Novák Attila, Endrédy István (2005): Automatikus ë-jelölő program.
In: III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005). Szegedi Tudományegyetem, 453-454

Automatikus *ë*-jelölő program

Novák Attila és Endrédy István
MorphoLogic Kft. 1126 Budapest Orbánhegyi út 5.,
{novak,endredy}@morphologic.hu

A magyar helyesírás nem jelöli a nyílt *e* és az egyes nyelvváltozatokban még élő félzárt *ë* fonéma különbségét, mivel az érvényes helyesírási norma kialakításának alapja egy olyan nyelvjárás volt, amelyben az *ë* 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 *ë* megkülönböztetése sem a beszédprodukció, sem a beszédértés szintjén.

Az *ë*-ző nyelvváltozatokat anyanyelvként használó magyar beszélők egy része szükségét érzi, hogy ezt a fonémát írásban is megkülönböztesse. Az *ë*-k írásbeli jelölését szorgalmazó legismertebb személyiség Kodály Zoltán volt. Jelenleg is létezik egy alapítvány[1], amely *ë*-jelölt szövegeket, illetve kiejtési szótárakat és szójegyzékeket ad ki. Ők kérték fel a MorphoLogicot arra, hogy készítsünk egy olyan eszközt számukra, amellyel *ë*-jelölést nem tartalmazó szövegeket lényegében automatikusan (egy utólagos félmanuális korrektúra lehetőségével) át lehet alakítani *ë*-jelölt szövegekké.

Az eszköz alapját egy olyan szóalaktani leírás képezi, amelyet a standard magyar köznyelv morfológiaielemző-adatbázisának kiegészítésével hoztunk létre. A magyar magánhangzó-rendszer ismeretében a toldalékok rendszerének megfelelő módosítását MorphoLogic Humor elemzőprogramjához készített nyelviadatbázis-kezelő keretrendszer segítségével (Novák, 2003 [1]) nem volt nehéz elvégezni. Ugyanakkor a tövek *e* hangjainak jelölését, illetve az elöl képzett harmóniájú nyitótövek azonosítását nyelvi kompetencia hiányában nem mi, hanem az alapítvány munkatársai, Buvári Márta és Mészáros András végezték.

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 olyanra, 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 egy a jobb egérgomb megnyomására feltűnő kontextusmenü használatával nagyon könnyen felülbírálható. 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. Kontextuális tényezőket nem veszünk figyelembe a jelöltek rendezésénél, de így is általában nagyon kevés kézi utómunkára van szükség a szöveg végle-

[1] Bárczi Géza Kiejtési Alapítvány

ges 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 segíti, amelyek segítségével 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 nem beszéli:

csënd (főnév) / csend (te azt)
szemetek (főnév) / szëmëtëk (főnév (birtokos alak))
illetékësëk (főnév) / illetékësek (melléknév)
finnek (olyannak (melléknév)) / finnëk (olyanok (melléknév))

Ez a vállalkozás példaértékű abból a szempontból, hogy hasonló módon esetleg más nyelvváltozatok leírására is lesz lehetőség.

Bibliográfia

1. Novák Attila.: Milyen a jó humor? Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), pp. 138–145, Szegedi Tudományegyetem, 2003.