Kalcsó Gyula
Eszterházy Károly Főiskola
kgyula@ektf.hu

RÉGI MAGYAR SZÖVEGEK NORMALIZÁLÁSI LEHETŐSÉGEI

A Magyar Antikvakorpusz

Előadásom a magyar nyelvű könyvnyomtatás első fél évszázadában megjelent nyomtatványok reprezentatív korpusza, a Magyar Antikvakorpusz[1] fejlesztésének egy fázisát, a normalizált szövegváltozatok előállításának folyamatát mutatja be. A gyűjtemény első verziója 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 eredetekröl és nemzetségekröl 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

[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,

12 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

---

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., 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.

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
- 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.

A normalizálás fogalma

A normalizálás az eredeti betűhű szóalakok egységesítése és mai hangjelölésre konvertálása. Közismert példákkal szemléltetve: *Latiatuc feleym → Látjátok feleim,* vagy: *Vylag uilaga → Világ világa.* A normalizálás elvégzése több okból is szükséges: a helyesírási következetlenségek (sőt: esetlegességek) miatt drámaian visszaesik a gépi feldolgozás hatékonysága; a mai magyarra kidolgozott nyelvtechnológiai eszközök így adaptálhatók a régi szövegekre. Ha találnánk olyan eljárást, amelynek segítségével a rendkívül időigényes, és nagy szakértelmet kívánó manuális átírási munka kiváltható, akkor a szükséges emberi erőforrás alkalmazása leszűkíthető.

A gépi normalizáláshoz voltaképpen a korabeli betűk és betűkapcsolatok mai megfelelőjét kell megkeresnünk. Elvileg lehetséges volna a korpusz összes karakteréhez, valamint karakterbigramjához és –trigramjához manuálisan hozzárendelnünk a mai megfelelőket[12]. A gondot az okozza, hogy egy-egy karakterhez vagy sztringhez több mai megfelelő is hozzárendelhető, valamint ugyanazt a hangot többféle karakterrel vagy sztringgel is jelölik, ráadásul akár ugyanazon nyomtatványban is, következetlenül.

Mivel ez a szövegnormalizáló konverzió analóg több klasszikus nyelvfeldolgozási probléma során jelentkező feladattal, így érdemesnek tűnik az azokban sikerrel alkalmazott módszerek adaptálása és eredményességének vizsgálata. Több ponton is

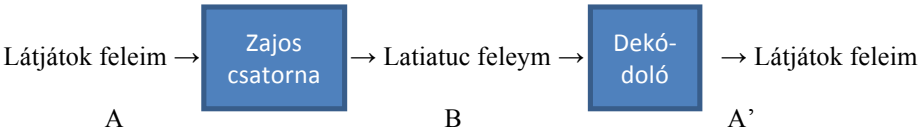
[11] Az OSZK állományában lévő művekről készült fakszimilék teljes listája fellelhető az http://regi.oszk.hu/hun/szakmai/hasonmas/hasonmas_index_hu.htm internetcímen.

[12] Tapasztalatunk szerint a korszak nyomtatványaiban hármas betűkapcsolatnál összetettebb hangjelölés nem található.

rokon feladatra vállalkoztak a Magyar Tudományos Akadémia Nyelvtudományi Intézetének kutatói, akik a Magyar Generatív Történeti Szintaxis[13] nevű projekt keretében felépítettek és normalizáltak egy ómagyar kódexkorpuszt. A normalizálásra nézve l. Oravecz – Sass – Simon (2009). Ők annak eldöntéséhez, hogy a lehetséges átírások közül adott esetben melyik a helyes, egy valószínűségi alapú paradigmát alkalmaztak Shannon zajoscsatorna-modellje (Shannon, 1948) és a Bayes-szabály (Denkinger, 1990) segítségével. Módszerük átdolgozásával sikerült egy viszonylag sikeresnek mondható normalizáló algoritmust létrehozni a korai magyar nyomtatványok normalizálására.

Shannon zajoscsatorna-modellje

Shannon zajocsatorna-modelljét oly módon alkalmazhatjuk, hogy az eredeti, betűhív szöveget (B) úgy tekintjük, mint egy zajos kommunikációs csatornán átment, eltorzított változatot.



A cél egy olyan dekódoló algoritmus megalkotása, amely a torzításokat kiküszöbölve „helyreállítja” a normalizált helyesírású változatot (A’).

A Bayes-szabály

A dekódolás során valószínűségi értékeket alkotunk az ún. Bayes-szabály segítségével. A tétel egy feltételes valószínűség és a fordítottja között állít fel kapcsolatot. Legegyszerűbb formája:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

A dekódoló feladata annak az A karaktersorozatnak a megtalálása, melyre a P(A|B) feltételes valószínűség maximális, vagyis:

A’ = argmax P(A|B), azaz:

A’ = argmax P(B|A)P(A).

Látható, hogy a képletnek két eleme van: a P(B|A) valószínűség, amely az eredeti formák, valamint a zajos csatorna torzulásai közti megfelelések valószínűségét jelenti,

[13] http://www.nytud.hu/oszt/korpusz/mgtsz.html

valamint a P(A), amely mai helyesírású változatokat jelenti. Az előbbit nevezzük csatornamodellnek, az utóbbit forrásmodellnek.

A csatornamodellt úgy állítottam elő, hogy a korpusz karaktereihez, valamint karakterbigramjaihoz és -trigramjaihoz mai betűket és betűkapcsolatokat rendeltem, valamint megadtam a megfelelés valószínűségét a korpuszbeli gyakoriság alapján számítva. Forrásmodellként több rendelkezésre álló, nagy mennyiségű szöveget tartalmazó mai helyesírású korpusz használható (pl. a Szeged Korpusz: http://www.inf.u-szeged.hu/projectdirs/hlt/hu/szegedcorpus%202.0.html).

Az eredmények

Az algoritmus működésének eredményeként a korpusz szóalakjai esetén az adott sztringhez tartozó lehetséges megfelelések valószínűségi értékeit kapjuk meg. Az esetek többségében a legnagyobb valószínűségű megfelelés valóban helytálló, vannak azonban olyan szóalakok, amelyek áthidalhatatlan problémát jelentenek. Például a rövidebb sztringek esetében gyakran több, hasonló valószínűségű megfelelés is lehetséges: *fwl* → *föl*, *fül*. A forrásmodell esetében további problémát jelent, hogy a korai nyomtatványokban vannak azóta kihalt vagy jelentősebben módosult nyelvi egységek (morfémák, lexémák), amelyek esetében kérdéses a valószínűségi értékek helytállósága.

Az algoritmus segítségével normalizált szöveg tehát kézi korrektúrára szorul, de még így is jelentősen csökkenthető a manuálisan elvégzendő munka mennyisége. A korrektúra után a korpusz alkalmassá válik a mai magyarra kifejlesztett nyelvtechnológiai eszközökkel történő elemzésre, az elemzett és annotált szövegek pedig alkalmasak sokoldalú lekérdezések, keresések végrehajtására.

Irodalomjegyzék

Shannon, C. E. 1948. *A Mathematical Theory of Communication.* Bell System Technical Journal, 1948, 27(3): 379–423.

Oravecz Csaba – Sass Bálint – Simon Eszter 2009. *Gépi tanulási módszerek ómagyar kori szövegek normalizálására.* In: Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): A VI. Magyar Számítógépes Nyelvészeti Konferencia előadásai. Szeged: Szegedi Tudományegyetem. 317–324.

Denkinger Géza 1990. *Valószínűségszámítás.* Budapest: Tankönyvkiadó.