Kalcsó Gyula
Eszterházy Károly Főiskola
kgyula@ektf.hu

A MAGYAR ANTIKVAKORPUSZ FEJLESZTÉSE

Előadásom a magyar nyelvű könyvnyomtatás első fél évszázadában megjelent nyomtatványok reprezentatív korpuszának, a Magyar Antikvakorpusznak[1] a fejlesztését mutatja be. A gyűjtemény első változata 2001 és 2005 között jött létre, amikor PhD-tanulmányaim során a korai magyar nyelvű nyomtatott írásbeliség alaktani variánsainak nyelvészeti vizsgálatához egy plain text példatárat készítettem. Az első fennmaradt, magyar nyelvű szövegrészeket is tartalmazó nyomtatvány Christoph Hegendorff Donatus-nyelvtanának *(Rvdimenta grammatices Donati…,* RMNy. I. 7.)[2] 1527-es krakkói kiadása, amelyben – valószínűleg Sylvester János jóvoltából – a német és a lengyel mellett magyar fordításban is szerepelnek a nyelvtani példák. Ezzel indul útjára a magyar nyelvű nyomtatott írásbeliség, és fejlődik töretlenül; ettől számítva a 16. század végéig összesen több mint 900 magyar nyelvű nyomtatványt tart számon a könyvtudomány. Az első néhány évtizednek nyilvánvalóan kiemelkedő jelentősége van: ekkor alakulnak ki azok az alapvető normák, amelyek a későbbi könyvnyomtatást meghatározzák. Az első fél század minden bizonnyal még a kísérletezés időszaka: ezt jól mutatja a sajtó alól kikerülő könyvek száma is: 1576-ig mindössze 196-ról tudunk, a század utolsó negyedében azonban évről évre megsokszorozódik a kiadott művek száma. Az RMNy. sorszámozása szerinti utolsó mű, amelyet a korpusz összeállításakor figyelembe vettem, Valkai András 1576-ban, Kolozsvárott kiadott históriás éneke, a *Genealogia historica regvm Hungariae… Az az az magyar királyoknac eredetekről és nemzetségekről való szép historia* (RMNy. I. 368.)[3].

A korpuszépítés első lépéseként összegyűjtöttem az időszak magyar nyelvű nyomtatványainak adatait. Ebben a *Régi magyarországi nyomtatványok* (RMNy.) c. bibliográfia I. kötete volt segítségemre. Az RMNy. szerint 196 legalább részben magyar nyelvű nyomtatvány jelent meg ebben az időszakban[4], ebből 152 maradt fenn[5]. További 12

[1] http://korpusz.ektf.hu

[2] Itt és a továbbiakban a Régi magyarországi nyomtatványok c. bibliográfia (RMNy.) sorszáma szerint hivatkozom a művekre.

[3] Az RMNy. I. kötete szerint az 1576-ban megjelent művek közül még a 370. sorszámú nyomtatvány is magyar nyelvű, ám csak töredékes formában maradt fönn, így kihagytam a korpuszból.

[4] Itt és a továbbiakban is a megjelenés RMNy. kronológiai sorrendjében: RMNy. I. 7., 8., 11., 12., 13., 14., 15., 16., 17., 18., 19., 20., 21., 22., 23., 24., 25., 26., 27., 39., 63., 47., 48., 49., 57., 58., 65., 64., 70., 74., 77., 78., 81., 85., 88., 80., 86., 88a, 91., 90., 92., 95., 98., 96., 99., 100., 101., 102., 103., 109., 108., 125., 137., 144., 151., 150., 154., 155., 156., 158., 159., 160., 161., 162., 165., 170., 164., 166., 169., 171., 172., 178., 185., 173., 181., 182., 183., 184., 186., 191., 191b, 192a, 193., 195., 192., 194., 196., 206., 205., 207., 208., 213., 218., 219., 220., 222., 230., 237., 238., 240., 229., 232., 233., 257., 240a, 241., 243., 242., 246., 253., 255., 259., 263., 266., 268., 260., 264., 265., 269., 273., 276a, 277., 281., 282., 283., 293., 276., 279., 280., 284., 286., 288., 289., 290., 294., 295., 296., 297., 298., 299., 301., 303., 304/1., 304/2., 307., 308., 308a, 308b, 311., 312., 314., 315., 318., 316., 321., 324., 331., 319., 320., 322., 323., 326., 327., 328., 333.,

műből csak kisebb töredékek tanulmányozhatók[6]. A nagyobb töredékeket beleszámítva tehát 140 fennmaradt könyvből és könyvtöredékből áll a magyar nyelvű nyomtatott írásbeliség első fél századának teljes állománya. Ebből 103 művet választottam ki a számítógépes korpusz összeállításához, így a korszakban megjelent művek több mint fele, a ránk maradt művek több mint kétharmada reprezentálva van. A 37 kimaradt mű négy csoportba sorolható:

- szótárak, nyelvtanfordítások, amelyekben csak szavak, legfeljebb szószerkezetek szerepelnek, ezért funkcionális szempontokat is érvényesítő morfológiai vizsgálatra alkalmatlanok (8 mű)[7];
- újrakiadások és újraszedések, amelyek előzményeikkel lényegében megegyeznek – függetlenül attól, hogy azonos helyen, azonos nyomdász adta-e ki őket (11 mű)[8];
- azonos helyen, közel azonos időben (legfeljebb öt éven belül), azonos szerzőtől, azonos műfajban, azonos nyomdász által kiadott művek közül minden esetben egyet választottam ki, pl. Méliusz Juhász Péternek a debreceni nyomdában 1562-ben négy vallásos prózai műve is megjelent, ezek közül hármat kihagytam (17 mű)[9];
- egyetlen műhöz nem tudtam a korpusz építése során semmilyen betűhív formában hozzájutni: Balassi Bálint *Beteg lelkeknek való füves kertecské*jéhez (RMNy. I. 318.), amelynek egyetlen fennmaradt példánya csak 2006 februárjában került vissza Magyarországra a több mint félszázados szovjetunióbeli, illetőleg oroszországi „lappangás" után. Mivel fotómásolat nem készült róla, valamint 2006-ig betűhív kiadása sem jelent meg[10], ezért kénytelen voltam lemondani a korpuszban szerepeltetéséről.

A 103 kiválasztott szöveg tehát minden tekintetben a lehető legteljesebben reprezentálja a fennmaradt nyomtatványokat: minden szerzőtől, minden kiadási évből, minden nyomdából, minden nyomdásztól, minden műfajból szerepelnek művek a korpuszban, így eleget tesz a minőségi reprezentativitás követelményének. A nyelvészeti vizsgálatokhoz elegendő volt a szövegekből reprezentatívnak tekinthető mennyiségű kiválasztott részleteket rögzíteni. A mintavétel elvei a következők voltak:

- minden műből legalább ezerszavas minta szerepeljen (az ennél rövidebb nyomtatványok teljes terjedelmükben kerüljenek be);
- a terjedelmesebb műveknek legalább 5%-a (azaz átlagosan húsz oldalanként egyoldalnyi részlet) kerüljön a korpuszba;

337a, 338., 339a, 334., 335., 337b, 340., 341., 342., 343., 344., 345., 339., 346., 347., 348., 349., 350., 355., 351., 352., 353., 357., 358., 359., 360., 362., 364., 367., 368., 370.

[5] Az elveszett nyomtatványok: RMNy. I. 19., 20., 22., 23., 25., 26., 27., 47., 48., 57., 58., 65., 70., 81., 85., 137., 150., 161., 165., 191., 193., 195., 206., 230., 237., 238., 240., 257., 263., 266., 268., 273., 276a., 277., 281., 282., 283., 293., 308., 316., 321., 331., 338., 339a.

[6] RMNy. I. 12., 18., 24., 88., 159., 170., 178., 185., 191b, 192a, 364., 370.

[7] RMNy. I. 7., 14., 21., 39., 103., 166., 240a, 241.

[8] RMNy. I. 11, 99, 172, 255, 265, 276, 327, 335, 337a, 352, 357.

[9] RMNy. I. 182., 183., 184., 196., 232., 242., 253., 279., 286., 298., 301., 312., 314., 323., 333., 347., 355.

[10] 2006-ban a Balassi Kiadónál megjelent az első hasonmás kiadás, ezt azonban a korpusz első változatában már nem tudtam figyelembe venni.

- minden műből több helyről szerepeljenek szövegrészletek (de lehetőség szerint ne legyenek a minták túlzottan széttöredezettek);
- a többszerzős művekből – amennyiben az egyes részek szerzői azonosíthatók – lehetőség szerint minden szerzőtől legyen részlet.

Ily módon egy 238 877 szövegszóból (1 176 826 betűhelyből) álló korpuszt választottam ki. Ez 43 ismert és tíz ismeretlen szerző 80 művének, valamint 13 többszerzős nyomtatványnak mintegy a huszadrészét jelenti. Tíz rövidebb szöveg teljes terjedelmében szerepel.

A nyomtatványokat a következő forrásokból tanulmányozhattam. Bizonyos műveknek rendelkezésre áll fakszimile kiadása[11]. Más esetekben az OSZK valamint az MTA könyvtára mikrofilm-állományán, illetőleg a filmekről készült digitális másolatokon keresztül vizsgálhattam a szövegeket. Néhány esetben – a művekről mikrofilm nem lévén – közvetlenül az eredeti mű alapján kellett a kiválasztott részek átírását elvégezni. Így mind a 103 esetben vagy közvetlenül a nyomtatvány szövege, vagy az arról készült fotómásolatok alapján készíthettem el az átiratokat.

Miután 2007-ben megvédtem a korpuszon végzett első nagy vizsgálat alapján írott PhD-disszertációmat, a gyűjtemény internetes publikálása és továbbfejlesztése mellett döntöttem. A webes megjelenítés szerteágazó problematikájából az előadás kizárólag néhány részkérdést érinthet. Ezek közül az első a karakterkezelés. A régi nyomtatványok szövege igen nagy számban tartalmaz még olyan betűket, amelyeket a karakterkódolási szabványok (mint amilyen pl. a Unicode[12]) nem kódolnak. Ez azt jelenti, hogy az egységes megjelenítés a felhasználók számítógépén nehézséget okoz. Megoldást kizárólag az jelenthet, ha az oldalon használt karakterkészlet(ek)et valamilyen módon be tudjuk ágyazni a weboldalba, vagy a felhasználót rá tudjuk venni a speciális fontkészlet telepítésére. Mivel ez utóbbi a felhasználók számítástechnikai jártasságának a függvénye, ezért célszerű az előbbit előnyben részesíteni.

A korpusz kódolásakor egy Unicode-alapú nemzetközi ajánlást követtem. A *Medieval Unicode Font Initiative* (MUFI)[13] alapvetően a középkori (de a korai nyomtatványokban is megjelenő) speciális grafémák kódolását próbálja szabványosítani. Kétféle megoldást kínálnak: a Unicode-ban még nem kódolt karakterek esetében privát kódot (a Unicode kódtábla ún. *Private Use Area*ja[14] kódpontjainak a konszenzusos felhasználását), valamint a Unicode konzorciumhoz benyújtott javaslatok révén a grafémák felvételét a Unicode szabványba. A MUFI 3.0-val kompatibilis fontok közül Andreas Stötzner lipcsei grafikus *Andron Scriptor Web*[15] nevű készletére esett a választásom, mivel ennek reneszánsz arculata jól illik a korpuszhoz. Az *Andron Scriptor Web*et a felhasználók letölthetik, és a saját gépükre telepíthetik. Aki ezt nem teszi meg, az is helyesen látja azonban a szövegeket, mivel Simo Kimmunen *cufón*[16] nevű projektje segítségével a fontot beágyaztam az oldalba. Ez esetben a szövegek vektorgrafikus kép-

[11] Az OSZK állományában lévő művekről készült fakszimilék teljes listája fellelhető az http://regi.oszk.hu/hun/szakmai/hasonmas/hasonmas_index_hu.htm internetcímen.
[12] http://unicode.org
[13] http://mufi.info
[14] http://unicode.org/charts/PDF/UE000.pdf
[15] http://mufi.info/fonts/#Andron
[16] http://cufon.shoqolate.com

ként jelennek meg, így nem másolhatók. Az eljárás lényege, hogy egy online generátor[17] segítségével a gépünkről feltöltött fontkészletet egy FontForge-szkripttel[18] SVG-fonttá[19] konvertáljuk, ezután az SVG-útvonalakból VML-útvonalakat[20] készít a szkript. Az eredményként kapott fájlt (JSON[21]) majd feltöltjük a weboldalunk szerverére. A JSON-ban tárolt VML-eket egy mellékelt Javascripttel[22] weboldalakba ágyazhatjuk. A szkript a weboldal megadott részeit fogja átalakítani vektorgrafikus elemmé (a HTML 5-ös[23] <canvas>-szá[24]), amelyben a szerverre feltöltött VML alapján rajzolja ki a megfelelő karaktereket.

A korpusz fejlesztésének sarkalatos pontja az annotáció. A különböző szövegek internetes közlése természetesen önmagában is nagy segítség a kutatók számára, ám az igazán értékes, weben is elérhető korpuszok azok, amelyeket gazdag annotációval látnak el. A hozzáadott információ, valamint annak kereshetősége új távlatokat jelent a kutatásban. Napjaink legszélesebb körben használatos általános jelölőnyelve az eXtensible Markup Language (XML)[25], amely tulajdonképpen az SGML egyszerűbb és rugalmasabb változatának tekinthető. A korpuszok annotálását általában XML-kódolással oldják meg, biztosítva ezzel a kompatibilitást, a cserélhetőséget és a hordozhatóságot. Az antikvakorpusz kódolására az egyik legjelentősebb nemzetközi digitális filológiai társaság, a *Text Encoding Initiative* (TEI) XML-kódrendszerét választottam ki. A TEI-t 1987-ben három számítógépes nyelvészeti és irodalmi kutatásokkal foglalkozó angolszász tudományos társaság, az Association for Computers and the Humanities (ACH), az Association for Computational Linguistics (ACL), és az Association for Literary and Linguistic Computing (ALLC) indította el. A projekt feladata irányvonalak kifejlesztése, terjesztése volt a géppel olvasható szövegek kódolására, közvetíthetőségére, és cserélhetőségére, valamint javaslatok tétele új szövegek kódolására. A TEI fejlesztésében mára számos tudományos társaság és tanszék vállal szerepet, évente konferenciákat tartanak, az egyes részterületeket – például a kéziratok kritikai kiadását vagy a karakterkódolást – munkabizottságok vizsgálják. A közös munka eredménye mindig egy DTD, vagyis dokumentumtípus-deklaráció, amely a nyelv jelölőelemeit és egymáshoz való viszonyukat határozza meg. A TEI-t elsősorban általános tartalmú szövegek, szépirodalmi művek, kritikai kiadások, történeti források, illetve élőszöveg-átiratok elektronikus feldolgozására alkalmazzák. Megjelenése óta öt verzióját adták ki, ezek közül a legutóbbi – TEI P5, mely 2007-ben jelent meg – tartalmazza az XML-támogatást is. (Vö. Bíró 2005.)

A korpusz webes megjelenítésére a *Drupal*t választottam, amely egy nyílt forráskódú, PHP-alapú tartalomkezelő rendszer. Moduláris felépítése biztosítja rugalmas bővíthetőségét és testreszabhatóságát, gyakorlatilag bármilyen közösségi tevékenységeket is lehetővé tevő weboldal (ún. web 2.0-s oldal) felépíthető a segítségével. Megoldható vele

[17] http://cufon.shoqolate.com/generate/
[18] http://fontforge.sourceforge.net/scripting-tutorial.html
[19] http://www.w3.org/TR/SVG/
[20] http://www.w3.org/TR/NOTE-VML.html
[21] http://www.json.org/
[22] http://cufon.shoqolate.com/js/cufon-yui.js?v=1.09i
[23] http://www.w3.org/TR/html5/
[24] http://www.w3.org/TR/html5/the-canvas-element.html#the-canvas-element
[25] http://www.w3.org/XML/

a karakterek megjelenítését biztosító cufón, valamint a TEI-XML-fájlok kezelése is. Ez utóbbit az *XML Content* nevű modullal lehet megvalósítani, amely képessé teszi a *Drupal* rendszert XML-fájlok feltöltésére, validálására és megjelenítésére. Szükséges hozzá még a *Content Construction Kit* (CCK) nevű modul is, amelynek segítségével tartalomtípuskén definiálhatjuk az XML-fájlokat a *Drupal* számára.

A fenti technológiák segítségével felépített korpusz olyan közösségi webhellyé válthat, ahol a kutatók egy virtuális kutatókörnyezetben foglalkozhatnak a 16. századi magyar nyelvű nyomtatványokkal. Az együttműködés teheti lehetővé a korpusz további bővítését (a teljes 16. századi magyar nyelvű nyomtatványanyag lehet a végső cél) és fejlesztését (pl. a szövegek összekapcsolását az eredeti dokumentumokról készült jó minőségű képekkel, azaz ún. digitális fakszimilék létrehozását).

Felhasznált irodalom

Bíró Sz. 2005. *Szövegfeldolgozás XML alapokon.* Neumann Kht. http://www.tankonyvtar.hu/informatika/szovegfeldolgozas-xml-080906-159 (letöltve 2011. szeptember 30.)

Kalcsó Gyula: *Magyar Antikvakorpusz.* http://korpusz.ektf.hu (letöltve 2011. szeptember 30.)

Content Construction Kit Drupal-modul. http://drupal.org/project/cck (letöltve 2011. szeptember 30.)

Drupal tartalomkezelő rendszer. http://drupal.org/ (letöltve 2011. szeptember 30.)

Régi magyarországi nyomtatványok 1473–1600 (RMNY). *http://www.arcanum.hu/oszk/lpext.dll/?f=templates&fn=main-h.htm&2.0* (letöltve 2011. szeptember 30.)

Medieval Unicode Font Initiative. http://www.mufi.info/ (letöltve 2011. szeptember 30.)

Text Encoding Initiative. http://www.tei-c.org (letöltve 2011. szeptember 30.)

XML Content Drupal-modul. http://drupal.org/project/xmlcontent (letöltve 2011. szeptember 30.)

Kinnunen, Simo. *Cufón – fonts for the people.* http://cufon.shoqolate.com/ (letöltve: 2011. szeptember 30.)

Unicode standard. http://unicode.org (letöltve: 2011. szeptember 30.)