DÖMÖTÖR ADRIENNE – GUGÁN KATALIN –
NOVÁK ATTILA – VARGA MÓNIKA

Kiútkeresés a morfológiai labirintusból – korpuszépítés ó- és középmagyar kori magánéleti szövegekből[1]

The paper introduces a novel annotated corpus of Old and Middle Hungarian (16–18th centuries), the texts in which were selected in order to approximate the vernacular of the given historical period as closely as possible. The corpus consists of testimonies of witnesses in trials and samples of private correspondence. The texts are not only analyzed morphologically, but each file contains metadata that facilitate sociolinguistic research. The texts were manually normalized and morphosyntactically annotated using the Hungarian morphological analyzer Humor originally developed for Modern Hungarian but adapted to analyze Old and Middle Hungarian morphological constructions. The paper discusses some of the typical problems that occurred during the normalization procedure and their tentative solutions. Besides, we also describe the query interface. Displaying the original, the normalized and the parsed versions of the selected texts, the first fully normalized and annotated historical corpus of Hungarian is freely accessible at the address http://tmk.nytud.hu/.

Keywords: Historical corpus, Old Hungarian, Middle Hungarian, corpus annotation, morphological analysis, corpus query tool

Kulcsszavak: Történeti korpusz, ó- és középmagyar kor, morfológiai elemzés, keresőfelület

[1] A korpuszt 2010 és 2014 között az OTKA K 81189 számú, *Morfológiailag elemzett nyelvtörténeti korpusz a magánéleti nyelvhasználat köréből* című projektum keretében hoztuk létre az MTA NyTI Finnugor és nyelvtörténeti osztálya több tagjának és külső munkatársaknak a bevonásával. A munkát 2015-től a K 116217 számú, *Versengő szerkezetek a középmagyar élőnyelvben: változók elemzésén alapuló megközelítés* című NKFI-OTKA pályázat részmunkálataként folytatjuk, néhány külső munkatárssal együtt. – A cikk a Language Resources & Evaluation c. folyóirat 2017-es számában angol nyelven megjelent tanulmányunk átdolgozott változata (DOI 10.1007/s10579-017-9393-8).

1. Bevezetés

A tanulmányunkban bemutatandó munkálat legfontosabb célja az volt, hogy olyan elektronikus korpuszt hozzunk létre, amely segítséget nyújt az ó- és középmagyar kori informális nyelvhasználat kutatásához. A korpusz forrásául bírósági jegyzőkönyvek tanúvallomásai és magánlevelek szolgáltak, mivel ezekről a szövegtípusokról tételezhető fel, hogy leginkább megközelítik az élőszóbeli nyelvhasználatot. Annak érdekében, hogy lehetővé tegyük a grammatikailag strukturált lekérdezést, az anyagot morfológiai elemzéssel láttuk el. Ennek megvalósításához a Humor elnevezésű – a mai magyar standardra kidolgozott – morfológiai elemzőt használtuk (Novák 2003; Prószéky – Novák 2005), amelyet tovább kellett fejleszteni, hogy az ó- és középmagyar korban élő, de ma már nem használatos szótöveket és morfológiai szerkezeteket is kezelni tudja. Más (elsősorban konfigurációs) nyelvek történeti korpuszai szintaktikai annotációt is tartalmaznak, mi azonban bizonyos megfontolások alapján nem léptünk túl a morfoszintaktika szintjén. (Az okokról a 7. alfejezetben lesz szó.) A magyar nyelv gazdag alaktani rendszere ugyanakkor lehetővé teszi, hogy a morfológiai elemzések alapján számos szintaktikai jelenségre is rá lehessen keresni.

A korpuszépítés során először a szövegeket karakterfelismerő (OCR) technikával és kézi utóellenőrzéssel digitalizáltuk, majd tagmondatokra bontás után normalizáltuk, vagyis a mai standardhoz közeli változatra írtuk át őket. Ezután a morfológiai elemző segítségével annotáltuk a szövegeket, és az elemzéseket egyértelműsítettük, ami részben automatizált, részben kézzel végzett folyamat. (Az egyes munkafázisokat részletesebben l. lentebb.) Végül az elemzéseket kézzel ellenőriztük és javítottuk. Munkánk eredményeképpen a korpusz tartalmazza az eredeti szövegeket, normalizált változataikat és morfológiai elemzésüket.

Tanulmányunk szeretné bemutatni a korpuszépítési munkának mind a nyelvtörténeti-szociolingvisztikai, mind a számítógépes nyelvészeti vonatkozásait. Az alábbiakban először a korpusz anyagáról szólunk, majd leírjuk a szegmentálás és a normalizálás folyamatát a munka során felmerült nehézségekkel együtt. Ezek után azzal foglalkozunk, hogyan adaptáltuk szövegeinkre a morfológiai elemzőt, illetve itt milyen problémákba ütköztünk. Bemutatjuk továbbá az automatikus és a kézi egyértelműsítő eljárásokat, valamint a korpuszkezelőt, amelynek segítségével az elemzett korpusz kereshető és javítható.

A korpusz anyagát, célját, felhasználási lehetőségeit, valamint a munkálatok közül elsősorban a normalizálást és a kézi egyértelműsítést korábban több tanulmány is az itteninél részletesebben tárgyalta (vö. Dömötör 2009–2011, 2011, 2014 stb.).

A Történeti magánélet korpusz (TMK) a http://tmk.nytud.hu/ linken szabadon hozzáférhető.

2. A korpuszépítés

2.1. Előmunkálatok

Az adatok százainak vagy akár ezreinek összegyűjtése és rendezése korábban a nyelvtörténeti kutatás nem éppen magas presztízsű, de annál időigényesebb fázisa volt. Annak érdekében, hogy ezt a munkaszakaszt megkönnyítsük és lerövidítsük, 2008-ban elhatároztuk, hogy pályázatot adunk be ómagyar nyelvtörténeti adatbázis építésére. Alapvető célkitűzésünk volt, hogy az adatokat ne csak a szavak szintjén, hanem morfoszintaktikai kategóriáik szerint is lehessen keresni. Mindenek előtt azonban látni szerettük volna, hogy a körvonalazandó projekt egyes munkafázisai mennyi időt és erőforrást igényelnek. Ezért afféle pilot-projektként kódexek rövid szakaszait kezdtük feldolgozni, építve a munkacsoport tagjainak korábbi tapasztalataira, amelyeket részben a TNyT. ómagyar fejezetein dolgozva, részben a kódexek szövegkiadásor szereztünk.

Eleinte az összes munkafolyamatot manuálisan végeztük: a választott szövegeket tagmondatokra és szavakra bontottuk, majd az egyes szavakhoz kézzel hozzárendeltük a morfológiai elemzéseket. Ám hamar beláttuk, hogy lényegesen célravezetőbb lenne, ha a szövegeknek elkészítenénk egy olyan átiratát, amelyet automatikus módszerekkel is elemezni lehet. Így kapcsolódott be a munkálatba Novák Attila, aki a mai magyarra kifejlesztett elemzőprogramot átalakította úgy, hogy az ómagyar szövegek normalizált változatát is elemezni lehessen vele. Már az előmunkálatok során lefektettük a szövegek átiratának elkészítési elveit, és az első elemzett mutatványok (a Jókai-, a Müncheni, a Gyöngyösi kódex, a Margit-legenda, a Cornides-, a Keszthelyi és a Sándor-kódex egy-két lapnyi részlete) hosszú ideig elérhetők voltak a Nyelvtudományi Intézet honlapján (http://www.nytud.hu/oszt/finnugor/mutatvany2.html). Hamarosan azonban egy párhuzamos pályázat, a Magyar generatív történeti szintaxis részeként megkezdődött az ómagyar kori szövegek már elektronikus formában meglévő változatainak összegyűjtése, illetve az ilyen változattal még nem rendelkező szövegek digitalizálása. Pályázatunkat így végül részben egy másik korszakra és eltérő regiszterre terveztük meg. A munkát azonban az ómagyar szövegek feldolgozása során kikristályosodott szempontrendszerre támaszkodva kezdhettük meg, míg az ó- és középmagyar szövegek feldolgozására adaptált elemzőt a Magyar generatív történeti szintaxis projektum keretében készült korpusz építése során is fel lehetett használni.

2.2. A források

Az ómagyar korból fennmaradt szövegek túlnyomó többsége latinból fordított egyházi szöveg. Emiatt is kell különös figyelmet kapnia egy másik regiszternek: a nyelvtörténet szempontjából kiemelkedő fontosságú magánéleti nyelvhasználatnak. Olyan korszakokból, amelyekből beszélt nyelvi adatok nem állnak rendelke-

zésre, a nyelvtörténésznek a magánéleti nyelvhasználat közvetett forrásait kell hasznosítania: olyan szövegtípusokat, amelyek az élőnyelvi használathoz a legközelebb állnak. Az ómagyar kori magánéleti regiszter anyaga igen szűkös, a középmagyar kori viszont kifejezetten terjedelmes, mégis viszonylag kevés kutatás merít belőle témákat. A korpuszt boszorkányperek szövegeiből, nemesek, diákok, irodalmi alkotók és jobbágyok leveleiből állítottuk össze. A gyűjtemény így párhuzamba állítható például a következő angol nyelvű korpuszokkal: Corpus of English Dialogues,[2] The Corpus of Early English Correspondence,[3] illetve a korai modern portugál és spanyol leveleket tartalmazó P.S. (Post Scriptum) projektummal.[4]

Bár a bírósági jegyzőkönyvek megszerkesztett iratok, szóbeli vallomásokon alapulnak: a tanúknak minél pontosabban vissza kell idézniük mindent, ami számottevő lehet az adott eljárásban. Így gyakran elevenítenek fel korábban elhangzott párbeszédeket, vitákat, veszekedéseket, fenyegetéseket stb. A magánlevelek pedig kétségtelenül a leginkább dialogikus és interakciós szövegtípust képviselik (vö. Pahta et al. 2010: 7).

Az első magyar nyelvű magánlevél a 15. század legvégéről maradt fenn, míg az első bírósági jegyzőkönyv a 16. század első feléből ismeretes. Emiatt – a fentebbiekkel is összhangban – a korpusz anyaga nem lehet kiegyensúlyozott: az ómagyar kori rész terjedelme messze alatta marad a középmagyar kori szövegmennyiségnek.

2.3. A szövegválogatás és a metaadatok

A korpuszba beépítendő szövegek kiválasztásakor nagy szerepük volt a szociológiai változóknak, mivel az anyagot szociolingvisztikai vizsgálatokra is alkalmassá kívántuk tenni. A jegyzőkönyvek keletkezési kora és helye lefedi az egész középmagyar időszakot, illetve a nyelvterület jelentős részét. Az írás dátuma és helyszíne az adatbázisban metaadatként segíti, hogy az adott szövegre jellemző nyelvváltozatot, nyelvjárást azonosítani lehessen. A boszorkányperekben vádlottként vagy tanúként részt vevő beszélők társadalmi pozíciója szintén ismeretes: a szereplők jellemzően alacsony társadalmi státusú nők és férfiak. A magánlevelek további szempontokat kínálnak a nyelvhasználók jellemzésére, ezek szintén metaadatként kerülnek be az adatbázisba. Ilyenek a feladó neme, társadalmi státusa és a címzetthez való viszonya, a címzett neme és társadalmi státusa, valamint a leírás módja is (a levelet a feladó saját kezűleg vetette-e papírra, vagy diktálás, másolás útján jött-e létre). Bár a keletkezési hely itt is szerepel a metaadatok között,

[2] http://www.helsinki.fi/varieng/CoRD/corpora/CED/index.html
[3] http://www.helsinki.fi/varieng/domains/CEEC.html
[4] http://ps.clul.ul.pt

ennek szociolingvisztikai jelentősége jóval kisebb, mivel a záradékokban olvasható lokalizáció általában nem jelent nyelvjárási kötődést, hiszen nagyon sokszor a levélíró átmeneti tartózkodási helyét nevezi csak meg.

Történeti források esetében a korpuszépítőnek le kell mondania arról, hogy a szociolingvisztikai változók tekintetében kiegyensúlyozott és reprezentatív anyagot hozzon létre – ahogy ezt már korábban mások is megfogalmazták: „Elmúlt korokra vonatkoztatva szinte lehetetlen pontosan meghatározni a teljes megcélzott nyelvhasználó csoportot, ami pedig alapvető fontosságú a reprezentativitás szempontjából (…) annak érdekében, hogy statisztikailag érvényes adatokat kapjunk. A fennmaradt szövegek – nyelven kívüli véletlenek következtében – a teljes népességnek csak kis, random részéhez köthetők. Így egy történeti korpusz megközelítőleg sem tudja megragadni a teljes nyelvi változatosságot." (Claridge 2008: 247; cf. Meyer 2002: 37). Végül is minden korpusz kompromisszum az ideális és a lehetséges között (Hunston 2008: 156).

A Történeti magánéleti korpusz terjedelme jelenleg mintegy 5,9 millió karakter (850 ezer szövegszó) az idegen nyelvű részeket nem számítva.[5] Az anyag 49,5%-a bírósági jegyzőkönyvekből, 50,5%-a levelekből származik. A korpusz folyamatosan tovább bővül, elsősorban a tanulmány elején megadott OTKA-projektek keretében.

2.4. Az annotálás

A tanulmányunk tárgyát képező korpusz az első teljes egészében morfológiailag elemzett magyar nyelvű nyelvtörténeti adatbázis.[6] Az annotálás folyamatát (digitalizálás, tagmondatokra bontás, normalizálás, morfológiai elemzés, egyértelműsítés) a következő három részfejezetben mutatjuk be. Az automatikusan előállított morfológiai elemzést – beleértve a lemmatizációt, a morfoszintaktikai jellemzők elemzését és az elemek részekre bontását – kézzel ellenőriztük és javítot-

[5] Bár a korpuszok terjedelmét általában szövegszóban (token) szokás meghatározni, célszerűnek tartjuk karakterszámban is megadni, hiszen ez utóbbi segítségével jobban összehasonlítható az egyes nyelvek korpuszainak mérete. A tokenszámot ugyanis nagyban meghatározza az adott nyelv típusa. Ezen kívül bizonyos korpuszokban a központozási jelek is önálló tokennek számítanak. A Történeti magánéleti korpuszban a tokenszám a normalizált szövegváltozat elemzett szövegszavainak mennyiségét jelöli. Ez némileg különbözik az eredeti szövegváltozat szószámától; részben az egybeírás-különírás különbségei miatt, részben pedig azért, mert a normalizált szövegváltozatból kimaradnak az idegen nyelvű (zömmel latin) részletek.

[6] Szerepelnek elemzett szövegek a fentebb említett Magyar generatív történeti szintaxis keretében készült korpuszban is (http://omagyarkorpusz.nytud.hu/hu-intro.html). A nyilvánosan elérhető adatok szerint (http://omagyarkorpusz.nytud.hu/hu-texts.html) 2.003.082 tokent tartalmaz, ennek 3,5%-a van elemezve (71.022 token). A TMK jelenleg 850. 000 token terjedelmű anyagának egésze elemezve van.

tuk egy web alapú interfész segítségével (erről majd az 5.1. részben lesz szó), valamint az erre a célra létrehozott felületen végrehajtott lekérdezésekkel (l. a 6. részben).

3. A szöveg előkészítése

3.1. Digitalizálás

A magánéleti regiszter forrásául szolgáló szövegek eredetileg kéziratosak, a korpuszépítéshez azonban nyomtatásban megjelent kiadásaikat használtuk fel. (Digitális kiadásaik nem voltak.) Első lépésként beszkenneltük a szövegeket, majd a FineReader nevű karakterfelismerő (OCR) program segítségével konvertáltuk őket. A szokatlan karakterek és diakritikus jelek nagy száma miatt a feladat nem volt mindig egyszerű.

3.2. Szegmentálás

A digitalizálást követően a szövegeket manuálisan tagmondatokra bontottuk. A tagmondathatárok azonosítása ugyanis a grammatikai fogódzók miatt kevésbé tekinthető problematikusnak (elsősorban az állítmányokra és vonzataikra, bővítményeikre gondolva), mint a mondathatároké. A mondatokra tagolás ebben az esetben sokkal szubjektívebb, önkényesebb lett volna, mivel az eredeti dokumentumokban – a korszaknak megfelelően – mind a nagybetűhasználat, mind a központozás következetlen vagy akár teljesen hiányos, a modernizált közlésekben pedig vitatható módon rekonstruált. A normalizált változatban a mondatokra bontás pusztán technikai célokat szolgált, ez ugyanis a korpuszban használt kereső szövegtagolási alapegysége. A normalizált sorban jelölést kaptak még az adott tagmondatba ékelődő további tagmondatok, illetőleg a mondatátszövődés különböző esetei.

A mai helyesírás szerint történő normalizálás azzal is jár, hogy az eredeti sor és a normalizált verzió szószáma eltérhet. Az ilyen eseteket szintén speciális jelöléssel láttuk el annak érdekében, hogy az eredeti és a normalizált verzió szóalakjai pontosan megfeleltethetők legyenek egymásnak, ennek pedig előfeltétele, hogy az eredeti tagmondat és annak normalizált verziója azonos számú szót tartalmazzon. A balra dőlő törtvonal (\) olyan szóalak mögött áll az eredeti sorban, amely a mai magyarban egybe lenne írva a rá következő szóalakkal, és a fordított esetre is (azaz amikor az eredeti szövegben olyan szavak vannak egybeírva, amelyek a mai magyarban külön lennének) megvan a hasonló eljárásmód (a két szó közé tett @ jellel). Speciális jelölést kapott továbbá, ha az adott szövegrészlet töredékes ({\…}), nem magyarul van (pl. {\!lat!29. Septembris 698.}; ha törölt (pl. Vörös {\uy} Vyz) vagy később beszúrt (pl. Chris{%t}ina) szövegrészt tartalmazott, illetve ha többféle értelmezés volt lehetséges (l. lentebb). A norma-

lizált szövegeket programmal ellenőriztük, hogy kiszűrődjenek a téves megfeleltetések, és ezeket manuálisan javítottuk.

3.3. Normalizálás

Történeti korpuszok leírásakor a normalizálás fontosságát számos szerző hangsúlyozza: olyan átiratokat hozunk létre a szövegekből, amelyek helyesírásukat és fonológiai jellemzőiket tekintve egységesek, egyrészt kibővítve a keresési lehetőségeket, másrészt előkészítve a morfológiai elemzés fázisát. A legnyílvánvalóbb megoldás erre a feladatra a mai helyesírásra való modernizálás lehet; így a szövegek kezelhetők az olyan morfológiai elemzőprogramokkal, amelyeket egy-egy nyelv modern standard verziójára fejlesztettek ki (l. például McEnery – Hardie 2010; Lüdeling – Kytö 2008; Bennet et al. 2010; Hendrix – Marquilhas 2011; Archer et al. 2015). Ez igen időigényes, ugyanakkor megkerülhetetlen feladat, így különféle szoftvereket fejlesztettek ki a folyamat megkönnyítésére, amelyekben manuálisan normalizált tanulókorpuszok segítségével azonosították az írásváltozatokat (Schneider 2002; Rayson et al. 2007; Baron et al. 2011; Archer et al. 2014, 2015; Lehto et al. 2010; Bollmann 2013). Az itt tárgyalt korpusz esetében azonban ezt a lehetőséget elvetettük. Egyrészt a forrásaink helyesírásukat és nyelvjárási hovatartozásukat tekintve igen sokfélék. Majdnem annyi különböző rendszer mutatkozik meg a szövegekben, mint ahány nyelvhasználó azonosítható, ha egyáltalán beszélhetünk egységes rendszerről ebben az időszakban (hasonló problémák merültek fel más hasonló típusú korpuszok építésénél, l. például Hendrickx and Marquilhas 2011). Emellett az automatikus normalizálás elfedne számos lényeges morfológiai kétértelműséget, kettősséget (l. lentebb). Ráadásul több esetben nem volt magától értetődő, hogy az adott karaktersor morfémát tartalmaz, vagy pusztán ejtésbeli változat (pl. *sohon – sehon*; nem lesz belőle *soha* v. *sehol*). Mindezeket mérlegelve a manuális átírás mellett döntöttünk.

A normalizálás fő elve minden egyes esetben az eredeti morfológiai szerkezet megőrzése volt. Csak akkor cseréltünk tehát le morfémákat, hogyha allomorfoknak tekinthetők, beleértve a nyelvjárási variációkat is. Ez ugyanakkor a gyakorlatban összetett problémává vált, mivel a forrásokra jellemző változatosságot éppúgy tekintetbe kellett venni, mint az éppen zajló nyelvi változásokat. Bizonyos kettősségek így is fennmaradtak, ezeket úgy kellett kezelni, hogy a normalizált változat elfogadható bemeneti forrás legyen a morfológiai elemző számára, mégis látsszon, hogy az adott formának több olvasata is lehetséges.

Az alábbiakban egy-egy listát mutatunk be azokról a tipikus esetekről, ahol általános döntést kellett hoznunk: normalizáljuk-e az adott elemet, vagy sem:

Normalizáltuk:
– pusztán helyesírásbeli változat: *szöllő* → *szőlő*, *szarnyajval* → *szárnyaival*; *szállott* → *szállt*
– pusztán nyelvjárási ejtésbeli változat: *gyüvés* → *jövés*; *files bagoly* → *fülesbagoly*; *igenessen* → *egyenesen*
– nevek: *Ersik, Ersok, Ersek* → *Erzsók*
– latin kifejezések, amelyek a magyar szövegbe kerültek magyar toldalékkal: *occurál* → *okkurál*

Nem normalizáltuk:
– történeti szempontból releváns morfológiai információt tartalmazó formák: *oldalaglast* → *oldalaglást*
– latin kifejezések, amelyek formailag nem integrálódtak a magyar szövegbe: *alioquin comburentur*
– kétértelmű formák (éppen zajló változás és/vagy helyesírási kettősség miatt, l. lentebb).

Ezekkel az általánosabb döntésekkel az volt a célunk, hogy az elemző számára a lehető legalkalmasabb szövegváltozat jöjjön létre, azaz minél kevésbé töredékes mondatszerkezeteket kapjon bemenetként; minél több információt lát az elemző, illetve az automatikus egyértelműsítő program, annál pontosabb elemzést kínálhat. A nevek helyesírása például nagy változatosságot mutat, így ezeket is szükséges volt normalizálni (szemben például Hendrix – Marquilhas 2011 megoldásával, ahol a 'név' címkét kapták elemzésként), hiszen gyakran toldalékolt formában fordultak elő, így rajtuk is végre kellett hajtani az automatikus morfológiai elemzést, kezdve az egységes alakú szótő azonosításával. A kódváltást szintén problémaérzékenyen kezeltük. A latin szótöveken is gyakran fordultak elő magyar toldalékok, s ez szükségessé tette az elemzést. Másfelől számos latin elemen nincs explicit magyar morfológiai jelölés, de ennek hiánya (alanyesetű főnevek, határozószók) nem feltétlenül jelent kódváltást. Ezeket tehát elemeztük, ugyanakkor a hosszabb idegen nyelvű szakaszok javítása és elemzése nem volt része a projektumnak. Ez utóbbiakat – a rövidebb, formailag a magyar szövegbe nem épülő kifejezésekhez hasonlóan – a normalizált sorban idegen nyelvűként jelöltük ({ } jelek közé téve), így annotálatlanul maradtak.

Az alábbiakban részletesebben is foglalkozunk az eredeti szövegváltozatban levő két- és többértelműségek kezelésének módjaival.

3.3.1. Kétértelműségekkel kapcsolatos eljárások a normalizált sorban

A) Kettősségek elfedése. – Bizonyos esetekben kénytelenek voltunk a normalizálás során elfedni, hogy az adott elemnek alapvetően többféle elemzése is lehetne. Azokon a szöveghelyeken, ahol a kérdéses elem egyaránt lehet igekötő és határozószó, ott – ige előtti helyzetben – következetesen az egybeírást és az igekötős e-

lemzést választottuk (az eredeti szövegváltozatban azonban megmutatkozik a kettős értelmezés, hasonló problémákról más korpuszokban l. Bennet et al. 2010).

Egy másik olyan eset, ahol a kétértelműséget a normalizálás elfedi, a határozott névelő vs. mutató névmás használatának egyes eseteihez köthető. A korszakban az *az házat* egyaránt értelmezhető volt 'a házat' és 'azt a házat' jelentésben, azonban ha a normalizált változatban az utóbbi olvasatot próbálnánk közvetíteni, az átírásba olyan elem is került volna, amely az eredetiből hiányzik, ez pedig sértené a morfémahűség elvét. Az ilyen esetekben tehát mindig határozott névelőként normalizáltuk és elemeztük a kétértelmű formát (noha az eredeti sorban látszik, hogy másfajta értelmezés is lehetséges).

B) Kettősségek jelölése speciális eljárásokkal. – A kétértelműségek egy másik típusa speciális eljárást igényelt mind a normalizálás, mind a morfológiai annotálás során. Ezekben az esetekben az adott normalizálási eljárás egy azt kiegészítő annotálási eljárással kapcsolódott össze (helyenként ez megkövetelte a morfológiai elemző címkekészletének kibővítését is, ezekről az esetekről a következő pontban esik majd szó). Az inesszívuszi *-bAn* és az illatívuszi *-bA* disztribúciója például a korpuszban (és a korpuszon belül az egyes forrásokban) eltér a mai írásos normától. Azokban az esetekben, amelyekben a szövegbeli használat különbözik a mai írott normától, megtartottuk a szövegben szereplő alakot, de aposztróffal jelöltük az eltérést a mai standardtól (pl. *házba'n*, ha az eredeti szövegben olyan szövegkörnyezetben fordult elő a lokatívuszi alak, melyben a kontextus illatívuszi használatot valószínűsítene, és *házba'*, ha a kontextus alapján a lokatívuszi alak lenne a várható, de a szövegben a latívuszi alak szerepel). A morfológiai elemző a jelölést érzékelve a kontextusnak megfelelő módon annotálja az adott alakot (azaz az illatívuszt igénylő környezetben illatívuszként, a lokatívuszt igénylő környezetben lokatívuszként). Ennek az eljárásnak az az előnye, hogy így ezek az alakok a formára és a funkcióra történő együttes keresés alapján könnyen listáztathatók.

A feloldhatatlan kétértelműség egy további példája a magánhangzók hosszúságának, illetve minőségének nem konzisztens jelöléséből következik. Mivel az ékezetek kitétele esetleges a szövegekben, az elbeszélő múlt egyes szám harmadik személyében a határozatlan és határozott ragozás gyakran nem különböztethető meg, így például az egyik leggyakoribb ige, a *monda* esetében sem. Közismert, hogy a korábbi századokban a mai magyar standardtól eltérő volt a kétféle igeragozás használatának szabályrendszere, így aztán fel sem merülhetett, hogy a mai magyaron alapuló intuíció alapján normalizáljuk ezeket az alakokat. Ezekben az esetekben repülő ékezetet használtunk (*monda'*) a normalizálás során, és az elemző az ilyen alakokat a tárgy határozottsága szempontjából kétértelműnek tekinti. Az E/1 befejezett múlt idejű alak (*mondtam*) és a T/2 elbeszélő múlt idejű alak (*mondátok*) szintén kétértelmű a tárgy határozottságának jelölése szem-

pontjából, de ebben az esetben a kétértelműség oka nem a helyesírási norma hiánya, hanem a paradigma szinkretizmusa. Ugyanakkor a kézi elemzés során (l. alább) ezekhez a formákhoz sem rendelhet az annotátor a mai intuíciója alapján morfológiai elemzést, mert egyértelmű, hogy a vizsgált korban a kétféle ragozás között részben eltérő szabályok szerint választottak a nyelvhasználók. Ezek a formák azonban csak a morfológiai elemzés és a kézi egyértelműsítés során jelölhetők meg inherensen kétértelmű szóalakként, míg azokat az alakokat, melyek esetében a helyesírás következetlensége okozza a kétértelműséget, már a normalizálás során jelölni kell.

Szintén a helyesírás következetlensége miatt egy betűsornak nemcsak két, hanem akár több értelmezése is lehetséges, így például a *halla* alak egyaránt lehet a *halla* (E/3 elbeszélő múlt alanyi ragozás), *hallá* (E/3 elbeszélő múlt határozott ragozás), és – a mássalhangzó-palatalizáció jelölésének következetlensége miatt – a *hallja* (E/3 felszólító mód határozott ragozás, vagy E/3 kijelentő mód határozott ragozás) írott formája. Mivel ez egy nem különösebben gyakori típus, amely egyrészt szóalakok egy szűk körére korlátozódik, másrészt pedig a többértelműség különféle jegykombinációkra vonatkozik, nem pedig csak egyetlen tulajdonságra, a fent bemutatott eljárás itt nem alkalmazható. Ezekben az esetekben a normalizáló kiválasztja a legvalószínűbbnek tartott értelmezést (bár ez maga is lehet egy kétértelmű forma, mint például a *hallá*), és ebben az esetben ez a választás már meghatározza, hogy a morfológiai elemző milyen címkét rendel az alakhoz. Ilyen esetekben a szóalak megcsillagozása hívja fel a korpusz használójának a figyelmét arra, hogy további értelmezések is lehetségesek.

Az inherens kétértelműség egy másik példája a birtokos személyragozás egy nyelvtörténeti-nyelvjárási változatához köthető, amely nagyon gyakori a korpuszban, és elfedi az egyes és többes számú birtok közötti különbséget. A *cselekedetitül* forma például egyaránt jelentheti azt, hogy 'cselekedetétől', és azt is, hogy 'cselekedeteitől', és sok esetben még a kézi egyértelműsítés során, a kontextus figyelembe vételével sem dönthető el, hogy melyik a szándékolt olvasat. Ebben az esetben azt a megoldást választottuk, hogy az eldönthetetlen eseteket speciális formában, az *-i* változat használatával normalizáltuk, tehát a jelen esetben *cselekedetitől*-ként. Ezeket a speciális, nem-standard formákat az elemző felismeri, és az ezekhez társított speciális címkével látja el (bővebben l. az alábbi szakaszt). Ebben a konkrét esetben a használt címke (PxS3.Pl?=i) egyaránt tükrözi, hogy az adott morfoszintaktikai jegy (azaz szám) szempontjából az alak kétértelmű (Pl?), és hogy a toldalék egy nem-standard, neutralizált formában jelenik meg.

A szavak egy speciális csoportjánál (elsősorban határozószóknál és határozói névmásoknál) a mai standard alakra történő normalizálás nem feltétlenül járt volna szigorúan értelmezve a morfológiai szerkezet leegyszerűsítésével, de nyelvtörténeti szempontból mindenképpen értékes morfológiai információ vált

volna nehezen kereshetővé. Például az *ahol* határozói névmás mellett gyakran bukkan fel az *ahon.* Bár a középmagyarban feltehetően már egyik alak sem volt morfológiailag kompozicionális, a kettő közötti különbség mégsem tekinthető pusztán nyelvjárási különbségnek, hiszen különböző primér ragokat őriznek (az-az eredetileg mégis különbözik a morfológiai szerkezetük). Emiatt nem tűnt észszerűnek az a megoldás, hogy a közöttük lévő különbséget neutralizáljuk azáltal, hogy mindkettőt *ahol*-ra, azaz a modern standard formára normalizáljuk. Ezt a jelenségtípust úgy kezeltük, hogy ilyen esetekben megtartottuk az eredeti alakokat a normalizált verzióban is (természetesen csak a morfológiailag releváns különbségeket őrizve meg), lemmáikat pedig összeindexeltük. A gyakorlatban ez azt jelenti, hogy ha a felhasználó a modern standard formára *(ahol)* keres rá, akkor a találati lista egyaránt tartalmazni fogja az *ahol* és az *ahon* előfordulásait is, s ezek közül aztán kiszűrheti az *ahon* előfordulásait, ha ezekre nincs szüksége. Reményeink szerint ez felhasználóbarát eljárásmódnak fog bizonyulni, hiszen segítségével a morfológiai kövületek sokkal könnyebben kereshetők. A szegmentált normalizált verzió, amely egyben a morfológiai elemző bemenete is, az 1. ábrán látható; a vastag betűvel szedett sor a szöveg eredeti változata, alatta a kézi normalizálás eredményeként előálló verzió olvasható, tehát az, amely megközelíti a mai magyar standardot.

1. ábra. Eredeti szöveg és átirata. Bosz. 1a. Abaúj-Torna megye, Szilas, 1736.
Félkövér betűvel szedve: az eredeti tagmondatok,
alattuk soronként: normalizált változatuk.

egy kis idő múlva estve\ feli
Egy kis idő múlva, estefelé,
<még világos vólt>
<még világos volt,>
Tehin\ gyüvéskor gyön Falubul edgy nagy Files\ Bagoly nagy czetajval\ patajval,
tehénjövéskor jön a faluból egy nagy fülesbagoly nagy csetajjal-patajjal,
fel az uton mentiben
fel az úton mentében,
<ahol a szöllő köszt volt,>
<ahol a szőlő között volt,>
oda\ gyött igenessen hozzája,
odajött egyenesen hozzája.
Ember nincsen
Ember nincsen,
aki meg tudna mondanj,
aki meg tudná mondani,
micsoda nagy nyelve ki\ mutatasaval, és feje ki\ Huzásaval, szarnyajval,
micsoda nagy nyelve kimutatásával és feje kihúzásával, szárnyaival,

és nyakat rea\ nyujtogatta,
és nyakát rányújtogatta,
ugy\ hogy elejben edgy karora szállott,
úgyhogy eleibe'n egy karóra szállt.
eszve\ szidta a Fatens,
Összeszidta a fatens:
Te Erdeg\ atta Bekenj menyel előlem
„Te ördögadta Bekéné, menjél előlem,
hagy békét,
hagyj békét!"
soha szöllőben nem maradhatott előtte,
Soha szőlőben nem maradhatott előtte,
hanem a szőllőbő ki kellet tirnem
hanem a szőlőből ki kellett térnem.
oldalaglast mentem
Oldalaglást mentem,
ugy mond
úgy mond.

3.3.2. A csoportmunka menete

Bár az egyes normalizált szövegváltozatok a normalizálást végző munkatársak egyéni munkájának eredményei, a normalizálás maga mégis csapatmunkára épült. A projekt résztvevői egyrészt a rendszeres megbeszélések során folyamatosan megvitatták azokat a problémákat, amelyekről a normalizálási útmutató (addig még) nem tartalmazott információt. Másfelől mindkét szövegváltozat (tehát a digitalizált eredeti és annak normalizált változata is) háromszoros ellenőrzésen ment keresztül a csoportban annak érdekében, hogy minél kevesebb hiba maradjon a szövegekben. Ugyanakkor természetes, hogy ha a normalizálást különböző emberek végzik kézzel, akkor a korpusz normalizált szövegváltozatai nem lesznek teljesen homogének. Annak érdekében, hogy minél kisebbek legyenek az egyéni különbségek, a normalizálási útmutatót folyamatosan frissítettük, amikor újabb és újabb, többféleképpen is kezelhető problémával találkoztunk, hogy az útmutató alapján mindenki a közösen választott megoldás szerint járhasson el a normalizálás során.

4. Morfológiai elemzés

A digitalizált és normalizált szövegeket a Humor morfológiai elemzővel dolgoztuk fel, melyet eredetileg a mai magyar köznyelv elemzésére fejlesztettek ki. A ó- és középmagyar szövegek elemzéséhez tehát módosítani kellett az elemző adatbázisát. Az elemző tő- és toldaléktárát kiegészítettük azokkal az elemekkel, a-

melyek időközben elavultak. A tőtárhoz több mint 5000 új lemmát adtunk hozzá, a toldaléktárba 50 új toldalék került be (nem számítva az allomorfokat). Ezen kívül a morfológiai elemző nyelvtanát kb. 20%-ban kellett módosítani, illetve kiegészíteni, hogy az elemző képes legyen a mai magyar köznyelvben már nem létező morfológiai konstrukciók kezelésére.

Jóllehet a toldalékok egy része nem tűnt el a nyelvből, de elvesztette produktivitását. Bár az ezeket a morfémákat tartalmazó szavak továbbra is részei a magyar szókincsnek, általában lexikalizált elemek a szótárban, gyakran az eredetihez képest módosult jelentéssel. Ugyan ezek a lexikalizált alakok jelen voltak a mai magyar köznyelv elemzésére szolgáló morfológiai elemző lexikonában, ezeket a toldalékokat produktív elemekként fel kellett vennünk a morfológiai elemző történeti szövegek annotálására szánt változatába.

Az egyik tényező, amely megnehezítette az eredeti morfológiai modellünk adaptálását, az volt, hogy nem állnak rendelkezésre megbízható leírások a paradigmák változásáról. Így magukból a szövegekből kellett kinyernünk az arra vonatkozó adatokat, hogy melyik toldalékallomorfok melyik tőallomorfokkal kapcsolódhattak össze. Bizonyos morfológiai (pl. bizonyos igenévi) konstrukciókkal kapcsolatban, amelyek már az ómagyar kor végére kihaltak a nyelvből, nagyon kevés adatot találtunk a forrásokban, és gyakran ezeknek a ritka részparadigmáknak olyan elemei is vannak, amelyekre más elemzés is adható. Emiatt sokszor nem volt nyílvánvaló az, hogy hiányzik a megfelelő elemzés.

Mint a 3.3.1. részben említettük, számos olyan toldalékot kellett felvenni, amelyeket inherensen többértelmű alakok elemzésénél használunk. Ezekben az esetekben az adott toldalékhoz tartozó címkében szereplő kérdőjel jelzi azt, hogy az adott szóalak többértelmű annak a grammatikai jegynek a szempontjából, amelyet a címke jelöl, például: *mondtam*{mond[V.Past.S1.Def?]}, *monda´*{mond[V.Ipf.S3.Def?]}, *kezivel*{kéz[N.PxS3.Pl?=i.Ins]}.

A morfológiai elemző fejlesztése során a legidőigényesebb feladat a tőtár bővítése volt. Amellett, hogy új lemmákat kellett felvenni, számos olyan lexikai tétel lexikonbeli reprezentációját is módosítani kellett, amelyek a mai magyar elemző tőtárában is szerepelnek. Az okok sokfélék voltak, némelyik tő a mai magyarban más szófajú, mint a történeti szövegekben, vagy bizonyos szintaktikai szerkezetekben másképp kell őket elemezni, mint a mai magyarban. Ezen kívül jóval magasabb volt a névmások száma a vizsgált időszakban, mint ma (pl. *tekegyelmed, tinagyságtok, tefelséged, egyetmásaik, ugyanőkegyelmük* stb.). Ezeknek a sok elemből álló és meglehetősen szabálytalan névmási paradigmáknak a leírása komoly kihívást jelentett, különös tekintettel arra, hogy a paradigmák számos eleme meglehetősen alulreprezentált volt a korpuszban.

Néhány olyan fejlesztést, amelyet a történeti szövegek annotálására irányuló projekt során végeztünk az elemzőn, a mai magyar szövegek elemzésére szolgáló elemzőváltozatba is érdemesnek láttunk átemelni. Ilyen módosítás volt példá-

ul az az annotációs séma, amelyet lexikalizált toldalékolt vagy toldalék nélküli főnévi alakok időhatározóként való használatához dolgoztunk ki; például *reggel, nappal*. Ezeknek az alakoknak egy része jelzővel módosítható *(fényes nappal)*. Ez utóbbi tény ezen szavak kettős természetére utal, amelyet úgy ragadtunk meg, hogy ezeket a szavakat főnevek speciálisan toldalékolt alakjaként annotáltuk a-helyett, hogy atomi határozószavakként vettük volna fel őket a lexikonba.

Az eredeti Humor elemző morfológiai címkekészlete alapvetően magyar nyelvű kategóriacímke-rövidítésekből áll. Ezt a címkekészletet kiegészítettük, hogy az ó- és középmagyar morfológiai szerkezeteket is lefedje, és a címkéket a nemzetközi nyelvészközösség számára érthető címkékké alakítottuk. Ugyanakkor a rendszerben használt morfológiai címkék nem követnek pontosan semmilyen nemzetközi szabványt. A lipcsei címkerendszerben (Leipzig Glossing Rules, LGR) javasolt címkékkel bizonyos mértékű átfedést mutat az elemzőben használt címkekészlet, de az LGR csak töredékét fedi le azoknak a morfológiai jegyeknek, amelyeket mi is használunk (nem csak az ó- és középmagyar elemző, hanem a mai magyar elemző esetében is). Ezenkívül a mindkét annotációs sémában szereplő morfológiai jegyek tekintetében is van eltérés a két rendszerben használt rövidítések közt. 2016-ban készült egy LGR alapú teljes annotációs rendszer a mai magyar köznyelvhez (Novák et al. 2017). Terveink között szerepel, hogy a korpuszban szereplő annotációkat ehhez a sémához igazítjuk.

5. Egyértelműsítés

A morfológiai elemző többértelmű annotációt generálhat (l. 2. ábra), ezeket egyértelműsíteni kell. A projektben a morfoszintaktikai annotáció egyértelműsítésére félig automatizált módszert alkalmaztunk (l. 5.2.): az automatikusan előegyértelműsített annotációt kézzel ellenőriztük és javítottuk. Ahogy a kézzel ellenőrzött anyag mennyisége nőtt, az automatikus egyértelműsítéshez használt statisztikai címkézőprogramot inkrementálisan (egyre nagyobb mennyiségű adatot felhasználva) folyamatosan újratanítottuk a korpuszon. A statisztikai egyértelműsítő program betanításához már meglévő annotált anyagra van szükség. Kezdetben ilyen nem állt rendelkezésre, ezért az elsőként feldolgozott korpuszrész egyértelműsítése teljesen manuálisan történt az 5.1. részben bemutatott kézi egyértelműsítő felület felhasználásával. A morfológiai elemzés eredményeként olyan annotáció jön létre, amelyben az adott tagmondatot már három sor reprezentálja: az eredeti szöveg és a normalizált változat mellett a morfológiai annotáció is megjelenik. A lemmát minden szónál szögletes zárójelben álló morfológiai címkék követik, ahogy a 2. ábrán látható.

2. ábra. Egy szövegrészlet automatikus morfológiai elemzést követő, de a kézi egyértelműsítést megelőző annotációja. A többértelmű szavak elemzése félkövér (a képernyőn zöld félkövér) betűvel jelenik meg.

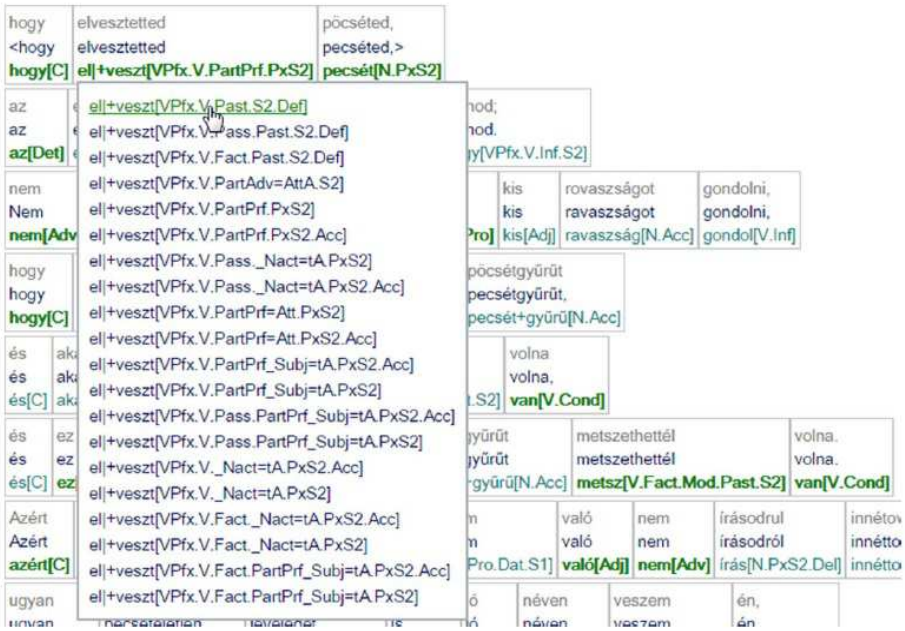
Ezen	Fatens	vallya	azt	hũti	után,
Ezen	**fatens**	**vallja**	**azt**	**hite**	**után,**
ezen[Det|Pro]	fatens[N]	**vall[V.S3.Def]**	**az[N|Pro.Acc]**	hit[N.PxS3]	után[PP]

A történeti szövegek elemzéséhez használt kiterjesztett morfológiai elemző – azonos részletességű elemzések mellett – átlagosan több (2,21) elemzést rendel a szavakhoz, mint a mai magyar elemzésére használt elemző (1,92). Ugyanakkor szélsőséges esetekben az elemző által adott lehetséges elemzések száma az átlagosnál sokkal magasabb lehet. A korpusz tartalmaz olyan, a mai magyarban már nem létező igenévi és passzív szerkezeteket, amelyeknek a morfológiai elemzőhöz adása jelentősen megnövelte néhány viszonylag gyakori igealak lehetséges elemzéseinek számát (l. a 3. ábrán). A nagyobb mértékű többértelműség több tényező következménye:

– a történeti elemző kevésbé normatív: megenged olyan nem standard vagy dialektális konstrukciókat is, amelyek egybeesnek valamely szabályos alakkal (például a suksükölést),
– számos azonos alak szerepel a megnövekedett igei paradigmában, többek között a tömegesen többértelmű faktitív-passzív részparadigma[7] legtöbb eleme és igenévi alakok,
– ezekhez járul még a fentebb leírt sokféle típusú inherens többértelműség.

[7] Ez a többértelműség a mai magyarból a passzív konstrukció kihalása miatt hiányzik.

3.ábra. Az *elvesztetted* igei alak lehetséges elemzései a kézi egyértelműsítés előtt.



5.1. Kézi egyértelműsítés

A korpusz kézi ellenőrzéséhez, illetve a projekt kezdetén a szöveg kézi egyértelműsítéséhez egy olyan, böngészőben működő felületet hoztunk létre, amelyben az egyértelműsítési és normalizálási hibák hatékonyan javíthatók. A rendszer a dokumentumot a korábban említett, könnyen és természetes módon (balról jobbra) olvasható interlineáris annotációs formában jeleníti meg a kézi ellenőrzést, illetve egyértelműsítést végző felhasználó számára. Az adott szóhoz úgy lehet másik elemzést választani, hogy az egérmutatót a szó fölé helyezzük, és az így megjelenő, az adott szó lehetséges elemzéseit tartalmazó listában a megfelelő elemre kattintunk. A lista kizárólag olyan releváns elemzéseket tartalmaz, amelyeket a rendszer hátterében működő webszerveren futó morfológiai elemző az adott szóhoz rendel. Ez a magyar esetében nagyon fontos, mert több ezer lehetséges címke közül lehetetlen lenne a megfelelőt kiválasztani, nem beszélve arról, hogy a lemma előállítása sem mindig triviális, például az ikes-iktelen többértelműségek miatt, mint *tör~törik, múl~múlik* stb. vagy a *-z* képzős igék esetében, ahol az elemző mindkét változatot produktívan generálja, akár létezik az ikes változat, akár nem *(megigéz~megigézik).* A megjelenített eredeti, illetve normalizált szóalak, valamint az elemzés kézzel is szerkeszthető az adott elemre kattintva, és

a szerkesztés után a szóra duplán kattintva a webszerveren futó morfológiai elemzővel újra lehet elemeztetni a kézzel javított szóalakot. Ezt követően pedig már az új elemzések közül választhatjuk ki a megfelelőt a frissített listából. Az elemzés kézi szerkesztésére akkor lehet szükség, ha a morfológiai elemző nem ismeri az adott szóalakot, vagy ha a visszaadott elemzések között nem szerepel az adott szövegkörnyezetben elvárt elemzés. Mivel a morfológiai elemző tőtárát folyamatosan bővítettük a feldolgozott szövegekben szereplő szókincsnek megfelelően, viszonylag ritkán van szükség az elemzés kézi szerkesztésére.

Lényeges tulajdonsága a kézi egyértelműsítő felületnek, hogy a szavakra és tagmondatokra bontással kapcsolatos hibák javítására is alkalmas. Ez egyrészt azért fontos, mert inherens különbség van az eredeti és a normalizált szövegváltozat szavakra bontásában (a tokenizálásban). Másrészt pedig azért, mert a normalizált változat gondos ellenőrzése után is előfordulhat, hogy a szöveg nem megfelelően van szavakra bontva. A tagmondatokra bontással kapcsolatban elsősorban a beágyazott tagmondatok megfelelő annotációjával kapcsolatban merült fel probléma. A szavakat, illetve tagmondatokat a szükséges helyen kettévágni, illetve összeolvasztani is lehet.

4. ábra. A böngészőben működő egyértelműsítő felület.

Ha a morfológiai elemző több lehetséges elemzést rendel egy szóalakhoz, akkor a statisztikai egyértelműsítő program a legvalószínűbb elemzést automatikusan kiválasztja (l. az 5.2. részben), de ezekben az esetekben az elemzés mindig zölddel kiemelve jelenik meg, szemben az egyértelmű szavak kék annotációjával (l. 4. ábra). A kézi annotátorok feladata, hogy az elemzett szöveget ellenőrizzék abból a szempontból, hogy a program választása helyes volt-e az adott szövegkörnyezetben. A 3. ábrán egy olyan igealak látszik, amelynek számos lehetséges elemzése van, amelyek közül a rendszer nem a megfelelőt választotta. Az 5. ábrán ugyanez a tagmondat látható az adott igealak helyes kézi egyértelműsítése után.

5. ábra. Kézzel egyértelműsített, javított tagmondat.

hogy	elvesztetted	pöcséted,
<hogy	elvesztetted	pecséted,>
hogy[C]	el|+veszt[VPfx_V_Past_S2_Def]	pecsét[N_PxS2_Acc]

A szövegek kézi ellenőrzése mellett a morfológiai elemző adatbázisának ellenőrzése és bővítése is munkaigényes feladat volt a projekt során, és ez szoros együttműködést kívánt a csoport tagjai között. Visszatérő probléma volt a morfológiai elemző által használt címkék finomhangolása; többek között új elemzést kellett hozzáadnunk bizonyos névutói alakokhoz (például: *az kenésnek utána*). Ez felvet egy olyan dilemmát, hogy bár a lehetséges címkék számának a növelése az adott esetben pontosabb morfológiai és szintaktikai elemzést tesz lehetővé, ugyanakkor megnehezítheti a különböző korpuszokból származó adatok összehasonlítását. Helyzetének megkönnyítése érdekében a felhasználót a lehető legrészletesebben tájékoztatjuk a címkék értelmezéséről a korpusz felhasználói leírásában.

Az automatikus annotálórendszert úgy alakítottuk ki, hogy lehetőséget biztosítson arra, hogy a munka folyamán megváltoztathassuk az alkalmazott annotációs séma egyes részleteit, ha úgy látjuk, hogy erre szükség van. Az egyik ilyen módosítás például a korábban említett időhatározók annotációjának megváltoztatása volt. A módosított annotációt az adott változtatást megelőzően egyértelműsített szövegekbe is viszonylag könnyen át tudjuk vezetni. Ezt az biztosítja, hogy a szövegek újraelemzésekor a program automatikusan a korábban választotthoz leghasonlóbb elemzést választja ki. Ugyanakkor minden olyan szót speciális kiemeléssel jelöl meg, amelyek esetében az újraelemzés az annotáció megváltozásával járt, hogy a kézi annotátorok könnyen ellenőrizhessék ezeket a pontokat. Azokban az esetekben, ahol az annotációs sémát mélyrehatóbban megváltoztattuk, és ahol ez az egyszerű hasonlóság alapú heurisztika várhatóan nem adott volna kielégítő eredményt,[8] kifinomultabb módszert alkalmaztunk az annotáció frissítésére: automatikusan generált reguláris kifejezésekkel cseréltük le a régi elemzéseket, amelyet a morfológiai generátor kézzel ellenőrzött kimenetének felhasználásával hoztunk létre.

5.2. Automatikus egyértelműsítés

Az első néhány dokumentumot teljesen kézzel egyértelműsítettük a webböngészőben működő eszköz segítségével. Amikor megfelelő mennyiségű anyag összegyűlt ahhoz, hogy egy statisztikai egyértelműsítő eszközt betanítsunk, ennek

[8] Például amikor bizonyos képzett alakokhoz a korábbinál részletesebb elemzés hozzárendelése mellett döntöttünk.

segítségével előegyértelműsítettük az annotációkat, és a kézzel kijavított annotált szövegeken inkrementálisan újratanítottuk az egyértelműsítőt. Először a rejtett Markov-modellen alapuló HunPos szófaji címkéző eszközt használtuk (Halácsy et al. 2007). A HunPos nem tud lemmatizálni, csak egy morfológiai címkét rendel a szavakhoz, ezért a következő egyszerű módszerhez folyamodtunk, hogy teljes morfológiai elemzést kapjunk: a csak címkékkel annotált szöveget újraelemeztettük a morfológiai elemzővel, és a címkéhez leghasonlóbb elemzést választottuk. Ez a módszer viszonylag jó eredményt adott, de volt vele egy probléma: a hasonlóság alapú sorrendezés mindig a rövidebb lemmákat részesítette előnyben. Ez az egyik leggyakoribb lemma-többértelműségi osztály, az ikes-iktelen igepárok esetében nem adott megfelelő eredményt, mert a mindig az iktelen változatot választó algoritmus az ebbe a többértelműségi osztályba tartozó gyakori igék nagy részénél nem a megfelelő lemmát választotta.

Később a HunPos címkéző programot lecseréltük a hasonló statisztikai modellt alkalmazó PurePos egyértelműsítő programra (Orosz – Novák 2013), amely számos további hasznos képességgel rendelkezik. Képes arra, hogy morfológiailag elemzett bementet dolgozzon fel, vagy annotáció közben hívja meg az integrált morfológiai elemzőt. A program tanítóanyagában nem szereplő szavak esetében ezekre az elemzésekre korlátozza az adott szóhoz rendelhető címkék halmazát ahelyett, hogy csak a szó végződése alapján próbálná a lehetséges címkéket megjósolni. Ez a magyarhoz hasonlóan gazdag morfológiájú nyelvek és kisméretű tanítókorpusz esetén nagyon nagy mértékben javítja az egyértelműsítő pontosságát. Emellett a PurePos lemmatizálásra is képes. A tanítóanyagban szereplő lemmák esetén azok gyakorisága alapján választ a morfológiai elemző által adott lemmák közül. A morfológiai elemző számára ismeretlen és a tanítóanyagban sem szereplő szavakat is tudja lemmatizálni. Ehhez a tanítókorpuszból megtanult, a szó végződésén alapuló lemmatizáló modellt használ.

A PurePos egyértelműsítő pontosságát egy 84000 szavas részkorpuszon értékeltük ki. 67000 szónyi anyagon betanítva és 17000 szón kiértékelve 95,9%-os szópontosságot kaptunk. A tagmondatok 81,5%-a nem tartalmazott annotációs hibát, azaz csak minden ötödik tagmondatban kell kézzel hibát javítani. A program pontosságának egyik előfeltétele, hogy már a morfológiai annotáció előtt szerepeljen lexikonjában a korpuszban előforduló szinte valamennyi lemma. Így a tesztanyagban szereplő 17000 szónak mindössze 0,32%-a volt ismeretlen a morfológiai elemző számára.

6. A lekérdezőfelület

A korpuszhoz készített, böngészőben működő lekérdező-felületet nem csak arra tettük alkalmassá, hogy a szövegekben szereplő különböző nyelvtani szerkezetek, illetve maguk a nyelvtörténeti dokumentumok kereshetők és megjeleníthe-

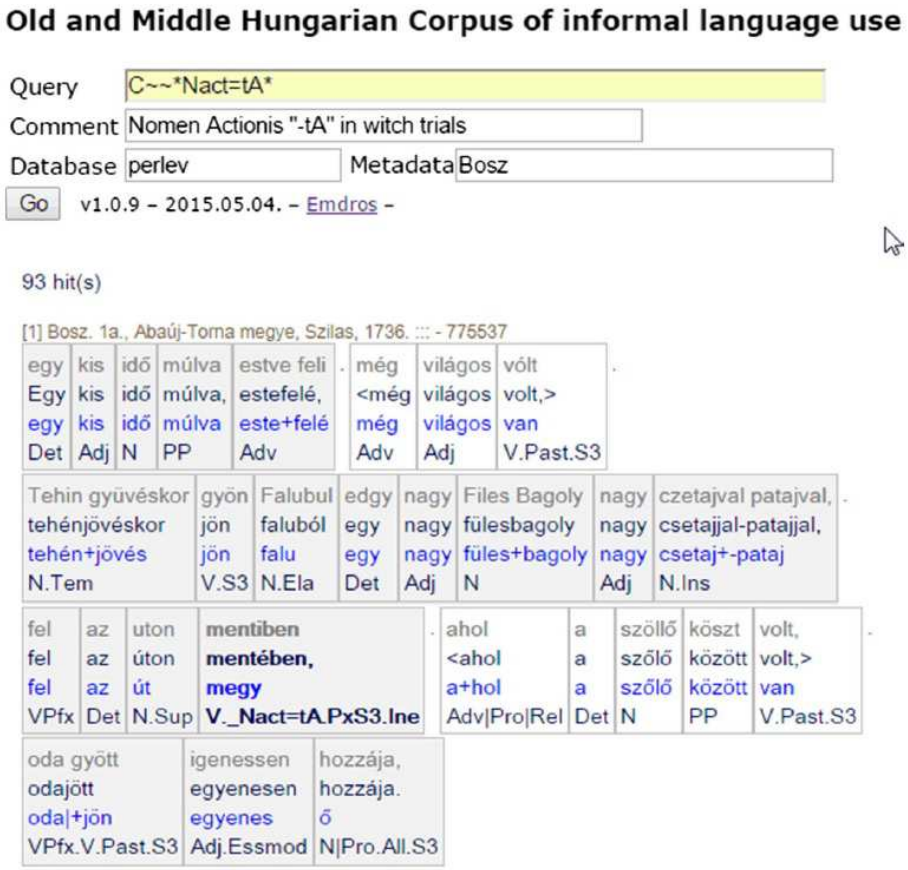
tők legyenek, hanem arra is, hogy hatékonyan használható legyen az annotációs hibák javítására is. Ha egy lekérdezés során hibásan normalizált vagy annotált eredmény jelenik meg a találatok között, az azonnal javítható a webszerveren futó morfológiai elemző segítségével, illetve az adott szó bármelyik jellemzője (eredeti vagy normalizált alak) módosítható, és a javítás azonnal bekerül a kereső által használt korpuszadatbázisba. Természetesen ez a javító funkció csak a megfelelő jogosultságokkal rendelkező annotátorok számára érhető el.

Gyors és hatékony módszer az annotációs hibák javítására, ha a lekérdező-felületen kifejezetten olyan szerkezeteket keresünk, amelyek nagy valószínűséggel hibás annotáció eredményeként álltak elő (pl. determinánst finit igealak követ, stb.), és a ténylegesen hibás eseteket azonnal kijavítjuk az adatbázisban. Ezután a javított korpusz kiexportálható az adatbázisból, és a statisztikai egyértelműsítőt újratanítjuk. A 6. ábrán látható egy példa arra, amikor a keresőfelületet a visszaadott találatban szereplő annotációs hiba kijavítására használjuk. Nem csak az egyes szóalakok és azok annotációinak a javítására van lehetőség, hanem a tagmondatokra bontással kapcsolatos hibák javítására is.

6. ábra. A lekérdezés eredményeként kapott találatban észrevett annotációs hiba kézi javítása.

A korpuszlekérdező által használt adatbázis az Emdros korpuszkezelő és -lekérdező eszközön alapul. A lekérdezéshez az Emdros beépített MQL nevű lekérdezőnyelvén megfogalmazott, vagy a lekérdező-felületen szereplő grafikus elemek segítségével összeállított lekérdezések mellett a haladó felhasználók egy olyan, általunk definiált lekérdezőnyelvet is használhatnak, amelynek segítségével az MQL-nél sokkal tömörebb formában megfogalmazhatóak a lekérdezések (l. 7. ábra). Jól megfogalmazott lekérdezések segítségével hatékonyan kereshetünk példákat sokféle szintaktikai szerkezetre, annak ellenére, hogy a korpusz csak morfoszintaktikai annotációt tartalmaz.

7. ábra. A lekérdező-felület.



A korpusz a grammatikai annotáció mellett számos metaadatot tartalmaz az egyes szövegek szerzőjére, címzettjére, ezek nemére, társadalmi helyzetére és egymáshoz való viszonyára, a szöveg keletkezésének helyére és idejére vonatkozóan. Kereséskor ezeket a metaadatokat is használhatjuk az eredmények szűrésére. A keresett szó, kifejezés vagy grammatikai szerkezet a találatban kiemelve jelenik meg, és alapértelmezésként a találatot magában foglaló mondat jelenik meg kontextusként. Egy kattintással megjeleníthető az adott találathoz tartozó teljes dokumentum, amelyben a kiemelés ugyanúgy megjelenik. A korpuszkezelő képes nemfolytonos tagmondatok kezelésére is; gyakran ilyenek a beágyazott alárendelő tagmondatokat tartalmazó mondatok. De a korpusz sok olyan példát is tartalmaz, amelyben közbevetett mellérendelt szerkezetek jelennek meg, illetve ahol az alárendelt tagmondat topikja megelőzi a főmondatot, amelynek eredményeként a főmondat ágyazódik be az alárendelt tagmondatba. A 7. ábrán látható példa több megszakított tagmondatot is tartalmaz. A csúcsos zárójelek kö-

zött szereplő tagmondatok beékelődnek az általuk megszakított tagmondat topikjába vagy a topik és a komment közé. Az Emdros ezeket a megszakított tagmondatokat nemfolytonos objektumként ábrázolja, amelynek nem része az adott tagmondatot megszakító másik tagmondat.

7. A TMK és más nyelvű történeti korpuszok

Számos történeti korpusz, például a Penn Corpora of Historical English,[9] a Tycho Brahe Parsed Corpus of Historical Portuguese,[10] a Welsh Prose[11] korpusz, a University of Ottawa parsed corpus of historical French,[12] az Icelandic Parsed Historical Corpus (IcePaHC),[13] a The Parsed Old and Middle Irish Corpus (POMIC),[14] a Parsed Corpus of Early New High German[15] és a Penn Parsed Corpora of Historical Greek (PPCHiG)[16] a morfológiai annotáció mellett szintaktikai annotációt is tartalmaz. Ezeknek többsége a Penn Treebank annotációs rendszerének valamilyen adaptált változatára épít, amely a Kormányzás és kötés elméletén alapuló, mondatösszetevőket annotáló séma. Ezzel szemben a TMK csak morfoszintaktikai annotációt tartalmaz, s ez egy elsősorban gyakorlati szempontokat mérlegelő döntés eredménye. Egyrészt a morfológiai elemzéshez már megvoltak az alkalmazható eszközök, hiszen rendelkezésre állt a mai standard magyarra kifejlesztett morfológiai elemző, bár ezt természetesen valamenynyire módosítani kellett ahhoz, hogy a történeti szövegeket is elemezni tudja. Másfelől nem-konfigurációs szintaxisából következően a magyar nyelv szintaktikai annotációja meglehetősen problémás kérdés. Az egyetlen olyan korpusz, amely mondattani fák segítségével elemzi a magyart, a függőségi nyelvtan keretében dolgozik (Vincze et al. 2009).[17]

A projekt erőforrásait lényegesen meghaladó feladat lett volna, hogy kidolgozzuk a szintaktikai annotációhoz tartozó szabályrendszert is, és felvállaljuk, hogy vagy kézzel végezzük el a korpusz szintaktikai annotációját, vagy pedig finanszírozzuk egy új szintaktikai elemző fejlesztését. A szintaktikailag elemzett

[9] https://www.ling.upenn.edu/hist-corpora/
[10] http://www.tycho.iel.unicamp.br/corpus/en/index.html
[11] http://www.rhyddiaithganoloesol.caerdydd.ac.uk/en/
[12] http://www. voies.uottawa.ca/corpus_pg_en.html
[13] http://linguist.is/icelandic_treebank/Icelandic_Parsed_Historical_Corpus_(IcePaHC)
[14] http://www.dias.ie/index.php?option=com_content&view=article&id=6586&Itemid=224&lang=en
[15] https://enhgcorpus.wikispaces.com/
[16] http://www.ling.upenn.edu/*janabeck/greek-corpora.html
[17] A korpuszban használt annotációs séma a függőségi relációk igen szűk készletére épül, amelyben igen sok különböző függőségi viszony össze van vonva.

magyar nyelvtörténeti korpusz egyik alapvető előfeltétele éppen az lenne, hogy legyen egy szintaktikailag annotált korpusz a mai magyarról, amely általánosan elfogadott annotálási elvekre épít – csak ezután lehetne ezeket az elveket kiterjeszteni, illetve módosítani úgy, hogy a nehezebben értelmezhető történeti szövegekre is alkalmazhatók legyenek. Ugyanakkor a magyar nyelv gazdag morfológiájának köszönhetően már a morfológiai annotáció megléte is lehetővé teszi sokféle szintaktikai jelenség vizsgálatát. Egy megfelelően összeállított keresőkérdés pusztán a morfoszintaktikai jegyek alapján is eredményezhet olyan találati listát, amely döntő többségében releváns találatokat tartalmaz egy adott szintaktikai probléma vizsgálathoz.

8. Összegzés

Elvileg a korpuszépítés folyamata lezártnak tekinthető, ha az automatikus elemző kimenetének kézi egyértelműsítése elkészül, és a korpuszhoz illeszkedő keresőfelület hiba nélkül működik. A gyakorlatban azonban mindig van még lehetőség a fejlesztésre. Magától értetődő, hogy a korpusz bővítésének tulajdonképpen nincs határa. Így azonban az újabb és újabb szövegek feldolgozása során újabb és újabb olyan esetekkel találkozunk, amelyek a normalizálás szempontjából problémát okoznak, és így mind a normalizálási útmutató, mind pedig az automatikus normalizálásban szerepet játszó eszközök folyamatosan bővülnek, illetve időnként módosulnak is. Az utóbbi esetben pedig a korábban, más elvek szerint normalizált szövegek elemzését is frissíteni kell, hogy a végeredmény egységes legyen. További célunk, amelyen jelenleg is folyik a munka, hogy a korpusz keresőfelülete lehetővé tegye a szövegekhez társított metaadatok szerinti keresést is, amelyek a szociolingvisztikai és dialektológiai vizsgálatokat teszik könnyebbé. Ha pedig az adatok nemcsak nyelvészeti, hanem a standard szociolingvisztikai faktorok szerint is kereshetők és osztályozhatók lesznek, akkor ez lehetővé teszi majd a változószabály-elemzést, azaz annak feltárását, hogy egy nyelvi változó változatai közötti választást milyen faktorok és faktorcsoportok határoznak meg. Terveink között szerepel az is, hogy egy későbbi írásunkban olyan konkrét kutatási kérdéseket mutatunk be, amelyek a TMK segítségével vizsgálhatók, s ebben további segédlet is szerepel majd a keresőfelület használatához.

Összegzésképpen elmondható, hogy a projekt két kulcsfogalmának a rugalmasság és a dokumentáció bizonyult. Elengedhetetlen volt, hogy megtaláljuk az arany középutat a filológiai, leíró és diakrón adekvátság, valamint azon lehetőségeink között, amelyek a morfológiailag annotált korpusz létrehozásához rendelkezésünkre álltak az eszközöket és adatreprezentációs módokat illetően. A történeti nyelvészek (akik a normalizálást, ellenőrzést és kézi egyértelműsítést végezték) és a számítógépes nyelvész (aki a morfológiai elemzésért és a keresőfelület fejlesztéséért volt felelős) közötti kooperáció mindkét féltől igényelt némi alkalmazkodóképes-

séget, a folyamatos egyeztetések pedig helyenként olyan döntéseket eredményeztek, amelyek ellentétesek voltak a korábbi döntésekkel. Ez együtt járt egy újabb feladattal: a korábbi elvek szerint normalizált és elemzett szövegek frissítésével. A normalizálással és annotálással kapcsolatos szabályok folyamatos, precíz dokumentálása pedig azért volt kulcsfontosságú, mert ez a csapatmunka koordinálása mellett a későbbi felhasználók számára is nélkülözhetetlen lesz, hiszen ennek segítségével lehet a keresési eredményeket megbízhatóan értelmezni. Úgy gondoljuk, hogy ugyanez, tehát a részletes dokumentáció egyben az annotált történeti korpuszok esetleges standardizációjának is előfeltétele.

Végső soron pedig azzal a meglehetősen banális megállapítással tudjuk tapasztalatainkat összegezni, hogy minél inkább felhasználóbaráttá kívánja tenni a korpuszépítő a munkája eredményét, annál több időt és humán erőforrást igényel az adatbázis-építés. A munka során tapasztalt bonyodalmak egy részét feltehetően „hozta magával" a választott forrástípus, azaz az ó- és középmagyar beszélt nyelvet leginkább megközelítő informális szövegek feldolgozása. Abban reménykedünk azonban, hogy hosszú távon megtérül a befektetett munka. A TMK több tekintetben is úttörő projektumnak tekinthető: elsőként dolgozott fel ilyen módon informális szövegeket tartalmazó történeti forrásokat, építése során számítógépes nyelvész munkatársunk kifejlesztette az ilyen típusú munkához szükséges elektronikus eszközöket, és ez az első teljes egészében normalizált és annotált magyar nyelvtörténeti korpusz. A normalizálásnak és annotálásnak köszönhetően az eszköz sokoldalúan használható kutatáshoz, oktatáshoz, vagy pusztán csak a szövegek böngészéséhez is.

Irodalom

Archer, Dawn – Kytö, Merja – Baron, Alistair – Rayson, Paul (2014), Normalising the corpus of English dialogues (1560–1760) using VARD2: Decisions and justifications. In: 35th ICAME conference, April 30–May 04, 2014. Nottingham. Abstract: http://eprints.lancs.ac.uk/72803/ (2017. 12. 19.).

Archer, Dawn – Kytö, Merja – Baron, Alistair – Rayson, Paul (2015), Guidelines for normalising Early Modern English corpora: Decisions and justifications. ICAME Journal. doi:10.1515/icame-2015-0001.

Baron, Alistair – Rayson, Paul – Archer, Dawn (2011), Quantifying early modern English spelling variation: Change over time and genre. In: Conference on new methods in historical corpora, University of Manchester. Presentation: http://eprints.lancs.ac.uk/60258/1/Presentation.pdf (2017. 12. 19.)

Bennett, Paul – Durrell, Martin – Scheible, Silke – Whit, Richard J. (2010), Annotating a historical corpus of German: A case study. In: Proceedings of the LREC 2010 workshop on Language Resources and Language Technology Standards, Valletta,

Malta, May 18, 2010, 64–68. http://www.ims.uni-stuttgart.de/institut/mitarbeiter/scheible/publications/lrec2010.pdf (2017. 12. 19.).

Bollmann, M. (2013), Spelling normalization of historical German with sparse training data. In: Proceedings of the Corpus Analysis with Noise in the Signal workshop (CANS 2013). http://ucrel.lancs.ac.uk/cans2013/abstracts/Bollmann.pdf (2017. 12. 19.).

Claridge, Claudia (2008). Historical corpora. In: Anke Lüdeling – Merja Kytö (eds), Corpus linguistics. An international handbook. Vol. 1, 242–259. Berlin–Nijmegen: Walter de Gruyter.

Dömötör Adrienne (2009–2011), Az alaktanig és tovább: korchmáros, kocsmáros, korchomáros és társai – morfológiailag elemzett történeti magánéleti adatbázis. Nyelvtudomány, V–VII. 13–19.

Dömötör Adrienne (2011), Nyelvtörténet, nyelvváltozat, adatbázis. In: Hegedűs Orsolya – Psenáková Ildikó (szerk.): Tudomány az oktatásért – oktatás a tudományért. I. Univerzita Konstantína Filozofa v Nitre, Fakulta stredoeurópskych stúdií, Nitra, 49–53.

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

Halácsy, Péter – Kornai, András – Oravecz, Csaba (2007), HunPos: An Open Source Trigram Tagger. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, 209–12. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics.

Hendrickx, Iris – Marquilhas, Rita (2011), From old texts to modern spelling: An experiment in automatic normalisation. JLCL, 26(2), 65–76.

Hunston, Susan (2008), Collection strategies and design decisions. In: Anke Lüdeling – Merja Kytö (eds), Corpus linguistics. An international handbook. Vol. 1. 154–168. Berlin–Nijmegen: Walter de Gruyter.

Lehto, Anu – Baron, Alistair – Ratia, Maura – Rayson, Paul (2010), Improving the precision of corpus methods: The standardized version of early modern English medical texts. In: Irma Taavitsainen – Päivi Pahta (eds), Early modern English medical texts. 279–290. Amsterdam: Benjamins.

Lüdeling, Anke – Kytö, Merja (eds). (2008), Corpus linguistics. An international handbook. Berlin–New York: Walter de Gruyter.

McEnery, Tony – Hardie, A. (2010), Investigating the journalism of the seventeenth century. http://www.lancaster.ac.uk/fass/projects/newsbooks/default.htm (2017. 12. 19.).

Meyer, Charles F. (2002), English corpus linguistics. An introduction. Cambridge: Cambridge University Press.

Novák, Attila (2003), Milyen a Jó Humor? In: Alexin Zoltán, Csendes Dóra (szerk.): I. Magyar Számítógépes Nyelvészeti Konferencia, 138–44. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged.

Novák, Attila – Rebrus, Péter – Ludányi, Zsófia (2017), Vincze Veronika (szerk.): Az emMorph morfológiai elemző annotációs formalizmusa. In XIII. Magyar Számítógépes Nyelvészeti Konferencia, 70–78. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged.

Orosz György – Novák Attila (2014), PurePos 2.0: egy hibrid morfológiai egyértelműsítő rendszer. In Tanács Attila; Varga Viktor; Vincze Veronika (szerk.) X. Magyar Számítógépes Nyelvészeti Konferencia, 373–377. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged.

Pahta, Päivi – Palander-Collin, Minna – Nevala, Minna – Nurmi, Arja (2010). Language practices in the construction of social roles in late modern English. In: Päivi Pahta – Minna Nevala – Arja Nurmi – Minna Palander-Collin (eds), Social roles and language practices in late modern English, (Pragmatics and Beyond NS 195). Amsterdam: Benjamins.

Petersen, Ulrik (2004), Emdros — a Text Database Engine for Analyzed or Annotated Text. In: COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, Volume II. Proceedings of COLING, 1190–93.

Prószéky, Gábor – Kis, Balázs (1999), 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 on Computational Linguistics, 261–68. ACL '99. College Park, Maryland: Association for Computational Linguistics.

Prószéky, Gábor – Novák, Attila. (2005), Computational morphologies for small Uralic languages. In: Inquiries into Words, Constraints and Contexts, 116–125.

Rayson, Paul – Archer, Dawn – Baron, Alistair. – Culpeper, Jonathan – Smith, Nicholas (2007), Tagging the bard: Evaluating the accuracy of a modern POS tagger on early modern English corpora. In: Proceedings of the Corpus Linguistics conference: CL2007. UCREL. http://eprints.lancs.ac.uk/13011/1/192_Paper.pdf (2017. 12. 19.).

Schneider, Peter (2002), Computer assisted spelling normalization of 18th century English. In: Pam Peters – Peter Collins – Adam Smith (eds), New frontiers of corpus research: Papers from the 21st International Conference on English Language Research on Computerized Corpora, Sydney, 2000, 199–211). Amsterdam: Rodopi.

Vincze Veronika – Szauter Dóra – Almási Attila – Móra György – Alexin Zoltán – Csirik János (2009), A Szeged Treebank függőségi fa formátumban. In: Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia. 127–138. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged.