Morfológiailag elemzett nyelvtörténeti korpusz a magánéleti nyelvhasználat köréből (a négy év alatt végzett munka és az eredmények bemutatása)

A munkálat célja és anyaga, az adatbázis terjedelme

A munkálat terve azzal a céllal született, hogy létrejöjjön egy megbízható, nagy terjedelmű, felhasználóbarát nyelvtörténeti adatbázis.

Időkörnek elsődlegesen a XVI–XVIII. századot, azaz a középmagyar kort választottuk (alkalmazkodva a korszak jelképes határaihoz: 1526–1772). Korszakválasztásunkat az indokolja, hogy a nyelvtörténeti kutatásoknak eddig kevésbé feldolgozott korszaka a középmagyar (a korábbi kutatásoknak elsősorban az ómagyar állt a középpontjában), ugyanakkor a szakirodalomban egyre fokozódó érdeklődés figyelhető meg a középmagyar kor iránt. Ez az első olyan korszak a magyar nyelv történetében, amelynek gazdag forrásanyaga lehetővé teszi, hogy különböző nyelvi regiszterek anyagát dolgozhassuk fel.

Választásunk ezen belül a magánéleti regiszterre esett, három okból is. Egyrészt: ez a nyelvhasználatnak az a rétege, amelyből a nyelvi változások meghatározóan kiindulnak; anyaguk tanulmányozása tehát kiemelkedően fontos a nyelvtörténeti kutatások számára. Másrészt: elsődlegesen ez lehet az anyaga a kutatásban az utóbbi években–évtizedekben a hazai és a nemzetközi színtéren is előtérbe került történeti szociolingvisztikai és történeti pragmatikai vizsgálatoknak. Harmadrészt: a középmagyar kori – bibliográfiában kiválóan összefoglalt, könyv formában könnyen hozzáférhető – nyomtatott művek mellett a magánéleti nyelvhasználat anyagából készült kiadványok összegyűjtésre, felkutatásra szorulnak.

A magánlevelek kapcsán úgy döntöttünk: a teljesség kedvéért feldolgozzuk az ómagyar korból fennmaradt (csekély számú) misszilist is; létrehozva ezáltal a magyar nyelv magánéleti adatbázisát egészen a kezdetektől a középmagyar kor végéig.

Korábbi nyelvtörténeti korszakok magánéleti regiszterét a magánlevelek és a bírósági jegyzőkönyvek képviselik. Bár – az írásbeli jelleg és az írásba foglalás szűrői következtében – egyik műfaj sem tükrözi közvetlenül az élő nyelvet, mégis ezek azok a források, amelyek a lehető legközelebb engedik a kutatót az élő nyelv, nyelvhasználat tanulmányozásához.

A források kiválasztásánál (a nyelvtörténet és a szociolinvisztika igényeit szem előtt tartva) ügyeltünk arra, hogy változatos időkört és földrajzi eloszlást képviseljenek, és mindkét nemet, valamint a lehető legtöbb társadalmi réteget érintsék. (Gyűjteményünkben magas hivatású főurak, családtagjaik, diákok, jobbágyok levelei is szerepelnek.)

Az eltelt négy év során a következő tíz gyűjtemény magyar nyelvű periratait és leveleit dolgoztuk fel hiánytalanul:

– 1–2. Schram Ferenc kiad.: Magyarországi boszorkányperek I–II. 1529–1768. Akadémiai Kiadó, Budapest, 1970.
– 3. Hoffmann Gizella szerk.: Peregrinuslevelek 1711–1750. Külföldön tanuló diákok levelei Teleki Sándornak. József Attila Tudományegyetem, Szeged, 1980.
– 4. Károlyi Árpád és Szalay József szerk.: Nádasdy Tamás nádor családi levelezése. Akadémiai Kiadó, Budapest, 1882.
– 5. Kincses Katalin kiad.:„Im küttem én orvosságot”. Lobkowitz Poppel Éva levelezése 1622–1640. ELTE Középkori és Koraújkori Tanszék, Budapest, 1993.
– 6. Eckhardt Sándor: Két vitéz nemes úr. Telegdy Pál és János levelezése a XVI. század végéről. Királyi Magyar Pázmány Péter Tudományegyetem, Budapest, 1944.
– 7–8. Kovács Ágnes szerk. / Csobó Péter [et al.] közread.: Károlyi Sándor levelei feleségéhez, 1704–1724. I–II. Kossuth Lajos Tudományegyetem, Debrecen, 1994.
– 9. Szabó T. Attila kiadásai: Jobbágylevelek. Magyar Nyelv, 32., 50–53. évf.

(passim)
– 10. Iványi Béla kiad.: A két Zrínyi Miklós körmendi levelei. Királyi Magyar Pázmány Péter Tudományegyetem, Budapest, 1943.

A pályázati időszakban felépült adatbázis terjedelme: csaknem **4 millió 340 ezer karakter**, ami maximálisan megfelel az előre tervezett szövegmennyiségnek. (Az tervezetben 3 millió 800 ezer–4 millió 600 ezer karakter közötti terjedelmet ígértünk.) Mivel a korpusz tovább bővíthető (és bővítendő), pluszmunkaként további 19 kötetet készítettünk elő a későbben lehetséges munkálatokra (ebből 7 beszkennelve és digitálisan olvashatóvá téve, 12 csak beszkennelve várja a feldolgozás folytatását).

Az adatbázis jellege és felhasználhatósága

Az adatbázis esetünkben nem pusztán digitálisan olvashatóvá tett szövegek gyűjteménye, hanem – és ez képezi a munkálat gerincét – szófaji és morfológiai elemzésekkel ellátott, bármely felhasználónak szabadon rendelkezésére álló forrás, adatgyűjtemény.

A korpusz tartalmazza tehát az eredeti szövegeket (pontosabban a könyv alakban megjelent szövegkiadások számítógépre vitt megfelelőjét), ezeknek a mai nyelvi sztenderdre általunk átírt változatát (erről l. lentebb is) és a szövegek minden egyes szava alatt azok szófaji és morfológiai jellemzését (a nemzetközi hagyományoknak megfelelő jelzésekkel; l. a rövidítésjegyzéket). Az így kialakult három soros lehetőségek mindegyike megjeleníthető, és bármelyikükben végezhető keresés.

Arra törekedtünk, hogy minél megbízhatóbb adatokat szolgáltassunk, és minél inkább felhasználóbarát grammatikai jellemzéseket adjunk. Igyekeztünk jól átgondolt – és a majdani felhasználó várható szempontjait is érvényesítő – kódolási rendszert kialakítani.

A kutatás eredményeit elsősorban különböző nyelvtudományi diszciplínák (és esetleg a rokon szakmák) művelői, tanárai fogják hasznosítani, illetve publikációk és előadások tanúbizonysága szerint hasznosítják már most is. A korpusz megkerülhetetlen forrásként kínálkozik többek között a nyelvtörténészeknek (elsősorban a történeti morfológia és számos kérdésben a történeti szintaxis művelőinek is); a történeti szociolingvisztika, a történeti pragmatika és szociopragmatika művelőinek, grammatikalizációkutatóknak, de adatait haszonnal tanulmányozhatják például a történeti lexikográfia vagy történeti szófajtan szakemberei is. Az adatbázis jól hasznosítható továbbá az egyetemi oktatásban is. Az elemzésben használt rövidítések nemzetközi jellege biztosítja, hogy a magyar nyelv külföldi kutatói is könnyen juthassanak információkhoz.

A létrehozott *Történeti magánéleti korpusz* nyilvános és ingyenes; az MTA Nyelvtudományi Intézet honlapjáról a következő címen lehet elérni: tmk.nytud.hu.

Az elvégzett munkafolyamatok rövid bemutatása

A technikai jellegű (bár szintén igen időigényes) folyamatokat ezúttal nem részletezném. Ezek voltak azok a munkafázisok, amelyek a szükséges források (szövegkiadások) felkutatásától és beszerzésétől a könyvek beszkennelésén keresztül a beszkennelt szövegek karakterfelismertetővel való feldolgozásáig, digitális rögzítéséig, majd korrektúrázásáig terjedtek. Ahogyan fentebb is jeleztem, ezekkel a munkákkal sikerült „előre dolgoznunk”, azaz a felépült korpuszhoz már felhasznált köteteken kívül – többletmunkaként – máris további 7 + 12 könyv digitális változatát készítettük elő a további feldolgozáshoz.

Az elvégzett szakmai feladatokat három pontban lehet összefoglalni.

(1) Először a digitálisan immár rendelkezésünkre álló szövegeket tagmondatokra tördeltük, majd az egész korpusz minden szövegének összes tagmondatát átírtuk a mai nyelvi sztenderdnek megfelelő változatára („normalizáltuk”), úgy azonban, hogy az eredeti szövegből egyetlen morféma se vesszen el (akkor sem, ha nincs mai magyar nyelvi megfelelője), illetve azokhoz egyet se tegyünk hozzá (akkor sem, ha a mai magyarban így kívánkozna). Bár az eredetihez való hűség megtartása magától értetődően hangzik, a gyakorlatban mindez napi gyakorisággal számos megoldandó kérdést vetett fel; ezek kezelésére készült a folyamatosan bővülő normalizálási szabályzat. Erre az igen munkaigényes fázisra azért volt szükség, hogy a mai szövegek feldolgozására kifejlesztett morfológiai elemző a nyelvtörténeti, nyelvjárási, egyéni különbségeket, valamint a hangjelölésében, helyesírásában is nagy változatosságot mutató szövegeket kezelni tudja.

A szövegnormalizálás jól képzett (és a korpuszépítésben is tapasztalatot szerzett) munkatársakat igényel, hiszen kinek-kinek a meglévő szaktudása mozgósításán túl folyamatosan tanulmányoznia kell a történeti szótárakat és nyelvtanokat, hogy a felmerülő kérdésekre megtaláljuk a lehető legmegbízhatóbb válaszokat. Heti gyakoriságú megbeszéléseken próbáltuk a problémákat tisztázni, és minden elkészült normalizált szöveget két további ember javított, hogy minimalizáljuk a hibák számát.

A projekt örömteli hozadéka, hogy a munkálatba igen jó eredménnyel beletanult több, a nyelvtörténet iránt elkötelezett fiatal kutató (l. lentebb is).

(2) A normalizált szövegek ezután a számítógépes nyelvészhez kerültek (aki egyben az adatbázis koncepciójának kidolgozója is: Novák Attila), hogy találkozzanak az általa kifejlesztett morfológiai elemzővel, illetve hogy az ezen találkozás során felmerülő további feladatok is megoldódhassanak (pl. a mai szövegek kezelésére kifejlesztett elemző bővítése a számára addig ismeretlen nyelvtörténeti elemekkel).

(3) Ezek után az elemzett szövegeknek még egy munkafázisra szükségük van: ez az ún. egyértelműsítés. Az automatikus elemző ugyanis egy-egy szóalaknak a formája szerint lehetséges összes elemzését felsorolja (ez nem ritkán 15-20 féle lehetőség), az adatbázisban azonban természetesen csak a megfelelő elemzésnek kell a szó alatt látszania. Első körben az egyértelműsítés szintén automatikus eszközzel történik. Ezen a fázison átment a felépült korpusz teljes anyaga. Az automata egyértelműsítés azonban bizonyos százalékban hibákat hagy maga után; ezért a szövegek még egyszer visszakerülnek a nyelvtörténész munkatársakhoz, akik – egyúttal mintegy utolsó ellenőrzés gyanánt – végignézik az összes elemzést, és kijavítják a helytelen egyértelműsítéseket (és bármely más hibát, amelyet még találnak). Ez a legutóbbi munkafolyamat a vártnál időigényesebbnek bizonyult, így nem végeztük el a korpusz egészén. (Az anyagnak mintegy a negyedét dogoztuk fel ekként is.) Bár tudjuk, hogy a teljes kézi egyértelműsítés általában nem jellemzi az adatbázisokat, mégis azt tartjuk igényes eljárásnak, hogy folyamatosan továbbdolgozzunk a kézi egyértelműsítésen és a fellelt hibák korrigálásán.

Publikációk, előadások

A nagyobb részben megjelent, kisebb részben megjelenés alatt álló hazai és nemzetközi publikációink listáját a megfelelő felületen felsoroltuk. (Lehetőség szerint az URL-jükkel együtt, sajnálatos módon azonban a felület visszautasított több működő URL-t is.)

A projekt négy éve során **31 publikáció**t jelentettek meg a projekt belső (a Nyelvtudományi Intézetben dolgozó), illetve külső (alkalmi szerződésekkel dolgozó) munkatársai. Ezek megoszlása:

– a projektet bemutató publikáció: 9 (6 nemzetközi, 3 magyar)
– a projekthez kapcsolódó, ill. anyagát felhasználó publikáció: 23 (3 nemzetközi, 20 magyar)

A publikációk rovatával kapcsolatban meg kell jegyeznem, hogy számos olyan tételt, amelyek gyűjteményes kötetben megjelent tudományos tanulmányok, kényetelen voltam „konferenciacikk”-ként feltölteni, mivel a felület – érthetetlen módon – nem tartalmazott a valóságnak megfelelő választási lehetőséget.

Mindezeken kívül született **9 egyetemi dolgozat** is (különböző PhD és MA kurzusokon).

A projekt munkatársai összesen **29 előadás**t tartottak a négy év során. Ezek megoszlása:

– a projektet bemutató előadás: 9 (4 nemzetközi, 5 magyar)
– a projekthez kapcsolódó, ill. anyagát felhasználó előadás: 20 (2 nemzetközi, 18 magyar)

Előadásaink időrendben a következők voltak:

– a projektet bemutató előadások:

Dömötör Adrienne: Az alaktanig és tovább: *korchmáros, korcsmáros, korcsomáros* és társai. (Morfológiailag elemzett adatbázis a középmagyar kori magánéleti nyelvhasználat köréből). SzTE, Szeged, Marótiné Korchmáros Valéria 70. születésnapja tiszteletére rendezett felolvasóest.2010. november 25.

Dömötör Adrienne: Morfológiailag elemzett adatbázis az 1772 előtti magánéleti nyelvhasználat köréből. A 81189 számú OTKA-projektum bemutatása. MTA Nyelvtudományi Intézet, Budapest, 2010. december 16.

Novák Attila: Kereső a morfológiailag elemzett nyelvtörténeti adatbázishoz. (Koreferátum.) MTA Nyelvtudományi Intézet, Budapest, 2010. december 16.

Dömötör Adrienne: Az ó- és középmagyar kori magánéleti nyelvhasználat morfológiailag elemzett adatbázisa. VII. Nemzetközi Hungarológiai Kongresszus, Babes–Bolyai Tudományegetem, Kolozsvár, 2011. augusztus 24.

Dömötör Adrienne: Nyelvtörténet, nyelvváltozat, adatbázis. Tudomány az oktatásért – oktatás a tudományért, Univerzita Konstantína Filozofa, Nyitra, 2011. október 21.

Novák Attila: Ómagyar és középmagyar szövegek morfológiai elemzése és egyértelműsítése. MTA Nyelvtudományi Intézet, Budapest, 2012. április 17.

Novák Attila–Wenszky Nóra: O & kozèpmaǵar zoalactanÿ èlèmzo. IX. Magyar Számítógépes Nyelvészeti Konferencia, SzTE, Szeged, 2013. január 8.

Novák, Attila–Orosz, György–Wenszky, Nóra: Morphological annotation of Old and Middle Hungarian corpora. ACL 2013 workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Szófia, Bulgária, 2013. augusztus 8.

Dömötör, Adrienne–Gugán, Katalin–Novák, Attila: Historical Morphology and Annotation: possibilities, procedures, constraints. 16th Diachronic Generative Syntax Conference. Research Institute for Linguistics, Hungarian Academy of Sciences, 3-5 July 2014 Budapest – Workshop: Converging Corpora: How to standardize historical corpora of typologically and genetically different languages, 1 July 2014.

– a projekthez kapcsolódó előadások:

Horváth Laura: A magyar határozói igeneves szerkezetek egy speciális típusáról. A nyelvtörténeti kutatások újabb eredményei VII., SzTE, Szeged, 2012. március 30.

Sipos Mária: „Kerem Aszert Nagisagodat…" – Kérem azért nagyságodat… Normalizált középmagyar szövegek az oktatásban és a kutatásban. Tudomány az oktatásért – oktatás a tudományért, Univerzita Konstantína Filozofa, Nyitra, 2011. október 21.

Varga Mónika: A határozói igenevek állítmányi szerepéről boszorkányperek szövegeiben. 8. Félúton konferencia, Budapest, ELTE, 2012. október 11.

Gugán Katalin: Lehet tévedtem. A tagmondattörlő grammatikalizációs folyamatokról és a *lévén*-ről. MTA NytI., 2012. november 27.

Varga Mónika: *Maguk között mulatván, azután eltűntenek* – a határozói igenév mondatbeli szerepeiről boszorkányperekben. MTA NytI., 2012. november 27.

Orosz György–Novák, Attila: Purepos 2.0: a hybrid tool for morphological disambiguation. International Conference Recent Advances in Natural Language Processing. Hisszar, Bulgária, 2013, szeptember 10.

Varga Mónika: A szövegkohézió tényezőinek vizsgálatáról boszorkányperekben: A referensfolytonosság kérdésköre. Nyelvelmélet és diakrónia – Műhelykonferencia, PPKE BTK, Budapest, 2013. november 19.

Mohay Zsuzsanna: Változás és változatosság a középmagyar korban. 17. századi misszilisek múltidő-használatának vizsgálata Máriássy András, Lobkowitz Poppel Éva és Bethlen Miklós levelei alapján. Nyelvelmélet és diakrónia konferencia, PPKE BTK, Budapest, 2013. november 20.

Faludi Gabriella: Többet ésszel. (Beszélgetés a nyelvtörténetről Dömötör Adrienne-nel.) Klubrádió, 2013. október 31. 13: 33.

Varga Mónika: A határozói igenévi állítmány – és ami körülötte van. 9. Félúton konferencia, Budapest, ELTE, 2013. okt. 3.

Dömötör Adrienne: *Ugyan az, ugyanaz:* kijelölő jelző és azonosító szerep. A nyelvtörténeti kutatások újabb eredményei VIII. 2014. április 4. Szeged.

Gugán Katalin: Nem kár volna foglalkozni velük: tagadószó-igemódosító-ige szórendű mondatok a középmagyarból. A nyelvtörténeti kutatások újabb eredményei VIII. 2014. április 3. Szeged.

Mohay Zsuzanna: Boszorkányperek múlt időben – középmagyar kori múlt idők és használatuk boszorkányperek szövegei alapján. A nyelvtörténeti kutatások újabb eredményei VIII. 2014. április 3. Szeged.

– különböző témájú, de a projektre hivatkozó előadások:

Dömötör Adrienne: Az adat a nyelvtörténetben. A nyelvtörténeti adat: érvény és értelmezés, MTA Nyelvtudományi Intézet, Budapest, 2010. október 1.

Haader Lea: A pragmatika és a grammatika egymásra hatása a kései ómagyar és a középmagyar korban. Grammatika és kontextus, Budapest, ELTE, 2011. április 21.

Dömötör Adrienne: Idéző szerkezetből keletkezett diskurzusjelölők a magyarban. A nyelvtörténeti kutatások újabb eredményei VII., SzTE, Szeged, 2012. március 29.

Gugán Katalin: Ki borotválja a borbélyt? (A magyar és a hanti igekötők grammatikalizációs folyamata: összehasonlítás és annak tanulságai.) A nyelvtörténeti kutatások újabb eredményei VII., SzTE, Szeged, 2012. március 29.

Dömötör Adrienne: Dráma és párbeszéd határán – az egyenes idézés történetének egy fejezete. Drámák határhelyzetben, Konstantin Filozófus Egyetem, Nyitra, 2012. szeptember 5.

Dömötör Adrienne: *Megvizsgálandó, mondván grammatikai változást mutat.* A *mondván* története. MTA Nyelvtudományi Intézet, Budapest, 2012. november 27.

Horváth László: Régi vonzat vénebb vonzat? A nyelvtörténeti kutatások újabb eredményei VIII. 2014. április 4. Szeged.

Az összes fentebbi tétel – publikáció, előadás, egyetemi dolgozat – tartalmazta a munkálat OTKA-számára való hivatkozást.

További hivatkozások is várhatók, egy aktuális kurzus leírása máris tartalmazza: http://pestibolcseszakademia.blog.hu/2014/09/17/betegseg_es_metafora_a_boszorkanyperekben_varga_monika_kurzusa, ahogyan a Nyelvtudományi Intézetben megtartandó projektzáró előadás címe és absztraktja is: http://www.nytud.hu/intprog.html.

Összesen 72 megjelenésről számolhatunk tehát be a projekt munkatársaitól. (Tudomásunk van külső felhasználók előadásaiban és tanulmányaiban megjelent hivatkozásokról is, ezek rendszeres gyűjtését azonban nem tekinthettük feladatunknak.)

A projekt további hozadékai

További két tényt, eredményt fontosnak tartok még megemlíteni. Egyrészt azt, hogy – az OTKA iránymutatásainak is megfelelően – sikerült a munkálatba bevonni és a különböző munkafázisokra betanítani nyelvtörténésznek készülő, tehetséges egyetemi és PhD-hallgatókat. Közülük négyen idekerülésüktől kezdve mindvégig a projekt (alkalmi szerződésekkel idekötődő) munkatársai maradtak, és készek a további együttműködésre is: Varga Mónika (jelenleg utolsó éves doktorandusz), Mohay Zsuzsanna (elvégezte a doktori

iskolát, dolgozatát írja), Horváth Laura (elvégezte a doktori iskolát, dolgozatát írja), Korompay Eszter (MA-hallgató).

Másrészt azt is szeretném rögzíteni, hogy a projektnek mintegy melléktermékeként – a terveken felül, többletmunkaként – egy önálló kiadványt is elkészítettünk (l. a publikációs jegyzékben). A forrásaink közé beemelendő jobbágylevelek – amelyek sok évtizede szétszórtan jelentek meg folyóiratszámokban – kínálkoztak arra, hogy könyv formátumban, a fakszimiléiket is felkutatva publikáljuk őket. Ebből a viszonylag egyszerűnek tűnő eljárásból hosszadalmas munka bontakozott ki, mivel az eredeti publikációk – annak ellenére, hogy kifejezetten nyelvészeti felhasználásra készültek – sok hibát, sőt kihagyásokat tartalmaznak.

Ez a kiegészítő munka kiválóan igazolja, hogy valóban hibátlan korpusz csak eredeti szövegek feldolgozásával építhető. Szövegközlésekből dolgozva a korpuszépítő ki van szolgáltatva a közlő gyakorlatának. Mégis kénytelenek voltunk vállalni ezt a hátrányt, hiszen ellenkező esetben sokszor ennyi idő alatt is a jelenlegi adatbázisnak csak töredéke készülhetett volna el.