

Automatikus *ĕ*-jelölő program

A magyar helyesírás nem jelöli a nyílt *e* és az egyes nyelvváltozatokban még élő félzárt *ĕ* fonéma különbségét, mivel az érvényes helyesírási norma kialakításának alapja egy olyan nyelvjárás volt, amelyben az *ĕ* fonéma már nem létezett. A mai budapesti köznyelvben sem szerepel ez a fonéma, és a magyar beszélők többségének nyelvi kompetenciájának nem része a nyílt *e* és a félzárt *ĕ* megkülönböztetése – sem a beszédprodukció, sem a beszédértés szintjén. Ráadásul, mivel az *e*–*ĕ* megkülönböztetés a köznyelvi beszélők számára szubfonemikus, észre sem veszik, hogyha beszédpartnerük *ĕ*-zik, feltéve, hogy az *ĕ*-ző beszélő magánhangzórendszerének egyéb elemei a köznyelvi kiejtési normának megfelelnek.

Az *ĕ*-ző nyelvváltozatokat anyanyelvként használó magyar beszélők egy része ugyanakkor szükségét érzi, hogy ezt a fonémát írásban is megkülönböztesse. Az *ĕ*-k írásbeli jelölését szorgalmazó legismertebb személyiség Kodály Zoltán volt. Jelenleg a félzárt *ĕ* ügyét a Bárczi Géza Kiejtési Alapítvány képviseli, amely *ĕ*-jelölt szövegeket, illetve kiejtési szótárakat és szójegyzékeket ad ki. Ők kérték fel a MorphoLogicot arra, hogy készítsünk egy olyan eszközt számukra, amellyel *ĕ*-jelölést nem tartalmazó szövegeket lényegében automatikusan (egy utólagos félmanuális korrektúra lehetőségével) át lehet alakítani *ĕ*-jelölt szövegekké. Az általunk létrehozott *ĕ*-jelölő eszköznek a segítségével az alapítvány munkatársai a kiadásra szánt szövegeket a korábbinál sok nagyságrenddel kevesebb munka befektetésével tudják elkészíteni.¹

Az eszköz alapját egy olyan szóalaktani leírás képezi, amelyet a sztenderd magyar köznyelv szóalaktani elemző adatbázisának kiegészítésével hoztunk létre. A magyar magánhangzórendszer ismeretében a toldalékok rendszerének megfelelő módosítását a MorphoLogic Humor elemző programjához készített nyelviadatbázis-kezelő keretrendszer segítségével (Novák, 2003 [1]) nem volt nehéz elvégezni, mert az eredeti leírás elég rugalmas formában és ugyanakkor teljesen pontosan írja le a magyar toldalékolási rendszert. A rugalmasságot az biztosította, hogy a leírás alapjául szolgáló nyelvészeti modellben az egyes toldalékok szótári leírása absztrakt mögöttes magánhangzókat tartalmaz. Ezeknek felszíni megvalósulása a hangrendi illeszkedés és a nyitás szabályait követi. A toldalékolás szempontjából az *ĕ*-ző nyelvváltozat pusztán egyetlen absztrakt mögöttes fonéma (a közép-ső nyelvváltozat rövid harmonikus magánhangzó) felszíni realizációja szem-

pontjából különbözik a köznyelvi változattól. Így a teljes toldalékrendszer szótári leírása teljesen megegyezik a két változatban, pusztán a hangrendi illeszkedést leíró szabályok különböznek. Ugyanakkor a tövek *e* hangjainak jelölését, illetve az elől képzett harmóniájú nyitótövek egy részének azonosítását nyelvi kompetencia hiányában nem mi, hanem az alapítvány munkatársai, Buvári Márta és Mészáros András végezték.² A leírást úgy alakítottuk át, hogy jelenleg egyetlen közös adatbázisból egy paraméter átállításával létre lehet hozni *e*-ző és *ë*-ző elemzőt is.

Az *ë* fonémát is tartalmazó kibővített adatbázis alapján készített módosított szóalaktani elemzőprogram képes az *ë*-jelölt szövegek elemzésére, igény esetén készíthető helyesírás-ellenőrző is ehhez a nyelvváltozathoz. Az adatbázis további módosításával hoztuk létre azt az eszközt, amely a szabályos magyar helyesírással írt szövegeket átalakítja olyanná, amelyben jelölve van a két *e* hang közti különbség.

A program többértelmű szavak esetében a legvalószínűbb változatot választja, de a döntése minden egyes többértelmű szó esetében a jobb egér-gomb megnyomására megjelenő kontextusmenü használatával nagyon könnyen felülbíráható. A jelöltek sorrendezése statisztikán, illetve kézzel hangolt jelöltségi sorrendeken alapul. Az alábbi három tényezőt vesszük csökkenő súlyozással figyelembe:

- – *ë*-jelölt szövegkorpuszból nyert szóalak-gyakorisági statisztikát,
- – az egyes tövekhez rendelt jelöltségi sorrendet és
- – az egyes toldalékmorféma-sorozatokhoz rendelt jelöltségi sorrendet.

Az elemzések sorrendezéséhez használt statisztika az elemzőtől függetlenül változtatható, hangolható.¹ A szó környezetét nem vesszük figyelembe a jelöltek rendezésénél, de így is általában nagyon kevés kézi utómunkára van szükség a szöveg végleges formára hozásához. Az utólagos kézi ellenőrzést segíti, hogy a program minden, az *ë*-jelölés szempontjából többértelmű szót megjelöl, és külön jelöli a számára ismeretlen *e* betűt tartalmazó szavakat is. A többértelmű szavak közötti választást a program olyan segédszavak megjelenítésével támogatja, amelyekkel minden olyan magyarul tudó felhasználó is könnyen ellenőrizni tudja, hogy a gép választása az adott kontextusban helyes-e, illetve el tudja végezni az egyértelműsítést, aki az *ë*-ző nyelvváltozatot egyébként nem beszéli:

csënd (főnév) / *csend* (ige)

szemetek (t. sz. főnév) / *szēmētēk* (birtokos személyjeles főnév)

illetékēsēk (főnév) / *illetékēsek* (melléknév)

finnek (olyannak (melléknév)) / *finnēk* (olyanok (melléknév))

Vállalkozásunk példaértékű abból a szempontból, hogy korábban leginkább csak az írott köznyelvi változathoz készültek számítógépes nyelvi eszközök. Programunkat egyébként nem csak *ë*-jelölt szövegek kiadásának előkészítésére lehet használni. Lehetséges további alkalmazása beszélt nyelvi korpuszok annotálásának támogatása, illetve az oktatás. Az *ë*-zö ejtés határozottan visszaszorulóban van. Azok a beszélők, akik az otthoni *ë*-zö környezetből tartósan kikerülnek, gyakran nemcsak egyre inkább leszoknak ennek a nyelvváltozatnak a használatáról, hanem az intuíciójuk is egyre inkább elhalványul, illetve az újonnan tanult szavak (pl. nevek) mentális reprezentációjából már sokszor hiányzik az *ë*-*e* megkülönböztetés. Ezt a folyamatot legalábbis lassíthatja a program használata.

Irodalom

Novák Attila: Milyen a jó Humor? *Magyar Számítógépes Nyelvészeti Konferencia 2003*. Szegedi Tudományegyetem, 138–145.