# KISEBB KÖZLEMÉNYEK

## A Nagyszótár történeti korpuszának morfológiai elemzéséről*

A cikkben bemutatandó projektum fő célja az Akadémiai Nagyszótár teljes elektronikus korpuszának morfológiai elemzése. (A korpusz jelenleg a XVIII. századtól [kb. 1772-től] 2000-ig tartalmaz különböző témájú, műnemű és műfajú szövegeket, l. a következő honlapon: http://www.nytud.hu/hhc.) Írásunkban részletesen beszámolunk erről a munkálatról és eddigi elért eredményeinkről.

Első lépésben a HUMOR morfológiai elemzőprogramot használtuk (l. PRÓSZÉKY GÁBOR, HUMOR — A Morphological System for Corpus Analysis. In Proceedings of the first TELRI Seminar in Tihany. Research Institute for Linguistics, Bp., 1996. 149—58). Minthogy a HUMOR a mai standardizált helyesíráson alapuló szövegekre készült, így nem vesz figyelembe diakrón szempontokat. Ennélfogva találnunk kellett egy olyan eljárást, amellyel a nem mai helyesírás szerinti szövegeket is tudjuk elemezni és ezen szavak lekérdezésének hatékonyságát is növelni. A cikkben azt fogjuk bemutatni, hogy miként tudtuk átültetni a nyelvi diakróniát egy másik nyelvre, a számítógép nyelvére.

**1.** A morfológiai elemzőprogram. — A HUMOR morfológiai elemzőprogram első verzióját a MorphoLogic Kft. osztályunkkal, az MTA Nyelvtudományi Intézete Lexikológiai

* Köszönetet mondunk Gerstner Károlynak és Mészáros Tamásnak a cikk megírásában nyújtott segítségéért.

és lexikográfiai osztályával együttműködve készítette el (vö.: PAJZS JÚLIA, The Use of a Lemmatized Corpus for Compiling the Dictionary of Hungarian Using Corpora, In: Proceedings of the 7th Annual Conference of the OUP & Centre for the New OED and Text Research. Waterloo, University of Waterloo, 1991. 129—36; PRÓSZÉKY i. m. 1996.; PRÓSZÉKY GÁBOR — KIS BALÁZS, Agglutinative and Other (Highly) Inflectional Languages. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. College Park, Maryland, USA, 1999. 261—8).

A morfológiai elemzőprogram működéséhez szükség van a magyar szókincset reprezentáló összes tőszóra, továbbá a magyar nyelvben előforduló összes felvehető szuffixumra és prefixumra, valamint olyan morfofonológiai és morfotaktikai szabályszerűségek alkalmazására, melyek az agglutináló magyar nyelv komplex morfológiai szabályait képesek leszűkíteni. Ezeket az információkat tartalmazza a HUMOR morfológiai elemzőprogram.

A HUMOR tőszavainak kiinduló halmazát „A magyar nyelv értelmező szótárá"-nak a szavai és a hozzájuk rendelt morfológiai kódok adták. Már itt láthatjuk a diakrón szövegeknél fennálló nehézséget: a XX. század szavainak és morfémáinak halmaza bizonyos tekintetben különbözni fog a XVIII. és XIX. századi szövegek szavaitól. Ezeket a lexémákat a HUMOR az esetek nagy többségében nem tudja elemezni.

**2.** A lekérdezhető elektronikus korpusz. — A jelenlegi lekérdezhető elektronikus korpuszon a HUMOR morfológiai elemző futott le. A szövegszavak századonkénti eloszlása a következő: a XX. századból 16 millió, a XIX. századból 6,8 millió, a XVIII. századból 1,7 millió szövegszó szerepel (szövegszón minden egyes szóelőfordulást értünk). A korpuszt különböző hosszúságú, műfajú és témájú szövegek alkotják, valamint a korpusz tartalmazza még a szövegeket felépítő logikai egységeket, az XML fájlokat. Az XML olyan metanyelv, amelynek segítségével szöveges dokumentumok tartalmi strukturálását végezhetjük el. Az XML szabványban (World Wide Web Consortium, XML Specification 1.0, W3C Recommendation, http://www.w3.org/TR/xml, 1998.) rögzített módon definiált XML alapú nyelvek (XML alkalmazások) segítségével adhatjuk meg a szöveges dokumentumok logikai szerkezetét. A szabvány szerint megjelölt szöveges dokumentumok tartalma számítógép által is értelmezhető. Ezt a megoldást alkalmazva a szövegkorpuszunk (adatbázisunk) több szempontból válik lekérdezhetővé (pl. bibliográfiai adatok lekérdezése, keletkezés dátuma, megjelenési év, műfaj stb.). A korpuszban a szövegeket ASCII-kódokkal rögzítettük (l. www.diffuse.org/chars.html#ASCII). Ez azt jelenti, hogy számok és betűk kombinálásával helyettesíthetjük a régebbi, illetve a mai magyar ábécé ékezetes és más speciális betűit: pl. *e1 = é; u2 = ü; u3 = ű*, *s41 = ſ*.

A korpuszban lévő szövegek illusztrálására bemutatunk két részletet. Egy a HUMOR számára tökéletesen elemezhető szövegrészlet Arany János Arisztophanész-fordítása (korpuszbeli azonosítója: 1900014008, kódolásunkban az első két számjegy jelöli a keletkezés évszázadát):

*Hogy fe1rfi egy se jo2n! re1g ke1ne ma1r. /*
*Itt a1llok, o1n-fehe1rrel vastagon /*
*Bekenve, sa1rga ko2nto2st ve1ve fo2l, /*
*Hia1ba, s holmi csa1b-dalt du1dolok, /*
*Va1rva1n, ha le1pre kaphatne1k vagy egyet. /*
*Mu1zsa1k, elo3 ha1t, ajkaimra mind! /*
*Sugalljatok egy io1ni*
*dallamot.*

Lássunk ugyanakkor egy XVIII. századi szövegrészletet is FÖLDI JÁNOS „Magyar nyelvkönyv, avagy grammatika" (1800112022) című művéből. Ez a szöveg a HUMOR-ral csak részleteiben elemezhető.

*Eggy Magyar Nyelvko2nyv jelenik-meg elo24tted, o2nno2n Nemzeti Magyar o24lto2zete1-ben. Minden ts43avarga1s, minden ta1volro1l valo1 eleibekeru24le1s ne1lku24l, egyedu24l ts43ak ve1ghez vitt Munka1mban elo24ttem tartott okaimro1l, e1s egyenes43en cze1lozott ta1rgyomro1l akarok e1n ez Elo3lja1ro1 besze1dben eggyu24gyu24en, nem fellengo24s43o2n, eggy Magyar egyenes43se1ggel, mine1l ro2videbben jelente1s43t tenni... Magyaru1l i1rva1n a' Nyelvko2nyvet, az azzal ja1ro1 Nevezeteket is Magyaru1l ki-tettem, e1s e' ve1gre eggyne1ha1ny u1j nevezetekkel Nyelvu2nket is gazdagi1tottam.*

A XVIII. századi szövegek ortografikus sajátságait jól tükrözi fentebbi példánk is. Itt egy-egy mással- vagy magánhangzót egy kétjegyű szám is követhet, pl. *s41* (azaz *ſ*) vagy *u24* (azaz *ü*). Itt olyan kódokat használunk, amelyek nem a mai magyar ábécé betűit reprezentálják. A szöveget végigolvasva olyan szóalakokat is találhatunk, amelyek a morfológiai elemzés tőtárában nem szerepelnek, pl. *eggy*. A HUMOR-ral tehát nem tudjuk a morfológiai elemzést elvégezni, ha olyan szám- és betűkombinációra bukkanunk, amely ma már nem létezik, pl. *o23* (azaz *ő*), vagy ha olyan szuffixumot vagy prefixumot találunk, amely ma már nem vagy nem a mai alakban használatos (*magyar-u1l*, lásd a fenti szövegben), vagy ha a szó (tőmorféma) alakilag régies formájú (*nevezetekkel* 'megnevezéssel').

Vizsgáljuk meg ezt a problémát közelebbről: vajon lehet-e valamilyen megfeleltetést találni a régi ortográfiával írt szövegszavak és a mai helyesírású szövegszavak között? Végső célunk az volt, hogy a XVIII., a XIX. és a XX. századi szövegeket egységes formában tudja a program elemezni és ezzel együtt a lekérdezések hatékonyságát növelni.

**3.** A XVIII. és XIX. századi szövegek standardizálásának módszere**.** — Arra törekedtünk, hogy a XVIII. és XIX. századi szövegek sajátosságainak feltárásához olyan szabályokat találjunk, amelyeket a számítógépes program fel tud dolgozni. Az alábbi példákkal szeretnénk bemutatni a feldolgozás módszerét.

A) A régi betűk problémájának feldolgozása. — Kezdjük első lépésben a *s41emmi* szóval. (A példákat ebben a fejezetben ASCII kóddal írtuk, hogy a programozástechnikai folyamatot jól be tudjuk mutatni. Az eredeti szövegben az *s41* jel az *ſ*-nek felel meg.) Az *s41* jelet a mai *s* betűre cseréljük, ekkor a szóalakunk már megfelel a mai helyesírás követelményeinek. Jelöljük ezt a cserét a következő szabállyal: *s41* → *s*. A szabály jelentése a következő: 'ha *s41*-et találunk, akkor cseréljük ki *s*-re'.

Az *u20*-as (eredeti szövegben: *ű*) kódot tartalmazó szavaknál már nem egyértelmű a választás, hogy melyik betűvel helyettesítsük ezt a kódot. Például a *bu20szke* (ma: *büszke*, ASCII kóddal *bu2szke*) vagy *bu20n* (ma: *bűn*, ASCII kóddal *bu3n*) szavak említhetők erre jó példaként. Ebben az esetben a szabályunk így hangzik: 'ha *u20*-at találunk a korpusz valamelyik szóalakjában, akkor azt *u2*-re vagy *u3*-ra kell cserélnünk', azaz *u20* → *u2, u20* → *u3, u20* → *u2, u3*. A magyar nyelv és helyesírás ismerőinek nem nehéz eldönteni, hogy melyik megoldást válasszák a második szabály alkalmazásánál. Nekünk viszont egy a számítógép nyelvére átültethető módszert kellett találnunk arra, hogy a gép elemezni tudja például a *bu20szke* szót. Az elemző nem képes arra, hogy megkeresse a jó eredményre vezető helyettesítést, de képes arra, hogy a morfológiailag jó vagy rossz szóalakot elfogadja, illetve elutasítsa. Ezt a tulajdonságát felhasználva jutottunk el konverterprogramunk alaptéziséhez: állítsuk elő az *u20* esetén azok korpuszbeli alternatív helyettesítéseit.

a) **bu20szke → bu2szke, bu3szke**

A nyíl jobb oldalán álló szóalakok közül a HUMOR a *bu2szke* alakot elemzi, a **bu3szke* alakot pedig nem. Természetesen fontos megtartani az archaikus *bu20szke* szóalakot is. A helyes elemzés mellett ezeket az információkat együtt tároljuk a korpuszban, így a mai szóalakokat keresve azok diakrón megfelelőit is megkaphatjuk. (Tapasztalataink során arra a következtetésre jutottunk, hogy az *u20* karakterkombináció cseréjénél elegendő azt *u2*-re és *u3*-ra cserélni. )

A fenti logikát követve olyan archaikus szóalakok estetében is jól tudjuk használni ezt a módszert, ahol nem a régi ábécé betűinek helyettesítését kell megoldanunk, hanem a mai helyesírástól eltérő szóalakokat kell egy mai szóalaknak megfeleltetnünk. Példaként említhetjük a *mellyen* szóalakot, amely igen gyakran fordul elő ilyen formában a még nem standardizált szövegekben. — A fenti megfeleltetési szabályhoz hasonló módon találhatunk szabályt erre az esetre is:

b) **lly → ly**.

A teljes korpuszon belül gyakori a *lly* hosszú mássalhangzó, illetve betű. Ez a többjegyű betű nemcsak a XVIII. századi ortográfiában, hanem a mai helyesírású szövegekben is előfordul, például *aszállyal*, *akadállyá* stb. Ha erre az esetre alkalmaznánk a *lly → ly* szabályt, akkor egy helyesírási szempontból helytelen szóalakot állítanánk elő: **asza1lyal (=aszályal)*. Így a fenti szabály alkalmazását bővítenünk kell, hogy ne állítsunk elő hibás formákat. Írjuk fel például az *aszállyal* szóalakra azt a szabályt, amely nem rontja el az elemzését! Ez esetben a megfeleltetési szabály a következő:

c) **lly → lly.**

Ezt a szabályt úgy értelmezhetjük, hogy 'helyettesítsük a *lly*-t *lly*-nal', azaz hagyjuk változatlanul. Az *u20 → u2, u3* szabályt két előzetes szabályból vontuk össze. Ugyanezt kell itt is tennünk, ha azt akarjuk, hogy a *mellyen* és az *aszállyal* típusú eseteket egy szabályban írjuk le:

d) **lly → ly, lly.**

Ezzel a szabállyal még nem teljes a *lly* megfeleltetésének problémája, de a bemutatott lépésekkel jól érzékelhető annak a menete, hogy milyen módszert választottunk a nem mai helyesírás szerint írott szövegek standardizálására. Egy-egy szabályszerűség megtalálásakor mindig arra törekedtünk, hogy minél több szempontból vizsgáljuk meg a korpuszt, hogy lehetőleg minden feltárható helyettesítést elvégezhessünk. Ezt számítógépes módszerrel és nyelvészeti ismeretek együttes alkalmazásával értük el. (Az alkalmazott szabályrendszer egy része a lentebbi táblázatban található, vö. **5.** pont.) Egy-egy szóalakon belül olykor több megfeleltetési szabályt is alkalmaznunk kellett.

Nevezzük el az *lly* és *u20*, azaz a cserélendő betűket *mintá*-nak. A program mintákat keres majd a szövegszavakban, s ha talál, akkor a szövegszó mintáját az általunk megadott mintákra cseréli. Ezt az eljárást mutatjuk be a *negyvennyolczbo1l* szón. Ehhez a szóhoz a következő szabályokat alkalmazhatjuk (a szabályokat ezúttal már a teljes korpuszra felírt szabályrendszerből vettük):

1. **nny → nyj-ny-nj-nny,**
2. **cz → c, cz,**
3. **o1l → ol, o1l.**

A jobb oldalon álló helyettesítések számát összeszorozva 4 × 2 × 2 = 16 variánst állíthatunk elő a *negyvennyolczbo1l* alakból.

Az első szabály alkalmazása után keletkező variánsok: *negyvenyjolczbo1l, negyvenyolczbo1l, negyvenyolczbo1l, negyvenjolczbo1l, negyvennyolczbo1l.*

Minden eddigi variánsalakra a második szabály alkalmazása:

| | |
|---|---|
| *negyvenyjolcbo1l* | *negyvenyjolczbo1l* |
| *negyvenyolcbo1l* | *negyvenyolczbo1l* |
| *negyvenjolcbo1l* | *negyvenjolczbo1l* |
| *negyvennyolcbo1l* | *negyvennyolczbo1l* |

Minden eddigi variánsalakra a harmadik szabály alkalmazása:

| | | |
|---|---|---|
| *negyvenyjolcbo1l* | *negyvenjolczbo1l* | *negyvenyjolcbol* |
| *negyvenyolcbo1l* | *negyvennyolczbo1l* | *negyvenyolcbol* |
| *negyvenjolcbo1l* | *negyvenyjolczbol* | *negyvenjolcbol* |
| *negyvennyolcbo1l* | *negyvenyolczbol* | *negyvennyolcbol* |
| *negyvenyjolczbo1l* | *negyvenjolczbol* | |
| *negyvenyolczbo1l* | *negyvennyolczbol* | |

Ezt a módszert követve számos más nyelvi esethez is tudtunk szabályt találni. Hangsúlyozzuk, hogy az általunk felírt megfeleltetési szabályok nem kötődnek szóösszetételi szabályokhoz sem (lásd *negyvennyolczbo1l*)! A program csak mintát keres, és ezért a minták cseréjénél a lehetséges variációkat állítja elő.

**4.** Nyelvészeti ismeretek alkalmazása a szabályok kialakítása során. — Bár a XVI—XVII. századi reformáció és ellenreformáció az ortográfia terén is sok változást hozott, a magyar helyesírás és hangjelölés akkor még nagyon messze volt a mai szabályokba fektetett állapotától. A XVIII. századi szövegeket vizsgálva megállapíthatjuk, hogy a következetesség inkább egy-egy szerző egy-egy művén belül érvényesül, a szélesebb körű, általánosabb szabályok azonban még ekkor is formálódóban voltak.

A megfeleltetési szabályok felírásához megpróbáltunk olyan nyelvi palettát összeállítani, amely képviseli azokat a maitól eltérő hangtani-morfológiai jellegzetességeket, amelyek a XVIII-XIX. századi nyomtatott írásbeliséget jellemzik. Természetesen a kategóriák kialakításába az egyedi esetek nem férhettek bele. Minthogy a régi nyelvezetben voltak még olyan igei és névszói paradigmák, valamint olyan írásváltozatok is, amelyek a mai standardban már nem léteznek, a morfológiai elemző ezeket az alakokat nem tudta elemezni, vagy pedig a mai összetételi szabályoknak megfelelően összetett szóként elemezte őket. Például: a) a HUMOR nem elemezte ezt: *ábrándi* % (a % jelöli azt, hogy a HUMOR nem ismerte föl a szóalakot); b) összetételként elemezte: *zokogásikon* = zokogás [FN] + ikon [FN]; *állék* = áll [FN] + ék [FN]; *kelének* = kel [FN] + ének [FN]; *írátok* = ír [FN] [MN] + átok [FN]; *hajtata* = haj [FN] + tata [FN]; *aggóda* = agg [FN] [MN] + óda [FN]; stb. — Megjegyzendő, hogy jelentések szerint az elemző nem tud keresni, így a szemantikailag nem létező összetett szavakat a gépi elemzésből nem tudjuk kizárni.

Az újonnan kialakított megfeleltetési szabályokat az alábbi típusokba soroltuk: pozicionális, ortográfiai, valamint fonológiai, orto-fonológiai és morfológiai. Ezeket a szabályokat a reprezentatív nyelvi metszet alapján állapítottuk meg.

A) Melyek a pozícióra (azaz az adott betűnek a szavakban levő elhelyezkedésére) jellemző sajátosságok, és milyen szabályokat írhatunk fel rájuk? Agglutináló jellegénél fogva a magyar nyelv ilyenfajta szabályok felírására csak csekély mértékben alkalmas. A szóvégi jelenségeket tekintve az alábbi változtatásokat állítottuk fel:

1. ***tt → t,*** *ábrázoltt → ábrázolt;*
2. ***ü → ű,*** *árnyszerü → árnyszerű*, *aczélhegyü → aczélhegyű;*
3. ***zz → z,*** *apollóhozz → apollóhoz;*
4. az ***n*** határozórag időtartamának változása: ***nn → n****: zajbann → zajban.*

Ezek a szabályok általában a magán-, illetve a mássalhangzók hosszúságára vonatkoznak. A táblázatban **P** jellel szerepelnek. Részint ide tartozhatnak még az olyan morfológiai szabályok, mint a *be, le, ki, meg, fel* igekötők leválasztása a szóvégről. Ehhez járul még az *-ik* rag és az *is* partikula, illetve kötőszó leválasztása is. Ezeket a szabályokat morfológiai szabályoknak hívjuk, bár inkább nevezhetnénk álmorfológiainak is, mivel nem morfológiai kategóriát jelöltünk ki, hanem hangzócsoportot szóvégi pozícióban, amelyre mellesleg morfológiai kritériumok is vonatkoznak. A pozicionális jellegüket is rögzítettük a táblázatban egy **$** jellel, azaz hogy szókezdő vagy szózáró helyzetről van-e szó. A morfológiai változtatásokat **M** jellel láttuk el. — Igekötő levágása a szóvégről: *állíthassonfel → állíthasson fel;* jel levágása: a birtoktöbbesítő jel és birtokos személyjel alternánsai: *-im → -aim/-eim, -jaim/-jeim: áldozatim → áldozataim;* lexéma leválasztása: *is* partikula és kötőszó leválasztása arról a szóról, amelyikre vonatkozik: *állapotbanis → állapotban is.*

B) Az ortográfiai és fonológiai megfeleltetési szabályok már nincsenek pozícióhoz kötve, a szó bármely részén előfordulhatnak. A szabályok ortográfiai és fonológiai jellege az esetek nagy többségében nem választható el élesen egymástól. Így létrehoztunk egy olyan kategóriát, amely az **FO** jelzést kapta, ahol az ortográfiai és fonológiai jegyek együttesen vannak jelen.

Elsősorban olyan jelenségek köré gyűjtöttük a jellemző előfordulásokat, mint az *l, m, n, r* nyújtó hatása magánhangzó után, az összeolvadásos hangok, a palatalizáció, a gemináció: tehát a jelentősebb minőségi és mennyiségi hangtani jelenségekről van itt szó, beleértve a korra jellemző ortográfiai sajátságokat is. Egyelőre azonban nem tudtunk olyan sajátos, egy-egy szerzőre vagy nyelvhasználói csoportra jellemző jelenségeket figyelembe venni, mint az *í*-zés vagy az *ö*-zés. Tehát mindenképpen a korra általánosságban érvényben lévő jelenségekkel kellett foglalkoznunk. Itt a következő alcsoportokat hoztuk létre:

1. magánhangzó időtartamának változása: *it → ít, ol → ól, ös → ős, ól → ol, ón → on: ámit → ámít, alattvalókón → alattvalókon;*

2. magánhangzó zártsági fokának változása: *ül → ől*: *afelül → afelől;*

3. palatalizált mássalhangzók jelölési variánsai: *lly → lyj, ly, lj; tty → tyj, ty, tj; nny → nyj, ny, nj; ggy → gyj; dgy → dj, gy, gyj; dj → dgy; ty → tyj; ly → l, lj: asszonyszemélly → asszonyszemély, beszélly → beszélj; adgy → adj, edgy → egy;*

4. különféle egyéb ortográfiai jelenségek: *cs → ts: hajcs → hajts*.

C) A t i s z t á n o r t o g r á f i a i szabályok közé az alábbi esetek tartoznak:

1. *ſ* cseréje *s*-re: *friſſ → friss;*

2. *o24 (ő )* és a maitól eltérő ékezetek cseréje *árvíztől → árvíztől;*

3. *cz → c: ábéczés → ábécés; tz → c: akátz → akác; ts → cs: áts → ács;*

4. *szsz → ssz: araszszal → arasszal.*

**5.** A s z a b á l y o k a t ö s s z e f o g l a l ó t á b l á z a t é r t é k e l é s e. — A táblázat összeállításakor 1.400.000 szóalakot vizsgáltunk, ebben mai és régi helyesírású szóalakok is találhatók. (Megjegyzések: $ = szövégi pozíció, O = ortográfiai szabály, FO = fonológiai-ortográfiai szabály, M = morfológiai szabály; hatékonyság = (helyes cserék száma / összes csere száma) × 100.)

| Szabály leírása | Szabály kódja | Nyelvtani jellemző | Helyes cserék | Összes csere | Hatékonyság |
|---|---|---|---|---|---|
| *o22 → o2, o3* | <m1> | O | 16211 | 22550 | 71,9 |
| *ly → l, lj, ly* | <m22> | FO | 3089 | 11909 | 25,9 |
| *im → im, aim/eim, jaim/jeim* | <m24> | M | 6659 | 18103 | 36,8 |
| *o1l → ol, o1l* | <m25> | FO | 5023 | 20324 | 24,7 |
| *s43 → s; s41 → s* | <m27> | O | 57375 | 88939 | 64,5 |
| *cs → ts, cs* | <m4> | FO | 1828 | 17790 | 10,2 |
| *ik$ → 0, ik* | <m45> | M | 4556 | 10031 | 45,4 |
| *it → i1t, it* | <m46> | FO | 25054 | 57487 | 43,5 |
| *cz → c, cz* | <m7> | O | 7695 | 13043 | 58,9 |
| *s43zs43z → szsz* | <m10> | O | 798 | 1162 | 68,6 |
| *s43s43 → ss, s* | <m11> | O | 2615 | 3776 | 69,2 |
| *cscs → ccs, cscs* | <m15> | O | 363 | 676 | 53,6 |
| *tyty → tty, tyty* | <m18> | O | 13 | 15 | 86,6 |
| *dgy → dj, gy, gyj, dgy* | <m19> | FO | 784 | 1178 | 66,5 |
| *ty → tyj, ty* | <m20> | FO | 747 | 4340 | 17,2 |
| *o1n → on, o1n* | <m26> | FO | 755 | 3769 | 20 |
| *le$ → 0, le* | <m33> | M | 2054 | 2054 | 100 |
| *fel$ → 0, fel* | <m35> | M | 114 | 114 | 100 |
| *id → id, aid, jaid* | <m36> | M | 200 | 730 | 27,3 |
| *be$ → 0, be* | <m37> | M | 325 | 325 | 100 |

| | | | | | |
|---|---|---|---|---|---|
| *meg$ → 0, meg* | <m42> | M | 239 | 239 | 100 |
| *u2$ → u3, u2* | <m433> | FO | 1955 | 3425 | 57 |
| *tz → c, tz* | <m6> | FO | 2725 | 7149 | 38,1 |

A szabályok alkalmazásának hatékonyságát bemutató lista eredményei azt mutatják, hogy a morfológiai megfeleltetési szabályok közel 100%-os hatékonysággal működnek. Az eredmény azonban félrevezető. Miért? A látszólagos sikernek technikai okai vannak. Az ilyen esetekben a HUMOR olyan morfémákat elemez, amelyek önállóan léteznek a program tőtárában, és ezeket így mindig sikeresen tudja is szegmentálni. Ha az elemzések tényleges hatékonyságát vizsgáljuk, akkor megállapíthatjuk, hogy a legsikeresebb elemzések az ortográfiai megfeleltetési szabályok alkalmazásával történtek. Itt tehát a tisztán ortográfiai helyettesítésekről van szó. Kevésbé sikeresnek mondhatók a fonológiai-ortográfiai megfeleltetési szabályok alapján működő elemzések. Ennek természetesen az az oka, hogy a fonológiai szabályok sokkal körülményesebben definiálhatók — főként ha nem is beszélhetünk nyelvi egységről —, mint a pusztán ortográfiaiak, amelyek mondhatni következetes cseréken alapulnak.

Megvizsgáltuk azt is, hogy a megfeleltetési szabályok közül melyeket milyen gyakorisággal alkalmaztunk. Legtöbbször az *s43*-at *(ſ)* cseréltük *s*-re, ezek a cserék azonos szabályba vannak belefoglalva. Várható volt ez az eredmény, hiszen a XVIII. században a nem egységes helyesírás ellenére rendkívül következesen használták ezeket a betűket. A második helyen az *it → ít* cserék állnak, a harmadik a *ts → cs* csere lett. Hogy a nagy előfordulás mellett miért volt csak 26%-os a szabályok alkalmazásának sikere, annak oka abban rejlik, hogy a *ts* betűkombináció nemcsak a *cs* hangot jelölte, hanem a magyarban a *ts* előfordulhat szóösszetételi határon (pl. *kamatszámítás*), illetve különböző morfémák találkozásánál (pl. *tartsa*) is. A negyedik helyen magánhangzójelek cseréi állnak, mégpedig az *o24 (ő) → ö, ő,* és más régies ékezetű betűk cseréje. Az ötödik helyre egy a XVIII. századra is jellemző fonológiai jelenség került: az *l, m, n, r* hangok magánhangzókat nyújtó hatásáról van szó, melyet írásban is jelöltek.

Az esetek nagy többségében nem pusztán egy cserét végeztünk el egy-egy szón belül: előfordult olyan eset is, amikor hét átalakításra is sor került, de a legvalószínűbb csak egy, illetve két csere volt szavanként.

**6.** Összegzés. — A kísérlet, hogy a történeti korpusz lekérdezéseinek hatékonyságát növeljük, biztató eredménnyel zárult. Azt a célt tűztük magunk elé, hogy a több mint kétszáz éves nyelvi diakróniából adódó nehézségeket a számítógép számára minimalizálni fogjuk. A mai standard nyelvre készült morfológiai elemzőt a hozzátoldott megfeleltetési szabályoknak köszönhetően olyan szövegekre is tudjuk immár alkalmazni, amelyek még jóval a nyelvi-helyesírási egységesülés előtt íródtak. Olyan szabályos eltéréseket kerestünk, amelyekkel a mai nyelvi és helyesírási formájukra tudtuk alakítani a régies szavakat. A teljességre nem is törekedhettünk a feldolgozandó anyag jellege miatt, azonban jelentős eredményeket értünk el az elemezhetőség és ezzel együtt a lekérdezések számszerű bővítésében. További feladatként arra törekszünk, hogy újabb átalakítási, megfeleltetési szabályokat tárjunk fel.

Egy másik lehetőséget kínál a HUMOR morfológiai elemző kivételeket tartalmazó adatbázis-komponense. Itt külön fájlban tüntettük fel azokat a szavakat, amelyekre nehéz transzformációs szabályt találni. Ilyen például a régies *vala* szó, amelynek mai megfelelője a *volt*. Ezen két módszer kombinálásával egy-két hónapos feltárómunkával, melyet lehetőség szerint nyelvészek végeznek el, lehet még javítani a korpusz elemezhetőségén. Meggondolandó azonban az, hogy mennyi energiát érdemes még a munkába fektetni, hiszen megítélésünk szerint ugyanannyi befektetéssel már csak egyre kevesebb esetet tudunk újradefiniálni, felvenni, ugyanis mindinkább közeledünk az egyedi jelenségek felé. Emiatt arra kell törekednünk, hogy a még fel nem tárt általános szabályszerűségeket próbáljuk megtalálni.

KISS GABRIELLA — KISS MARGIT — PAJZS JÚLIA