A Szószablya fejlesztés

Németh László

2003. október

Kivonat

A Szószablya fejlesztés célja, hogy a magyar weboldalak szövegtartalma alapján magyar szógyakorisági-szótárat, majd ennek felhasználásával nyitott forráskódú morfológiai elemző programot készítsen.

A szótövezésre képes elemző megkönnyíti, illetve minden korábbi eszköznél jobban lehetővé teszi az internetes, vagy intranetes weboldalak, dokumentumok indexelését, ellenőrzését, egyéb automatikus kategorizálását, nem véletlen, hogy az ilyen célra történő vállalati felhasználása már elkezdődött.

A Szószablya fejlesztés alapjául szolgáló; kibővített, javított Magyar Ispell helyesírási szótár és Hunspell helyesírás-ellenőrző pedig a magyar OpenOffice.org révén hozzájárul a nyitott forráskódú programok hazai terjedéséhez. A magyar morfológia elemző pedig az informatika jövőt képviselő területein (gépi fordítás, mesterséges intelligencia) nyújt majd pótolhatatlan segítséget.

1. A Szószablya fejlesztés

A Szószablya fejlesztés (`http://www.szoszablya.hu/`) 2003 márciusában indult a BME és a Matáv által létrehozott, a BME Szociológia és Kommunikáció Tanszék keretén belül működő Média Oktatási és Kutató Központ (MOKK) vezetésével.

Az egy évig tartó fejlesztés során alapvető, nyitott forráskódú (LGPL és GPL licencű) informatikai eszközök készülnek el a magyar nyelvhez. Ezek közül a legfontosabbak a *Hunspell* helyesírás-ellenőrző, a *Hunstem* szótövező, és a *Hunmorph* morfológiai elemző programkönyvtár és program. Elkészül a magyar weboldalak szövegtartalma alapján egy több millió szót tartalmazó *gyakorisági szótár* is. Elérhetővé válnak a a weboldalak feldolgozása során használt segédprogramok (*Hunnorm* szövegátalakító és -normalizáló, *Huntoken* mondatra és szóra bontó) és a *Hunp* nyelvfelismerő.

A következőkben azoknak az eredményeknek az ismertetésére kerül sor, amelyek már most is elérhetők, szabadon felhasználhatók, és (különösen az indexelésre használható tövező esetében) sok Linux fejlesztőt érintenek.

2. Hunspell helyesírás-ellenőrző

A Szószablya fejlesztésnek köszönhetően a Hunspell helyesírás-ellenőrző a kezelt szabályok és szókincs tekintetében is sokat lépett előre.

Csaknem félmillió válogatott weboldal szókincsét tartalmazza az a 4 millió szavas Szószablya tesztszótár, amivel a helyesírás-ellenőrző szókincsének bővítését, és hibamentességét is biztosítani lehet. A szókincs minőségét a legfrissebb szótárakkal

(Magyar helyesírási szótár és az új Magyar értelmező kéziszótár) való összevetés is szavatolja.

A Hunspell szótárának 0.96-os (2003. február) és a 0.99.4-es (2003. szeptember) változata esetében megközelítőleg 99,4%-ról 99,8%-ra sikerült növelni a program szövegellenőrzési képességét, vagyis míg korábban 1000 átlagos szövegszóból több, mint 6 helyes szót nem ismert fel az ellenőrző, most ez a szám körülbelül kettőre, vagyis harmadára csökkent.[1]

A Hunspell dokumentáció [10] részletezi az ellenőrző magyar helyesírási szabályoknak való megfelelését, összevetve a legelterjedtebb magyar helyesírás-ellenőrzővel is. Az MS Office XP-ben található Helyes-e? helyesírás-ellenőrző a következő nagyobb hiányosságokat mutatja a Hunspell-lel szemben:

1. Alapvető helyesírási szabályokat nem ismer. Helyesnek veszi a következő típushibákat: **lássd*, **írdd*, **fessd*, **késsd*; **mögémegy*, **eléväg*, **föléüt*, **teliönt*; **melegvíz*, **szépvers*, **magyarállam*; **fázósgyerek*, **egércincogott*, **olymacska*; **észszerű*, **mészszerű*, **viaszszerű*; **feddd*, **pufffélé*, **frissség*, **fessség*, **izzzad*, **továbbbotorkál*, **játsszza*; **Einsteinféle*, **Budapestszerű*.
2. A magyar nyelvtan kezelésében nem pontos, ezért elfogadja a következő típushibákat: **kézéé*, **vízéé* **tűzéé*; **kértet*, **nyitat*, **oktatatás*; **hölgyök*, **tölgyök*, **szüzök*; **hatalomi*, **birodalomi*; **meghalja*, **történed*, **évődi*; **kézség*, **jelentősség*, **tartóság*; **újnyi*, **nyulnyi*, **lovnyi*; **bokré*, **terhé*; **tesszél*, **visszétek*, **essz*; **tettról*, **hitthoz*, **vészja*; **fiája*, **híjája*, **őfenségéje*; illetve elutasítja: *Józsefné*, *Pálné*; *Béláék*, *unokámék*; *zsenijei*, *tepsijei*; *1.-t*, *11.-et*.
3. Egyéb súlyos helyesírási hibákat is elfogad (például **estéji*, **karvaj*, **szeméji*, **ünnepéjes*; **csevely*, **bolyár*, **bolytár*, **súlytó*; **elitélt*, **hivő*, **izület*, **neutrinó*, **egyivású*, **tikfa*, **hivat*, **irogat*; **fűzek*, **színtű*, **bíztat*, **kompatíbilis*; **hiu*, **savanyu*, **ramazuri*, **harangbugás*; **tinórú*, **kultúrált*, **kultúrálatlan*, **kultúrálódik*; **egyűvé*, **szűntető*, **tűntető*, **kűzdőtér*; **diagramm*, **vállfaj*, **zsupptető*, **mennyhal*; **analfabétizmus*, **polémikus*, **siserehad*, **bédekker*), sőt néhol ezzel együtt a helyes alakokat el is utasítja (például *mésszerű*, *viasszerű*; *csevej*, *puzón*; *Einstein-féle*; *trabantos*; *Rio Janeiró-i*, *New York-iakat*; *camembert-rel*).
4. Szókincse új szavakat nem tartalmaz. Nem ismeri például: *EU*, *ombudsman*, *klónozás*, *valóságshow*, *szafari*, *sztárol*, *kapucsínó*, *nonprofit*, *globalizáció*.
5. Szókincse a gyakori szavak tekintetében is hiányos: *karalábé*, *megnövekedett*, *mondotta*, *közzétesszük*, *legősibb*; *ebtartás*, *ebtenyésztés*; *hőszabályzó*, *hőháztartás*; *síkesztyű*, *síbaleset*; *sólepárlás*, *sómennyiség*; *ízérzékelés*, *ízfokozó*; *hókotrás*, *műhó*; *eperlé*, *léalma*; *aggaszt*, *apaszt*, *áraszt*, *ébreszt*, *fagyaszt*, *függeszt*, *horgaszt*, *kepeszt*, *lehiggaszt*, *sorvaszt*, *szállaszt* stb.
6. Szókincse nem terjed ki a magyar településnevekre (még a városokéra sem: *Gyomaendrőd*, *Szécsény*, *Vásárosnamény*, *Zalaszentgrót* stb.).
7. Szókincse nem terjed ki a szakszavakra. Hiányzik például: (matematika) *elempár*, *variancia*, *alterek*, *bijektív*, *attraktor*, *Cauchy*; (informatika) *alhálózat*, *élkiemelés*, *mikrocsip*, *flopi*, *Linux*; (biológia) *póc*, *domolykó*, *csilló*, *füzény* stb.

[1] Kevesebb, mint 4% helyesírási hibát (beleérve ebbe az idegen és ismeretlen szavakat is) tartalmazó weboldalak szókincsére vonatkozik ez a becslés. Ez magában foglalja a magyar nyelvű napilapokat és a szépirodalmi műveket is.

8. Elavult ékezet nélküli formában, vagy egyáltalán nem tartalmazza az idegen ékezetes szavakat. Hiányzik például: *Händel*, *Dvořák*, *Škoda*, *złoty* stb.
9. Értelmetlen szavakat tartalmaz. Például: **veé*, **sepr*, **sodr*, **éssz*, **mis*, **fesz*, **süv*, **lépt*, **tel* stb.
10. Javítási képességei a tipikus hibák csak csekély részére terjednek ki, szemben a Hunspellel, ami minden egy karakter távolságra lévő hibát, valamint az összes tipikus több karaktert érintő hibát javítja.
11. Zárt program. Alapszókincse a felhasználók által nem bővíthető, szemben a nyitott forráskódú Hunspell-lel, aminek parancssori változata az olyan tőszavak ellenőrzés közbeni felvételét is támogatja, amelyeket az ellenőrző az alapszavakhoz hasonlóan képes toldalékolt alakban is felismerni.

Az összehasonlítás alapján határozottan javasolható az MS Office lecserélése a Hunspell programkönyvtárát és helyesírási szótárát tartalmazó OpenOffice.org-ra [3].

3. Hunstem szótövező

A Hunstem szótövező programkönyvtár és alkalmazás a Hunspell forráskódjának és helyesírási szótárának bővítésével készül. Első komolyabb változata a valós nagyvállalati igényeknek megfelelően fel van készítve a szálkezelésre, és képes ismeretlen szavak feltételezett töveinek megállapítására, ami elsősorban tulajdonnevek és szakszavak tövezése esetében hasznos. A tövezés sebessége átlagos szövegen, AMD XP 1800+-os gépen 40 000 szó/másodperc.

A tövező programhoz elkészített mintaalkalmazás, illetve CGI program példát mutat webhelyek indexelésére, és helyi dokumentumok keresésére. A tövezésnek köszönhetően az indexállomány mérete kisebbre választható, valamint a módosuló tövet tartalmazó szóalakok kezelése sem okoz többé gondot (macska/macskát, bokor/bokrot, híd/hidat stb.). Egy komolyabb webhely esetén esetén a Hunstem használata kikerülhetetlen, de természetesen a tövező felhasználási területe nem korlátozódik az intra- és internetes információ-visszakeresés támogatására.

4. A fejlesztés tapasztalatai

4.1. Hardver- és szoftverkörnyezet

A magyar weboldalak feldolgozásáért felelős központi gép Debian GNU/Linux operációs rendszert futtat. 1,5 TB-os XFS fájlrendszerű háttértárolóval, 2 GB memóriával, 2 darab 1,8 MHz-es Intel Xeon processzorral rendelkezik.

A Szószablya fejlesztés során nyitott forráskódú programok kerültek felhasználásra. A magyar weboldalak a Larbin webrobottal lettek összegyűjtve [6]. A Hunnorm, illetve Huntoken feldolgozó szűrősor C-ben, illetve Flexben készült. A hatékonyság növelése céljából a szűrősor első tagja egy állományba kötegeli a feldolgozni kívánt (több millió) weboldalt, ami a szűrősor végén igény szerint újra állományokra bontható.

A gyakorisági listák a Glibc programkönyvtár tsearch() bináris fát kezelő függvénycsaládja segítségével készültek. Egyéb feladatokat pedig legtöbbször GNU segédprogramokkal meg lehetett oldani. Kisebb gondot okozott, míg ki nem derült, hogy a fejlesztői gép alapértelmezett awkja a mawk, ami milliós nagyságrendű adatok hasító-táblás kezelésében lényegesen rosszabb teljesítményt nyújtott, mint a GNU awk.[2]

A Hunspell helyesírás-ellenőrző a BSD licenccel rendelkező Myspell függvény-könyvtáron alapul, ami eredetileg az OpenOffice.org része. A Myspell az Ispell program működése alapján készült. Az Ispell (és így a Myspell) működése egy olyan mintafelismerő algoritmuson alapul, ami egy orosz nyelvű folyóiratban jelent meg először, és a folyóirat angol nyelvű kiadásával került be a köztudatba még a hatvanas években [7]. Az orosz nyelvű cikk írója Dömölki Bálint matematikus, aki meghatározó szerepet játszott a legendás magyar M–3-as számítógép megépítésében. A Dömölki-algoritmus az M–3-on folyó magyar nyelvészeti kutatások során született meg, közelebbről magyar költők műveinek pszicholingvisztikai, fonetikai elemzésekor [8]. A Dömölki-algoritmust használó Hunspell most a mai magyar köznyelvi szókincset tartalmazó szótár elkészítését teszi lehetővé.

4.2. A magyar web és feldolgozása

A letölthető magyar web durva becslés alapján 20–25 millió oldalból áll, amelynek jelentős része szövegtartalom nélküli, idegen nyelvű, vagy duplikált oldal. A weboldalak kódolása, nyelve, helyesírása különböző. Egy előzetes, a .hu tartomány alól letöltött 2,4 millió weboldalon alapuló vizsgálat érdekesebb számai a következőek voltak: Az oldalak mindössze 0,46%-a volt Unicode kódolású. A letöltött oldalak 6,5%-a nem tartalmazott egy szót sem, további 21% nem tartalmazott magyar ékezetes betűt. Átlagban a nem üres oldalak 347 szót tartalmaznak, 85%-uk legalább 20 szóból állt.

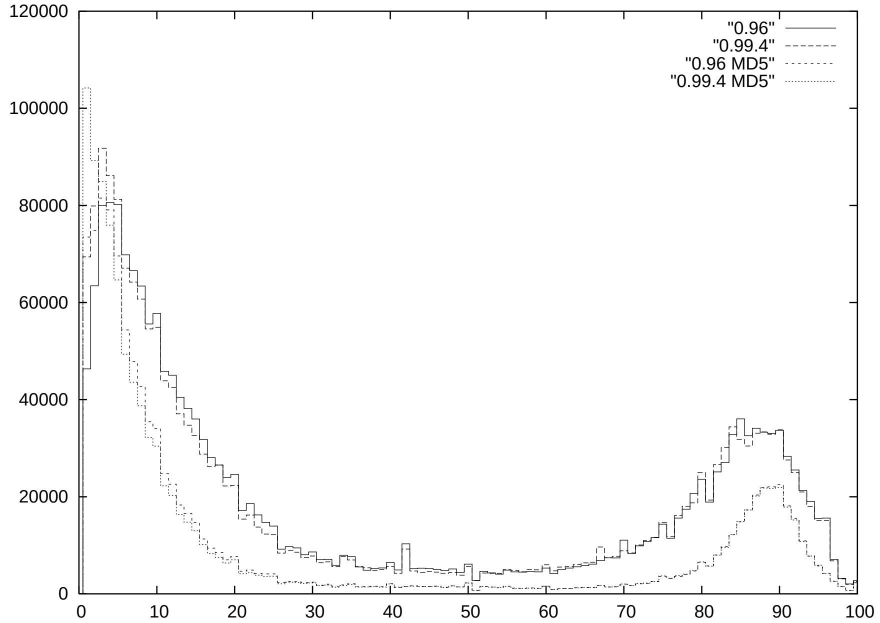
A magyar köznyelvi szókincs összegyűjtése céljából oldalszintű szűrésre került sor a Hunspell helyesírás-ellenőrző segítségével. Az oldalszintű szűrés hátterében az a feltételezés állt, hogy az oldalak szókincse konzisztens, és így a megfelelő (esetünkben az ellenőrző által biztosított köznyelvi) szókinccsel a jó oldalak kiválaszthatók [11].

A következő hisztogram a legalább 20 szövegszót tartalmazó weboldalak gyakoriságát ábrázolja a százalékos hibás szóarány függvényében.

[2] Az awk több helyen fel lett használva a Szószablya fejlesztésben: duplikátumszűrés, trigram adatok számítása, küszöb szerinti vágás és gyakorisági osztályok kiszámítása a később Gnuplot programmal megrajzolt diagramokhoz stb. Például a következő – gyakoriságok összesítéséhez használható – kis héjprogram a bemenet adott (alapértelmezésként első) számoszlopát összeadja, a bc programnak köszönhetően akár több milliárd számjegy pontosságig:

```
#!/bin/sh
awk '{ print "a+=" $'${1-1}' }
END {print "print a"} ' | bc
echo
```

A Unix, Linux és egyéb technológiákról, illetve történeti vonatkozásaikról figyelemre méltó áttekintést nyújt [9].



Az ábra két csúcsot tartalmaz. A bal oldali csúcs a jó helyesírású oldalaknak felel meg, a jobb oldali pedig a jelentős részében idegen nyelvű oldalaknak. Az MD5-tel jelölt görbék egy olyan oldalszűrés eredményét mutatják, ahol csak a pontra végződő mondatok lettek meghagyva a letöltött oldalak szövegében (illetve a kérdő, és felkiáltó mondatok, ha volt mellettük pontra végződő mondat), majd ezután lett alkalmazva rajtuk egy MD–5 hasítókódot használó duplikátumszűrés. A cél elsősorban az automatikusan hozzácsatolt címek (menüpontok, hírek) okozta minőségromlás kiküszöbölése volt.[3] Ez a módszer azonban általában véve növelte a tényleges összefüggő szövegtartalom súlyát.

A módszer eredményességét jól mutatja a kevésbé jó oldalak számának nagy mértékű csökkenése, a duplikátumok okozta kiugró kis csúcsok eltűnése, illetve a jó oldalak számának növekedése. A fenti adatok tükrében, standard szövegek átlagos hibaaránya alapján kiválasztható a magyar web azon része, ami a mai magyar köznyelvi szókincset tartalmazza. Természetesen nem kapunk hibátlan szókincset így sem, de arányában sokkal több helyes szót tartalmaznak ezek az oldalak. Jellemző, hogy az oldalszintű szűrésnek köszönhetően határozott javulás figyelhető meg a tipikus tévesztések arányában is: például míg a letöltött oldalak a szerviz szót az esetek 27%-ban hibásan (szervíz) tartalmazzák, ez az arány a minőség alapján kiválasztott oldalakon már csak 9%. A módszerekről és az eredményekről részletesen a Szószablya fejlesztés honlapja számol be.

[3] Ez legalább két nagy hibáért volt felelős: (1) duplikálás, amikor is a más tartománynéven, de ugyanazon helyről letöltött oldalak a különböző időpont miatt eltérő automatikus címeket tartalmaztak, bár tényleges szövegtartalmuk ugyanaz maradt; (2) szavak hamis gyakorisága, mivel az automatikus címekben szereplő szavak a tényleges gyakoriságuknál jóval nagyobb számban jelentek meg, például egy nagy szolgáltató minden egyes letöltött oldalán olvashatóak voltak.

5. Támogatók

A Szószablya fejlesztést az Oktatási Minisztérium és az Informatikai és Hírközlési Minisztérium ITEM pályázata, valamint a MATÁV támogatása tette lehetővé. A Szószablya alapját képező Magyar Ispell és Magyar Myspell fejlesztés kiemelt támogatói a Szószablya fejlesztés mellett az UHU-Linux Kft. és az IMEDIA Kft. A Szószablya fejlesztést megelőzően a TypoTEX könyvkiadó, Mátó Péter és a SuSE Linux Kft. nyújtott még anyagi támogatást a nyitott forráskódú magyar helyesírás-ellenőrző készítéséhez. Sokan hozzájárultak munkájukkal is a fejlesztéshez. Az együttműködők nevét a Hunspell dokumentáció, illetve forráskód sorolja fel. Külön köszönöm a Szószablya fejlesztés résztvevőinek a cikk megírásához adott számos tanácsot.

Hivatkozások

[1] A Szószablya fejlesztés honlapja: `http://www.szoszablya.hu/`

[2] A Magyar Ispell fejlesztés honlapja: `http://magyarispell.sourceforge.net/`

[3] Magyar OpenOffice.org: `http://office.fsf.hu/`

[4] UHU-Linux: `http://www.uhulinux.hu/`

[5] IMEDIA Kft: `http://www.imedia.hu/`

[6] Larbin webrobot: `http://larbin.sourceforge.net/`

[7] Domolki, B., *Algorithms for the recognition of properties of sequences of symbols.* Журнал Вычислительной Математики и математической Физики 5, 1 (1965), 77–97, fordítás: USSR Computational & Mathematical Physics 5, 1, Pergamon Press, Oxford, 1967, 101–130.

[8] Kovács Győző: *Válogatott kalandozásaim Informatikában*, Gáma-Geo Kft. és Masszi Kiadó, Budapest, 2002, 235–236. o.

[9] Ny. Bezrukov: *Softpanorama: (slightly skeptical) Open Source Software Educational Society*, `http://www.softpanorama.org/`

[10] Németh László: *Magyar Ispell dokumentácó*, `http://magyarispell.sourceforge.net/magyarispell.pdf`

[11] Halácsy P., Kornai A., Németh L., Trón V.: *Cleaning large corpora*, kézirat, benyújtva: DIMACS Workshop on Data Quality, Data Cleaning and Treatment of Noisy Data, 2003