegeket tartalmazó korpusz megfelelő. Ezért ezt a Magyar Nemzeti Szövegtár (Váradi, 2002) egyik alkorpuszából, mintegy 10 millió szóból, 65 millió karakterből állítottuk elő.

Adott E sztring esetén az (1) képlet szerinti $\hat{M}$ értéket kellett kiszámítanunk. Ehhez az eredeti betűhű szöveg minden tokenjéből a csatornamodell megfeleltetései alapján a lehetséges normalizált változatokat legeneráltuk, melyekhez a modell hozzárendelte a valószínűségüket is. Ennek alapján kaptunk egy rangsort a lehetséges változatokra, amelyet aztán a forrásmodell segítségével újrarendeztünk – így alakult ki az eljárás kimenete. (Az eljárás teljes leírásához ld. Oravecz et al. (2009, 2010).)

A kimenet minden egyes ómagyar szóhoz a legjobb n normalizált alakot tartalmazó lista. Ennek illusztrációja a 2. ábrán látható. A módszer valós haszna abban mutatkozik meg, hogy a manuális annotáció redukálható a felkínált alakok közötti választásra, ami jelentősen felgyorsítja a normalizálási munkát.

```
fwl (fül)=>                          ygen (igen)=>
-8,80780895229285       föl          -10,8729908279143       igén
-10,7227286786192       fel          -11,3178857141749       igen
-11,0558158154337       fül          -11,5989613202567       igény
-11,2756412387919       f◊l          -13,4229320257043       igyen
-12,4574295350367       fol          -14,3578433608162       igin
-12,790296695296        ful          -14,478835649955        igyén
-13,519092302452        fely
honneg (honnét)=>                    sabach (szabács)=>
-19,1117218113907       honneg       -17,2582527599661       szabács
-19,5230300429664       honnég       -18,1187648297282       sabács
-20,8376176340216       honnét       -18,6771909747334       szabacs
-21,8538140705439       honyneg      -19,1848409742852       sábacs
-22,2098585020436       honynég      -19,5520665992527       szabach
-22,5639991398073       hónneg       -19,9685260661797       szabách
```

2. ábra. Legjobb n listák különböző bemenetekre.

6. Morfológiai elemzés és egyértelműsítés

A normalizálásnak két fő célja van: egyrészt ez teszi lehetővé, hogy a sokféleképpen írt szavak összes előfordulását megtaláljuk, másrészt a normalizált szövegváltozat képezi a morfológiai elemző bemenetét. Mivel a normalizálás során az ómagyar szöveget mai magyarra írjuk át, az ez utóbbira kifejlesztett automatikus morfológiai elemzőt viszonylag könnyen tudjuk alkalmazni a nyelvemlékek feldolgozására. Jelen projektben a *Humor* elemzőt használjuk (Prószéky and Kis, 1999). Az egyik normalizálási alapelvünk, hogy minden morfológiai konstrukciót megtartunk, ezért természetesen ki kell bővítenünk a lexikont és a szabályhalmazt bizonyos ma már nem létező, de az ómagyarban még használt nyelvi jelenségek leírásával.

A morfológiai elemző kimenetének egyértelműsítését automatikusan végezzük, utólagos kézi ellenőrzéssel. A 2. fejezetben ismertetett Jakab-féle táblázatok konvertálásával előállt normalizált és morfológiailag egyértelműsített anyag tanítókorpuszként tud szolgálni egy gépi egyértelműsítő számára. Ennek a kimenetét aztán – a gépi normalizáló kimenetének kezeléséhez hasonlóan – kézzel ellenőrizzük.

Már a normalizálás során felmerül az a probléma, hogy vannak olyan ómagyar szóalakok, amelyeket a szövegkörnyezet alapján sem lehet egyértelműen normalizálni. Például: BécsiK 253.o.: *kic nē hallottac* [kik nem hallottak/hallották]. Mivel ebben a korban a magánhangzó hosszúságát nem jelölték, és a mondat itt véget ér, nem tudjuk, hogy a *hallottac* szóalak határozott vagy határozatlan ragozású. Az ilyen esetekben a normalizálás, valamint a morfológiai elemzés és egyértelműsítés során is megőrizzük a szóalak alulspecifikáltságát.

7. Automatikus vagy manuális?

Amint láttuk, egy jelentős méretű korpusz előállítása során számos nagy munkaigényű feldolgozó lépést kell megvalósítani. Az egyik lehetőség, hogy aprólékos manuális munkával szavanként dolgozzuk fel, és ellenőrizzük a korpuszt. Ugyanakkor a nyelvtechnológia célja éppen az, hogy bizonyos feladatokat a számítógép segítségével meggyorsítson, vagy egészében automatikusan megoldjon. A modern nyelvtechnológiai eszközök az alapszintű feldolgozó lépéseket (szavakra és mondatokra bontás, morfológiai elemzés) nagy sebességgel, nagy mennyiségű (akár milliárd szónyi) szöveget feldolgozva jó minőségben oldják meg.

Az automatikus nyelvtechnológiai módszerek két nagy csoportra oszthatók: szabályalapú, valamint statisztikai, gépi tanulási módszerekre. Mindkét esetben valamilyen módon a szabályszerűségeket próbáljuk feltérképezni; a két megközelítés között lényegében az a különbség, hogy az ember vagy a gép alakítja-e ki a szabályrendszert. A gépi tanulási módszerek egy jelentős csoportjában az algoritmusok egy mintahalmaz (ún. tanítókorpusz) alapján fedezik fel az összefüggéseket. Ezek az algoritmusok tehát a megfelelő nyelvi információval felcímkézett korpuszok segítségével taníthatók és tesztelhetők.

Az automatikus módszerek jó teljesítményt nyújtanak, de nem hibátlanok. A teljes hibamentesség nem érhető el, de bizonyos területeken (pl. tulajdonnév-felismerés) 95% fölötti teljesítmény is elérhető. Fontos látni, hogy az automatikus módszerek alkalmazása sok esetben egyáltalán nem jelent kompromisszumot a minőség tekintetében, mivel a manuálisan végzett elemzés, címkézés szintén nem hibamentes. Véletlenül is előfordulhatnak hibák az elemző, annotátor figyelmetlensége miatt, ennél fontosabb azonban, hogy a manuális elemzésnek is van egy minőségi határa. Azokban az esetekben, amikor ugyanazt a szövegrészt több ember párhuzamosan annotálja, egyértelműen megmutatkozik, hogy minél nehezebb(en megfogalmazható) egy annotálási feladat, annál kisebb az egyetértés az annotátorok között. Ilyen feladatok esetén már az emberi munka hibaszintjét közelítő automatikus megoldás is jelentős eredmény.

Abban, hogy egy nyelvfeldolgozási lépés megvalósításakor automatikus vagy manuális megoldáshoz folyamodunk, természetesen számít a feldolgozandó anyag mérete is. Kis méretnél reális alternatíva a manuális munka, illetve az automatikus elemzés manuális ellenőrzése, nagy méretnél azonban kizárólag az automatikus feldolgozásra hagyatkozhatunk. Bizonyos speciális vagy újszerű feladatoknál megbízható automatikus eszközök hiányában nagyobb a manuális munka létjogosultsága.

A fejezet további részében a jelen projektben alkalmazott szövegfeldolgozási lépéseket tekintjük át automatizáltságuk szempontjából (vö. 4. táblázat).

Az optikai karakterfelismerés (ld. 3.2. fejezet) feladatára a mai nyelvekre kifejlesztett megbízható automatikus eszközök állnak rendelkezésre. A fő nehézséget az ómagyar anyagban található különleges karakterek: a kombinált diakritikus jelek és a latin ábécén kívüli karakterek kezelése jelenti. Amint ez az 1. táblázatból látható, a tanítható OCR program az alap latin karaktereket kiválóan felismerte, a mellékjeles karakterek esetén azonban jóval gyengébb teljesítményt mutatott. Az OCR kimenetét hibamentessé kel-

lett tennünk, hogy a további feldolgozó lépések tiszta, zajmentes adatokon dolgozhassanak, ezért a hibákat kézi erővel javítottuk. A fenti két lépés együttese tekinthető *félautomatikus* karakterfelismerésnek, mely a begépelésnél (a hosszabb szövegek esetében) gazdaságosabbnak bizonyult.

A normalizálás átfogó nyelvtörténeti szakértelmet igényel, és rendkívül időigényes, emiatt megkíséreltük a manuális munkát automatikus eszközzel segíteni. A statisztikai algoritmus (ld. 5. fejezet) nehezen boldogul az egységes írásmód hiánya miatt nagyon szabálytalan ómagyar szöveg kezelésével, ezért azt a megoldást választottuk, hogy automatikusan felkínálunk valószínű normalizált alakokat, és az ezek közül való választás már kézzel történik. A normalizálás tehát szintén *félautomatikus*.

A meglévő robusztus mai magyar morfológiai elemzőre támaszkodva a morfológiai elemzés *automatikus*an történhetett. Az elemző adaptálásával megbízható ómagyar elemzőhöz jutottunk. Az adaptálás során egyrészt új tövekkel bővítettük az elemző szótárát, másrészt pedig új alakok kezelésére tettük alkalmassá az ómagyar ragozási paradigmáknak megfelelően.

Az utolsó feldolgozási lépést, az egyértelműsítést – melynek során az egyes szóalakokhoz rendelt több alternatív morfológiai elemzés közül választjuk ki a valóban érvényeset –, az OCR-ezéshez és a normalizáláshoz hasonlóan *félautomatikusan* végezzük.

Összefoglalva elmondhatjuk, hogy ha kellően robusztus eszközök állnak rendelkezésre, akkor előnyösebb a gazdaságos, automatikus megoldás választása. De a különféle automatikus módszerek megfelelő eszközök hiánya esetén is segíthetik a kézi munkát, azaz ilyenkor a félautomatikus megoldást választjuk. A tisztán manuális megoldáshoz akkor folyamodunk, ha különösen fontos a hibamentesség, illetve nincs elegendő/megfelelő tanítóanyag az automatikus módszerek tanításához.

8. A korpusz felépítése

A korpusz felépítése, vagyis az egyes szövegszavakhoz tartozó annotációs szintek párhuzamosan alakulnak a szövegfeldolgozottsági szintekkel, melyeket a 4. táblázatban láthatunk. Ezek alapján hat annotációs szintet és öt feldolgozó lépést különíthetünk el.

Ahhoz, hogy a korpuszban a nyelvi jelenségek kereshetők legyenek, vagyis az adatbázis használható segédeszköze legyen az elméleti nyelvészeti és nyelvtörténeti kutatásoknak, a releváns információkat elektronikusan előhívható és interpretálható módon kell tárolni. A kifinomult, nyelvészetileg releváns lekérdezések sok esetben különféle nyelvi szinteken megjelenő információra hivatkoznak. Hogy ezek mind elérhetőek legyenek, adatbázisunk párhuzamosan tartalmazza a 4. táblázatban látható szövegfeldolgozottsági szinteknek megfelelő nyelvi adatokat. Vagyis minden egyes szövegszóhoz a következő adatok tartoznak:

- betűhű forma (3): *adïjad*
- normalizált alak (4): *adjad*

(1)	kiadott kódex szkennelve
	→ *automatikus* OCR
(2)	nyers OCR-kimenet
	→ *kézi* javítás, kódolás
(3)	betűhű elektronikus forma
	→ *félautomatikus* normalizálás
(4)	normalizált forma
	→ *automatikus* morfológiai elemzés
(5)	szótövesített és morfológiailag elemzett forma
	→ *félautomatikus* egyértelműsítés
(6)	egyértelműsített korpusz

4. táblázat. Szövegfeldolgozottsági szintek.

- szótő (6) alapján: *ad*
- morfológiai elemzés (6): *[V.Sub.S2.Def]*

A korpusz anyaga vertikális fájlok formájában készül el. Ezek `.tsv` formátumú táblázatok, melyek soronként egy szövegszót tartalmaznak. Az egyes szövegfeldolgozottsági szintekhez tartozó információkat a megfelelő oszlopokban kódoljuk, ahogy azt az 5. táblázat mutatja (a példa a Bécsi kódexből származik, amelynek a morfológiai elemzése és egyértelműsítése még nem készült el, ezért nem szerepel benne a szótő és a morfológiai információ).

kéz	**könyv**	**oldal**	**fejezet**	**vers**	**betűhű**	**norm**	**ért**	**megj**
1	Rut	4	2	8	Es	és		
1	Rut	4	2	8	monda	mondá		
1	Rut	4	2	8	Booz	Boász		
1	Rut	4	2	8	[Noėminèc]			FAIL
1	Rut	4	2	8	Rutnac	Rutnak:		
1	Rut	4	2	8				
1	Rut	4	2	8	Halgassad	hallgassad,		
1	Rut	4	2	8	leaṅom·	leányom:		

5. táblázat. A vertikális fájlformátum.

A korpusz a különböző szinteken feldolgozott szövegen kívül számos metaadatot tartalmaz. Az elsődleges metaadatok az ún. *lókuszjelölők* (ld. az 5. táblázat első öt oszlopát), melyek megadják, hogy a dokumentumban hol szerepel az éppen aktuálisan keresett szövegszó. A lókuszjelölők szövegenként változnak, de annyiban megegyeznek, hogy mindig az eredeti kódex helyeire vonatkoznak, nem a nyomtatott kiadáséira. A például hozott Bécsi kódex esetében rögzítjük a kódexmásoló kezek sorszámát, valamint a bib-

liai könyv- és versszámozást is, hogy más bibliakiadásokban is visszakereshető legyen az adott szövegrész.

A vertikális fájl tartalmaz egy *értelmezés* mezőt is, amelybe a normalizált alak mai magyarra való „fordítása” kerülhet, például az ómagyar *jonh* szó mai magyar *szív* megfelelője. Az a tény, hogy külön mezőben rögzítjük az értelmezést, természetesen nem jelenti azt, hogy a normalizálás során nem történik értelmezés. Normalizálás és értelmezés szorosan összefüggenek, az utóbbi feltétele az előbbinek. Például az Ómagyar Mária-siralom *buthuruth* szavát csak akkor tudjuk normalizálni, ha rájövünk, hogy ennek a jelentése 'bútőr, a fájdalom tőre' (Korompay, 2003).

A megjegyzés rovat egyrészt szabad szöveges megjegyzések rögzítésére alkalmas, másrészt ide kerülnek a szöveghez tartozó egyéb metaadatok is különböző kódok formájában. A korpusz az alábbi metaadatokat tartalmazza:

- Ha a cím a szöveg része, akkor szövegként kódoljuk, és a megjegyzés rovatba `TITLE` kód kerül. Ha a cím nem a szöveg része, akkor lókuszjelölőként funkcionál, vagyis külön oszlopot kap.

- A szövegekben előforduló idegen nyelvű szavakat, amelyek a szöveg részét képezik, felvesszük a korpuszba, és a `LANG{nyelv}` címkét adjuk nekik, amellyel egyben azt is jelezzük, hogy ennek a szónak nem lesz morfológiai elemzése. Ha az idegen nyelvű szó magyarul ragozódik, akkor magyar szóként kezeljük, vagyis normalizáljuk, és elemezzük.

- A betűhű szövegváltozat a szkriptor javításait is tartalmazza. Ezeket a következőképpen jelöljük: szkriptor általi utólagos *betoldás* (kód: `ADD`), *szövegrekonstrukció* eredményeként létrejött betoldás (kód: `RECO`), az eredeti szövegben szereplő *áthúzott* szöveg (kód: `STRIKE`), a szkriptor által *elírt*, de nem áthúzott szó (kód: `FAIL`), *töredékes* szó (kód: `FRAG`). Ha csak a szó egy részét érinti a felsorolt jelenségek valamelyike, akkor kerek zárójellel megjelöljük a betűhű mezőben – és lehetőség szerint a normalizált mezőben is – a megfelelő szórészt. Például:

betűhű	**normalizált**	**megjegyzés**
uimagg(om)uc	imádju(n)k	ADD
sumha	soha	
nym	nem	
kyul		FAIL
hyul:	hűl.	
teun	tőn	
l		FRAG

A metaadatokkal ellátott vertikális fájlt XML-lé alakítjuk, így végezzük el a validációs lépéseket, melyek az adatbázis konzisztenciáját ellenőrzik. Egy következő átalakító lépés során alakul ki az alkalmas bemenet a korpuszkezelő rendszer számára.

3. ábra. A korpuszlekérdező felülete. A feltüntetett példában azokra a szavakra keresünk, melyeknél a normalizált alak kezdete a *jonh* sztring.

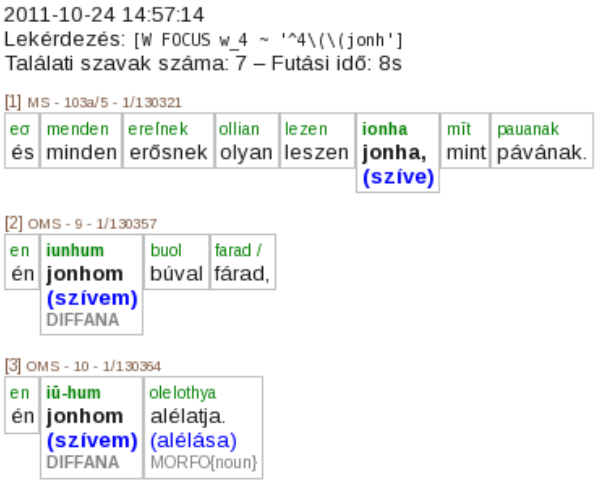
9. A korpuszlekérdező eszköz

A korpusszal párhuzamosan készül a hozzá tartozó korpuszlekérdező felület, amelynek segítségével a teljes ómagyar korpuszt kutathatjuk. Ez jelenlegi állapotában az *Emdros* (Petersen, 2004) korpuszkezelő rendszerre épül. A korpusztalálatok megjelenítése független a lekérdezéstől, abban az értelemben, hogy igény szerint bármilyen – a lekérdezésben esetleg nem is szereplő – szövegfeldolgozottsági szintet is megjeleníthetünk. Ezenfelül lehetővé tesszük a több szintre való egyidejű hivatkozást akár egy kérdésen belül is. Ha például az a kérdésünk, hogy milyen szavak szerepelnek egy igealak és egy igekötő között, akkor az elemzések szintjén (6) kell megfogalmazni a kérdést. Ha gyakorisági listát készítünk a korpusz egy részéből, akkor ezt megtehetjük például a szótövekből kiindulva, de rá lehet kérdezni közvetlenül az *nç* végű szavakra is, ekkor a (3) szinthez fordulunk (vö. 4. táblázat).

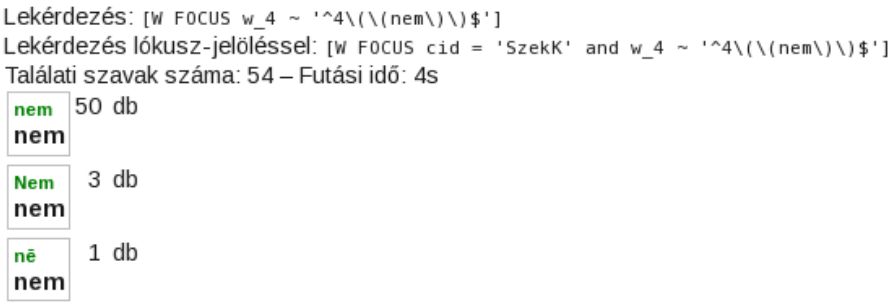
A lekérdező felület a 3. ábrán látható. A felület középső részén adhatjuk meg a lekérdezést, melyben hivatkozhatunk az egyes szövegfeldolgozottsági szintekre, akár többre is egyszerre. Az itt megadott adatokból az *OK* gomb megnyomásával áll elő maga a lekérdezés a bal oldali szövegmezőben az Emdros lekérdezőnyelvén. Ez utóbbi még utószerkeszthető, és a *Mehet* gombbal futtatható.

A 3. ábrán bemutatott lekérdezés eredménye a 4. ábrán látható. A találatok felett a lókuszjelölő található, mely a kódex azonosítójából, az oldalszámból és az adott szó egyedi azonosítójából áll. Az egyes találatokat táblázatos formában jelenítjük meg: fent a betűhű alakot (a felületen zölddel), alatta a normalizált alakot (feketével), majd az értelmezést (kékkel). A felületen (jobb oldalt) a konkordancia mellett alternatív megjelenítési formátumként a gyakorisági lista is beállítható. Az 5. ábrán erre látunk egy példát: a Székelyudvarhelyi kódexben kerestünk rá a *nem* normalizált alakra, és az eredményben a betűhű alakot is megjelenítettük. Láthatjuk, hogy ezen a kódexen belül szinte egységes (és a maival egyező) ennek a szónak a helyesírása, de egy esetben azért előfordul a nazalitást makronnal jelölő régies forma (*nē*) is.

Végül lássunk három ómagyar szintaxisra vonatkozó elméleti nyelvészeti kutatási kérdést, melynek megválaszolásához segítséget nyújthat a korpusz. Mindhárom esetben a (6)-os szintre vonatkozik a lekérdezés, mely a szótövet és a morfológiai elemzést



4. ábra. A 3. ábrán látható lekérdezés eredményének részlete: korpuszpozíciók, ahol a normalizált alak kezdete a *jonh* sztring, konkordancia formájában megjelenítve.



5. ábra. Példa a gyakorisági listás megjelenítésre.

tartalmazza – ennek használatával lehet a nyelvészetileg leginkább releváns kérdéseket feltenni.

A mai magyarban tagadás esetén az igekötő követi az igét (*nem jön be*), az ómagyar viszont az igekötő + tagadószó + ige (*be nem jön*) sorrendet használja legtöbbször (É. Kiss, 2010). Ezt a jelenséget mutatja az alábbi példamondat is: JókK 69.o.: *Ver touaba kÿ nem futott* [Vér továbbá ki nem futott.] A szófajok sorozatára vonatkozó megfelelő lekérdezés a mai magyar szórendre:

```
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
[W FOCUS w_6e ~ 'Vpfx']
```

A lekérdezés az ómagyar szórendre:

```
[W FOCUS w_6e ~ 'Vpfx']
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
```

A `w_6e` jellemzővel a (6) szinten elérhető morfológiai elemzésre kérdezhetünk rá, a tagadószó kódja `Mod`, az ige kódja `V`, az igekötőé pedig `Vpfx`.

Az ómagyarban a mai magyartól eltérő a névelőhasználat: sok helyen nem használnak névelőt, ahol ma igen (Egedi, 2010). Hogyan tudunk alátámasztani egy effajta hipotézist korpusz segítségével, azaz hogyan tudunk rákeresni arra, ami nincs ott? A megoldás az lehet, hogy két olyan szó kombinációjára keresünk rá, melyek között mai intuícióval várnánk a névelőt, de az ómagyarban a két szó névelő nélkül közvetlenül követi egymást. Ilyen konkrét helyzet lehet, mikor definit ige után tárgyesetű főnév áll, mint például ebben a mondatban: JókK 140.o.: *Es azert ewkewztewk zent ferencz czudalatost gÿczerÿuala teremtewtt* [És azért őköztük Szent Ferenc csodálatost dicséri vala Teremtőt.] Ilyen esetekre a megfelelő lekérdezés:

```
[W FOCUS w_6e ~ 'V.*Def']
[W FOCUS w_6e ~ 'N.*Acc']
```

A használt morfológiai kódok: ige: `V`; határozottság: `Def`; főnév: `N`; tárgyeset: `Acc`.

A harmadik kutatási kérdés a *se*-névmások tulajdonságairól szól. Míg a mai magyarban a tagadószó hordozza a tagadást, a *se*-névmások pedig csupán a tagadószóval egyeztetett alakok, a korai ómagyar korban a *se*-névmásoknak önmagukban is lehetett tagadó erejük (É. Kiss, 2010). Ha a *senki/semmi* után közvetlenül egy tagadószótól különböző szót találunk a korpuszban akkor jó eséllyel erre a jelenségre találtunk példát. Az alábbi szövegrészlet éppen ilyen: JókK 8.o.: *mendenestewlfoguan maganac semÿtt meg tarttuan* [Mindenestül fogván magának semmit megtartván]. Ebben az esetben a lekérdezés a következőképpen néz ki:

```
[W FOCUS w_6s ~ '^6s\(\(se[nm][km]i\)\)$']
[W FOCUS NOT(w_6e ~ '^6e\(\(Mod\)\)$')]
```

A *Régi Magyar Konkordancia* nevet viselő lekérdezőfelület szabadon elérhető a `http://corpus.nytud.hu/rmk` címen.

10. További feladatok

Elsődleges feladatunk a teljes ómagyar anyag betűhű szöveges formában való előállítása és kereshetővé tétele. A normalizálást, valamint a morfológiai elemzést és egyértelműsítést csak a korpusz egy részén fogjuk végrehajtani.

Az ómagyar szövegek eleve adott heterogenitása mellett további problémákat okoz az is, hogy a különböző korokban kiadott nyomtatott kódexátiratok tipográfiai kényszerűségek miatt azonos karaktereket eltérően jelenítenek meg. Terveink között szerepel ezen esetlegességek kiküszöbölése, vagyis a különbözőképpen jelölt karakterek azonos sztenderd Unicode-karakterrel való lecserélése.

A projekt vállalásai közé tartozik, hogy a korpusz arányos válogatást tartalmazzon a középmagyar kor (1526–1772) szövegeiből is. Ezen anyagok esetében már fontos szerepet játszik a reprezentativitás kérdése, ugyanis ebből a korból lényegesen több nyelvemlékünk származik, vagyis a teljes anyag feldolgozására ebben a projektben nem vállalkozhatunk. A középmagyar szövegemlékek kiválogatásánál két fő szempontot tartunk szem előtt: csak a már szöveges formátumban elérhető dokumentumokkal foglalkozunk, és ezeket Dömötör (2006) műfaji beosztását követve kategorizáljuk úgy, hogy minden regiszter képviselve legyen a korpuszban.

11. Összegzés

A nyelvi kulturális örökség feldolgozhatóvá és elérhetővé tételében kulcsfontosságú szerep jut a nyelvtechnológiának, amely (fél)automatikus módszereivel hozzásegíti a humán tudományok kutatóit olyan adatbázisokhoz, melyekben a nyelvészeti (és/vagy történeti, paleográfiai stb.) információk elektronikusan előhívható és interpretálható módon vannak tárolva. Az ilyen korpuszok sokkal kifinomultabb keresési lehetőségeket kínálnak, mint az egyszerű digitalizálás, amely általában kimerül a primér adat képként való beszkennelésében. A nyelvtechnológiai eszközökkel feldolgozott történeti szövegeknek további előnyei közé tartozik, hogy a kutatók egységes, akár egy egész korra jellemző, átfogó keresési eredményhez jutnak, mellyel elméleti feltevéseik könnyebben igazolhatóvá válnak.

A nyelvi kulturális örökség feldolgozása a nyelvtechnológusok elé számos kihívást állít. Az elektronikus formátumok előtti korból származó szövegek esetében az eddigieknél robusztusabb vagy teljesen új módszerekre van szükség. Vagyis a kulturális örökség digitalizálása során nem csak a már bevált módszerek új területeken való alkalmazása történik, hanem az új módszerek új kutatási kérdéseket is felvetnek. Ezek megoldásához a különböző tudományterületek képviselői közötti szoros együttműködésre van szükség, melyből meggyőződésünk, hogy hosszú távon minden résztvevő profitálhat.

12. Köszönetnyilvánítás

Az ómagyar korpusz építése a Magyar Generatív Történeti Szintaxis projekt keretében valósul meg. A projektet az OTKA NK 78074. számú pályázata támogatja. Köszönet-

tel tartozunk azoknak a nyelvtörténészeknek és kiadóknak, akik rendelkezésünkre bocsátották az általuk előkészített szöveges kódexátiratokat; továbbá mindazoknak, akik a manuális és/vagy automatikus szövegfeldolgozásban részt vettek. Külön köszönet Novák Attilának, aki a morfológiai elemzést és egyértelműsítést, valamint a Jakab-féle táblázatok átalakítását végzi.

Hivatkozások

Maria Clara Paixao de Sousa and Thorsten Trippel. Building a historical corpus for classical Portuguese: some technological aspects. In *Proceedings of the Vth International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, 2006. ELRA.

Adrienne Dömötör. *Régi magyar nyelvemlékek*. Akadémiai Kiadó, Budapest, 2006.

Barbara Egedi. A határozott névelő. Előadás a Mondattani jelenségek a Jókai-kódexben műhelykonferencián, 2010.

László Jakab. *A Jókai-kódex mint nyelvi emlék szótárszerű feldolgozásban*. Számítógépes Nyelvtörténeti Adattár 10. Debreceni Egyetem Magyar Nyelvtudományi Tanszék, Debrecen, 2002.

László Jakab and Antal Kiss. *A Guary-kódex ábécérendes adattára*. Számítógépes Nyelvtörténeti Adattár 6. KLTE Magyar Nyelvtudományi Tanszék, Debrecen, 1994.

László Jakab and Antal Kiss. *Az Apor-kódex ábécérendes adattára*. Számítógépes Nyelvtörténeti Adattár 7. KLTE, Debrecen, 1997.

László Jakab and Antal Kiss. *A Festetics-kódex ábécérendes adattára*. Számítógépes Nyelvtörténeti Adattár 9. Debreceni Egyetem, Debrecen, 2001.

Anna Adamikné Jászó. Az igenevek. In Loránd Benkő, editor, *A magyar nyelv történeti nyelvtana 2/1. A kései ómagyar kor. Morfematika*. Akadémiai Kiadó, Budapest, 1992.

Gabriella Kiss and Júlia Pajzs. An attempt to develop a lemmatiser for the Historical Corpus of Hungarian. In *Proceedings of CL 2001*, pages 443–451. University of Lancaster, 2001.

István Kniezsa. *Helyesírásunk története a könyvnyomtatás koráig*. Akadémiai Kiadó, Budapest, 1952.

Klára Korompay. Helyesírás-történet (az ómagyar korban). In Jenő Kiss and Ferenc Pusztai, editors, *Magyar nyelvtörténet*. Osiris Kiadó, Budapest, 2003.

Anthony Kroch and Ann Taylor. *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania, second edition, 2000. URL `http://www.ling.upenn.edu/hist-corpora/`. CD-ROM.

Pierre Kunstmann and Achim Stein. Le Nouveau Corpus d'Amsterdam. In Pierre Kunstmann and Achim Stein, editors, *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*, pages 9–27, Stuttgart, 2007. Steiner.

Tony McEnery and Andrew Hardie. *Lancester Newsbooks Corpus*, 2003. URL `http://www.lancs.ac.uk/fass/projects/newsbooks/default.htm`.

Csaba Oravecz, Bálint Sass, and Eszter Simon. Gépi tanulási módszerek ómagyar kori szövegek normalizálására. In Attila Tanács, Dóra Szauter, and Veronika Vincze, editors, *VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2009)*, pages 317–324, Szeged, 2009. SZTE.

Csaba Oravecz, Bálint Sass, and Eszter Simon. 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, Portugal, 2010. Faculty of Science, University of Lisbon.

Ulrik Petersen. Emdros – a text database engine for analyzed or annotated text. In *COLING 2004*, pages 1190–1193, 2004.

Gábor Prószéky and Balázs Kis. A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 261–268, College Park, Maryland, USA, 1999.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In *Proceedings of Corpus Linguistics*. University of Birmingham, 2007.

Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Peter Wynn Thomas, D. Mark Smith, and Diana Luft. Rhyddiaith gymraeg 1350-1425, 2007. URL `http://www.rhyddiaithganoloesol.caerdydd.ac.uk`.

György Volf. *Nyelvemléktár I.* A Magyar Tudományos Akadémia Könyvkiadó Hivatala, Budapest, 1874.

Martin Volk, Torsten Marek, and Rico Sennrich. Reducing OCR Errors by Combining Two OCR Systems. In *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, Lisbon, Portugal, 2010. Faculty of Science, University of Lisbon.

Tamás Váradi. The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 385–389, Las Palmas de Gran Canaria, 2002. European Language Resources Association.

Katalin É. Kiss. A tagadás. Előadás a Mondattani jelenségek a Jókai-kódexben műhelykonferencián, 2010.