A TEX és a LATEX magyar elválasztási modulja

Mayer Gyula

2002. július 10.

1. A `huwordshy.*` fájlok formátuma

A `huwordshy.*` fájlok magyar szövegekben előforduló szavakat tartalmaznak. Kettős célt szolgálnak: egyrészt TEX-rendszerű elválasztási szabályrendszer generálható belőlük, másrészt egy magyar `ispell` modul ellenőrzéséhez és javításához használhatók föl.

Szavakon a folyó szövegnek szóközökkel vagy újsorokkal határolt egységeit értjük. A szavak belsejében a megengedett elválasztási helyeket általában kötőjelekkel jelöljük.

A formátum kialakításánál figyelemmel kellett lenni a TEX elválasztási algoritmusának korlátaira. Ezért az egyszerűsítve kettőzött többjegyű mássalhangzók esetében a `\ccs{}`, `\ddz{}`, `\ggy{}`, `\lly{}`, `\nny{}`, `\ssz{}`, `\tty{}`, `\zzs{}` illetve `\ddzs{}` kódokat alkalmazzuk. A kötőjeles szóösszetételekben a kötőjel helyett egyenlőségjelet írunk, de minthogy a TEX elválasztási algoritmusát a kötőjel megakasztja[1], ilyen szót egyelőre csak keveset vettünk föl.

Minthogy az idegen eredetű szavak elválasztása a Szabályzat szerint kétféleképpen is lehetséges, elkezdtük különválasztani az ilyen szavakat egy `huwordshy.comp` (comp mint composita) nevű fájlban. Ebben a fájlban az összetétel elemeit `+` jellel kötjük össze, valamint TEX-stílusú megjegyzések is vannak.

A fönti problémát nem mutató, azaz csak egyféle módon (de persze akár több helyen is) elválasztható szavakat a – lényegesen nagyobb – `huwordshy.common` nevű fájl tartalmazza. Esetenként a lehetőség illusztrálására ebben a fájlban is alkalmazzuk a `+` jelet.

2. Megjegyzések néhány elválasztási problémához

2.1. Összetett szavak elválasztása

Szóhatár mellett, azaz szó elején és végén, egyetlen magánhangzó elválasztása a `\lefthyphenmin` és a `\righthyphenmin` paraméterekkel szabályozható, illetve megtiltható. Ezért, függetlenül a tipográfiai kívánalmaktól, megtehettük és meg is tettük, hogy egyetlen betű elválasztásának lehetőségét is föltüntessük (pl. `fi-a-i`). Az összetett szavakban, az összetétel határa mellett ez automatikusan nem tiltható meg, tehát az ilyen szavak esetében a szótagolhatóságtól gyakran eltértünk, és az elválasztást megtiltottuk: `csal-étek`, `fi-zio-te-rá-pi-a` stb., noha a magyar helyesírási szabályzat ([3]) megengedi a `rend-ő-ri`, `vas-u-tas`, `fel-a-dat`, `nagy-a-tá-di` ([3, §236]) típusú elválasztásokat. Ennek a kérdésnek általánosabb kezelését a `+` jel alkalmazása teheti majd lehetővé. Például a `bi-o+e-ner-gi-a` szó `bio-ener-gi-a` alakká lesz az elválasztási szabályok legenerálása előtt.

[1] A – mellékelt – `hypht1.tex` nevű fájlban leírt módon az ilyen szavak elválasztása is megoldható, de ennek ára van.

Mint föntebb említettük, az akadémiai MHSz az idegen eredetű összetett szavak esetében kétféle elválasztást tesz lehetővé: az egyik a szótagolás szerinti (§231), a másik az elemeik szerinti elválasztás (§237). Amennyiben legalább az egyik elem a magyarban önálló szó, akkor „az alakulat összetett voltára általában tekintettel vagyunk”, pl. `Sztálin-grád`, `kilogramm`, `pre-klasszikus`. Ezt a szabály tekintik érvényesnek akkor is, ha az önálló szóalak nem teljesen azonos az összetétel megfelelő elemével, pl. `Jugo-szlávia`, `Petrográd` vö. `szláv`, `Péter`, lásd [6, I. 724] is.

A többi esetre vonatkozóan a szótagolás szerinti elválasztás érvényes, amint az MHSz előző, tizedik kiadása – talán szándékos iróniával – fogalmaz, „így bánunk el természetesen azokkal az idegen szavakkal is, amelyek az átadó nyelvben ugyan összetételek, de az átlagos magyar beszélő nem érezheti összetett voltukat”, pl. `demok-rácia` [2, §326]. Az MHSz jelenleg érvényben lévő kiadásának előkészítése során már érvényre juthatott az a fölismerés, hogy valamiféle elképzelt „átlagos magyar nyelvérzék” alapján nem lehetséges a kérdés egyértelmű szabályozása, ezért – azt mondhatnók, hogy szociolingvisztikai szempontokat érvényesítve – ettől a fiktív mércétől mindkét irányba el lehet térni. Amennyiben az író/szedő/tördelő úgy látja, hogy pl. az `informál` és a `dezinformál` szavak összefüggése „kevésbé nyilvánvaló”, akkor „a szótagolás szerinti elválasztás is helyes”. Másrészről, „az ismertebb elő- és utótagok figyelembevételével történő elválasztás, kivált szaktudományi munkákban, szintén elfogadható”. Ennek a helyzetnek az a következménye, hogy a tördelendő szöveget alkotó és majdan olvasó közeg figyelembevételével esetileg lehet csak eldönteni, hogy melyik típusú elválasztást célszerű előnyben részesíteni, vagyis az elválasztási modulnak testreszabhatónak kell lennie. A szabályzattal kapcsolatban fölmerül viszont az az igény, hogy segítségével egyértelműen eldönthető legyen, melyik az (egyetlen) helyes elválasztás, s ezzel összecseng a TEX alkotójának az a törekvése, hogy a rendszer a lehető legnagyobb mértékben biztosítsa a dokumentumok hordozhatóságát és az adott telepítéstől független megjelenítését. Minthogy az elválasztási szabályokat (`\patterns`) a tördelőprogram – legalábbis a manapság használatos rendszerekben – egy elkülönített folyamat során dolgozza föl, az elválasztási modul testreszabása akadálya az egységes megjelenésnek. Praktikus okokból tehát két szabályrendszert generálunk: az egyik a MHSz keretei között, vagyis a fonetikai alapú előírások figyelembevételével igyekszik a nyelvi igényességet képviselni, s `huhyph-phonet.tex` az elnevezése.

Komolyan fölmerül azonban annak a lehetősége, hogy a szabályzatban kifejtett elveket – a megkezdett úton tovább menve – még általánosabban kellene megfogalmazni. Az „ismertebb elő- és utótagok” kifejezés semmivel sem kézzelfoghatóbb, mint az „átlagos magyar nyelvérzék”, és ugyanúgy kitett mindenféle változásoknak, ezért nehezen lenne védhető a nem eléggé ismert elő- és utótagok bármiféle fölsorolása. Pl. a `foto-gráfia` elválasztás megengedett, azonban a nyelvművelő kéziszótár ellenzi a `biblio-gráfia` elválasztást ([5, 144sk.]; a MHSz átdolgozott szótára nem tér ki e szó elválasztására, vagyis vélhetőleg megengedi mindkét változatot). A szabályzat előírni látszik a `demok-rata`, `prog-ram` elválasztást, de aligha tagadható, hogy a `-krata, pro-, -gram` szóelemek „kivált szaktudományi munkákban” joggal tarthatnak igényt az „ismertebb” minősítésre, vö. arisztokrata, technokrata (technológia, technó), probléma.

Mindezek miatt megteremtettük a számítástechnikai kereteket ahhoz, hogy az idegen eredetű elemekből létrehozott szavakkal kapcsolatban tükröződhessék ezek növekvő ismertsége, a nyelvhasználat tudatosabb volta és az átlagműveltség szintjének folyamatos emelkedése, s a szabályzat szellemében eljárva a műveltebb közönség számára egy a fonetikai szabályokat nem túlértékelő elválasztási szabálygyűjteményt (is) generáltunk. Aligha vitatható, hogy a TEX felhasználói az átlagnál jobban ismerik az idegen szavakat, ezért ezt a szabályrendszert a rövideb `huhyph.tex` névvel láttul el.

Ebben tehát gyakrabban érvényesítjük a szóelemző írásmódot: `ce-phal-algia` (§230), `de-mo-kra-ta` (§231), különösen a latin és a görög eredetű szavak esetében.

(Kevéssé tartanánk indokoltnak ezt az egyéb tulajdonnevek esetében (Michelangelo, Shakespeare)).

Az azonos alakú, de többféle elválasztású szavak esetében (pl. `megint`, `gépelem` §233, `felül`, sugárút) az elválasztást letiltjuk. Az adott szöveghelyen a `\-` parancs segítségével lehet az elválasztást engedélyezni.

2.2. Egyebek

A kettőshangzószerű kapcsolatok esetében (`au-tó`, `au-gusz-tus`, `Eu-rópa` stb.) a két magánhangzó között kerüljük az elválasztást. A szokásos kiejtés alapján nem tartozik e csoportba pl. a `ka-la-uz` és a `fe-udális` szó [4, 86], [8, 364].

Másrészről, a klasszicizáló ejtést követve megengedett az `a-e-ro-szol` elválasztás (MHSz s.v.: `ae-ro-szol`).

Ahol az `ssz` és hasonló betűkapcsolatok *nem* hosszú többjegyű mássalhangzót jelölnek (pl. `kis-szerű`), ott az elválaszthatóságot a szokásos módon jelölhettük. Ez érvényes az összetételhatáron találkozó azonos többjegyű mássalhangzókra is (pl. `jegy-gyű-rű` §7).

Az `András`, `templom` típusú szavak problémájáról vö. [8, 360–2], [9, 348].

3. Néhány nyelvhelyességi kérdés

3.1. Az idegen szavak helyesírása

Az idegen szavak esetében a műfaj jellege és egyéb okok következtében ingadozik a nyelvhasználat a magyarosított és a betűhű írásmód között.

3.2. A toldalékok hangrendi illeszkedése

P. o. bizonyos görög erdetű tulajdonnevek esetében a toldalékok hangrendje ingadozó.

A magyar nyelvben megfigyelhető szabályosságok szerint (pl. [7, 37], [8, 302–3]) mind a Szókratésszal (vagy Sókratésszal), mind pedig a Szokratesszel alak elfogadható. Minthogy azonban a ma érvényes átírási szabályok az utóbbi formát nem engedik meg, ezért az előbbieket vettük csak föl.

3.3. Más ABC-k átírása

A görög tulajdonnevek esetében kétféle átírási rendszer megengedett és használatos (pl. Szókratész vagy Sókratés, Püthagorasz vagy Pythagoras), ezért az adattárban mindkét átírás szerinti formák előfordulnak.

4. Az elválasztási modul generálása

A TEX program (egyelőre) automatikusan és általánosságban nem képes két elválasztási problémát kezelni: az egyik a szavak belsejében lévő kötőjel[2], a másik az egyszerűsítve kettőzött többjegyű mássalhangzók. Ezért ezek tekintetében le kellett butítanunk az alapadattárat: a kötőjelet egyszerű elválaszthatósági helynek tekintjük, az egyszerűsítve kettőzött többjegyű mássalhangzók elválasztását pedig kénytelenek voltunk letiltani. Az ehhez szükséges apró változtatásokat a `huwords2dic.flx` lex-programocska végzi el.

Az elválasztási szabálygyűjteményt a `patgen` programmal generáljuk. Minthogy forráskódján kis változtatást kell végrehajtani, ezt is föl kell installálni. A program része a `web2c`

[2] De v. ö. az előző lábjegyzetet.

disztribúciónak, melynek készrecsomagolt változata a teTeX csomagegyüttesben kényelmesen hozzáférhető. (Mi a fejlesztéshez a linuxos verziót használjuk, MSWindows alatt a MikTeX vagy az fptex rendszer javasolható, amennyiben megvan vagy elkészül a patgen adaptálása is.) A web2c-7.3.3/texk/web2c könyvtárban lévő patgen.web forrásfájlban meg kell emelni két paraméter értékét:

```
@!trie_size=99000; {space for pattern trie}
@!triec_size=54000; {space for pattern count trie, must be less than
```

Erre a legegyszerűbb módszer, ha a patgen.ch fájlt felülírjuk a jelen disztribúcióban szereplővel. Ezután a jelzett alkönyvtárban a make patgen paranccsal generáljuk le a programot. (Az általunk használt linuxos végrehajtható patgen programot mellékeljük, azonban nem biztos, hogy másképp telepített rendszeren működik, és óvatosságból amúgy sem javasoljuk használatát.)

A patgen program használatáról rövid tájékoztató a patgen2tutor.tex fájl, melynek kissé frissített verzióját mellékeljük. (E fájl egy olyan verzióját írja le a programnak, melynél a paraméterek kiosztása eltér a lent leírtaktól.) A részletes dokumentáció a patgen.web és a patgen.ch fájlban olvasható, és a weave program segítségével szépen kinyomtatható TEX-állomány generálható belőlük.

A patgen programot négy paraméterrel kell indítani, pl. a következőképp:

```
patgen hu.dic nulla.pat magyar.out magyar.tra
```

Az első paraméter a szótárfájl (dictionary), mely a helyesen elválasztott szavakat kell tartalmazza. A második paraméter szabályok (patternek) előre megadható készletét tartalmazó fájl; üres fájlt használunk, mivel ilyeneket véleményünk szerint nem érdemes megadni. A harmadik paraméter a legenerálandó patterneket tartalmazó kimeneti fájl neve. A negyedik paraméter annak a fájlnak a neve, mely lényegében az illető nyelv ABC-jének leírását tartalmazza. Belevettük az umlautos német ät is, hogy pl. a Händel név is elválasztható legyen.

A patterngenerálás folyamata finoman szabályozható online megadandó további paraméterekkel. Ezek általunk javasolt értékét egy ideiglenes fájlból olvassuk be. Egyszerűsítésül a generatehuhyph.sh szkript mindezt megteszi, valamint ellenőrzi is rögtön a legenerált patterneket.

A finomhangoló paraméterek megválasztásánál a nyers erő elvét alkalmaztuk, ezért a generálás folyamata meglehetősen időigényes: egy P4-es processzort majdnem nyolc percig igénybe vesz 1 GHz-en.

Meg lenne a lehetőség arra, hogy a szótagolás szerinti elválasztás szabályait mint előre betápláljuk a patgen programba, azonban ennek az lenne a hátránya, hogy kevésbé optimális lesz a legenerált pattern-halmaz, és a szótár bővülése során előbb kifutnánk a lehetséges kilenc elválasztási szintből, s nem lehetne hibátlanul elválasztani a szótár szavait. A szótagolás szerinti szabályokat is a szótár útján igyekszünk érvényesíteni, ezért némelykor idegen szavak is szerepelnek benne, pl. Entwicklung. A szabályzat szerint, ha magyar szövegben egy-egy idegen szó található, a magyar elválasztás szabályai szerint kell elválasztani (§230) és csak a többszavas kifejezések esetében kell az illető nyelv szabályait alkalmazni.

Minthogy [3, §231] kétféle elválasztást enged meg, ezért egy huhyph-phonet.tex nevű szabálygyűjtemény is készül, mely kétes esetekben nem az összetétel határán, hanem a hangtani szabályok szerint választ el.

5. Lokális TEX-fa használata

Gyakran fölmerülő probléma, hogy a felhasználónak nincsen root jogosultsága, vagy pedig csak kísérletezni akar, és nem szeretné elrontani a TEX telepítését.

Ilyen esetekben lokális /texmf/ fa használható. Hogy valamely adott fájltípus esetében mi a keresési sorrend, meglehetősen bonyolult algoritmus szabályozza (szerepet játszik

a végrahajtható állományok fordításakori alapbeállítás, környezeti változók és konfigurációs fájl). Az éppen érvényes beállítás a `kpsepath fmt` és hasonló parancsokkal kérdezhető le. Az elválasztási szabálygyűjtemény a fmt fájlokba statikusan beépítésre kerül, ezért problémát okozhat, ha a mindenkinek szóló pl. `/var/lib/texmf/web2c/` könyvtár prioritást élvez a lokálissal szemben. Ekkor a prioritási sorrendet kell megváltoztatni vagy más elnevezést adni a lokális fmt-fájljainknak. (Legegyszerűbb persze, ha a rendszergazdát meg tudjuk kérni, hogy törölje ki a fmt-fájlokat.)

6. Az elválasztási modul installálása

Miután a TEX-rendszert a szokásos módon telepítettük, keressük meg a `huhyph.tex` fájlt (unix/linux: `locate huhyph.tex`)! Egy szokásos helye lehet az `/usr/share/texmf/tex/generic/hyphen/` könyvtár. Ha nem lokális (per user) tex-fát használ, akkor a következőket superuserként kell végrehajtani. A régi fájlt óvatosságból nevezze át (pl. `huhyph3.tex`-hé) és utána az újat másolja a helyére! Ha korábban nem tette volna meg, a `language.dat` fájlban annak a sornak az elejéről, mely a `magyar` szót tartalmazza, törölje ki a százalékjelet. (Amennyiben többféle magyar elválasztást is telepíteni kívánunk, akkor új nyelvnevet kell kitalálnunk. A fonetikus elválasztás esetére a `magyarf` elnevezést javasoljuk. Ha kizárólag a fonetikus elválaszást használja, akkor a babel csomaggal való együttműködés úgy valósítható meg a legegyszerűbben, ha meghagyja a magyar nyelvnevet.) Ennek a fájlnak egyik szokásos helye a `/etc/texmf/` könyvtár, avagy a `/var/lib/texmf/tex/generic/config/` alkönyvtár,

Ezután a format-fájl(oka)t generálja le a szokásos módon. Unix/linux esetén pl. ennek módja, hogy a `texmf`-fa `web2c` alkönyvtárában kiadjuk az

```
...texmf/web2c> initex latex.ltx
```

és hasonló parancso(ka)t.

7. A telepítés ellenőrzése

A ellenőrzés első lépéseként célszerű ennek a dokumentumnak a forrását, azaz a `huhydoc.tex` fájlt letexhelni:

```
...>latex huhydok
```

Amennyiben a logban a `Babel <v3.7h> and hyphenation patterns for french, german, ... magyar, ... nohyphenation, loaded.` típusú üzenetben szerepel a `magyar` szó, akkor az installálás vélhetően sikeres volt.

Második lépésként alkalmazhatjuk a `\showhyphens{...}` parancsot, pl.

```
\showhyphens{kialakítás Händel Dessewffy folyamatábra szkript}
```

Ha a logfájlban e szavakban a megengedett elválasztási helyeken látunk kötőjelet, a telepítés minden bizonnyal sikeres volt.

8. Az elválasztás vezérlése

Mint a jelen dokumentum forrása is mutatja, a magyar elválasztás beállítására a legegyszerűbb módszer a `\usepackage[magyar]{babel}` parancs használata. Az opcionális paraméter vesszőkkel elválasztva több nyelv megjelölését is tartalmazhatja, és az utolsóként megadott lesz az alapértelmezésű nyelv.

A dokumentum belsejében a `\selectlanguage{nyelv}` utasítással válthatunk a nyelvek és elválasztási modulok között.

9. Az elválasztási modul ellenőrzése

A teljes körű ellenőrzés a `check` alkönyvtárban, a `checkhuhyph.sh` szkript segítségével végezhető el.

10. Felhasználási engedély

A magyar elválasztási modul jelen verzióját a TypoTEX Kft. készítette (©Budapest, 2001–2002) és a Gnu GPL (General Public Licence) szabályai szerint szabadon fölhasználható. Kérjük, hogy a modullal kapcsolatos észrevételeiket a `gam@cs.elte.hu` címre írják meg.

A magyar elválasztási modul elkészítését a Széchenyi Terv keretében az Informatikai Kormánybiztosság SZT-IS-10/3 pályázati számon támogatta.

Hivatkozások

[1] É. KISS Katalin–KIEFER Ferenc–SIPTÁR Péter: Új magyar nyelvtan, Budapest, 1998;

[2] M. T. AKADÉMIA, szerk.: A magyar helyesírás szabályai. 10. kiadás, Budapest, 1954, 1973[13];

[3] M. T. AKADÉMIA, szerk.: A magyar helyesírás szabályai. 11. kiadás, Budapest, 1984, 1994[11];

[4] DURAND Jacques–SIPTÁR Péter: Bevezetés a fonológiába, Budapest, 1997[98];

[5] GRÉTSY László–KEMÉNY Gábor, szerk.: Nyelvművelő kéziszótár, Auktor Kiadó, Budapest, 1996;

[6] GRÉTSY László–KOMLOVSZKY Miklós, szerk.: Nyelvművelő kézikönyv. I–II, Akadémiai Kiadó, Budapest, 1980–85;

[7] RÁCZ Endre–TAKÁCS Etel: Kis magyar nyelvtan, Budapest, 1978;

[8] SIPTÁR Péter: Hangtan, In: *Új magyar nyelvtan* [1], pp. 291–390;

[9] VARGA László: Péter Siptár and Miklós Törkenczy, The Phonology of Hungarian [bírálat], Magyar Nyelv 117/3 (2001) 345–8;

Tartalomjegyzék