A Szószablya projekt – www.szoszablya.hu

Halácsy Péter[1], Kornai András[2], Németh László[1], Rung András[3], Szakadát István[1], and Trón Viktor[4]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem, Média Oktatási és Kutató Központ, {halacsy,szakadat,rung}@mokk.bme.hu
[2] MetaCarta Inc., andras@kornai.com
[3] Budapesti Műszaki és Gazdaságtudományi Egyetem, Kognitív Tudományi Központ, rung@itm.bme.hu
[4] International Graduate College of Language Technology and Cognitive Systems Saarland University – University of Edinburgh
v.tron@ed.ac.uk

A 2003 márciusában indult Szószablya projekt[5] célja, hogy létrehozzuk a *Magyar Webkorpuszt* — egy minden korábbinál nagyságrenddel nagyobb méretű magyar nyelvű tokenizált szöveggyűjteményt —, az ez alapján készülő *Szószablya Gyakorisági Szótárat*, a szabadon elérhető (LGPL licencű) *hunmorph* morfológiai elemzőt, a *hunstem* szótövezőt és a *hunspell* helyesírás-ellenőrzőt és a programok által használt *hunlex* magyar helyesírási és morfológiai szótárat.

Az egyedülálló teljességű Magyar Webkorpusz alapanyagát – 2,4 millió weboldalt, 700 millió szövegszó (token) és 13 millió különböző szóalak (type) – a magyar webről (a .hu tartományból) gyűjtöttük 2002 decemberében a Larbin webcrawler programmal. A weboldalakat normalizáltuk és a nyers szövegtartalmat, valamint mondatonra és szótokenekre bontottuk. Teljesen automatikus módszerekkel a weboldal gyűjteményből 433 ezer jó minőségű magyar dokumentum került kiválasztásra (113 millió szövegszó, 4.5 millió szóalak). 2003 decemberében már elérhető lesz a Magyar Webkorpusz és a Szószablya Gyakorisági Szótár újabb verziója, amely ennél még egy nagyságrenddel nagyobb szövegmintán alapul.

A nyers és a kiválogatott korpuszok alapján elkészítettük a gyakorisági szótár két verzióját. Ezek tartalmazzák a szavak szövegszó- és dokumentum-gyakoriságát. A szótárakban megjelöltük azt 4 millió (a nyers szótárban) és 2,8 millió szót (a válogatott szótárban), amelyet a hunspell helyesírás-ellenőrző aktuális verziója helyesnek fogad el. A fel nem ismert szavak alapján megkezdődött a hunlex szótár intenzív bővítése. Becslésünk szerint a hunspell jelenlegi verziója a magyar weboldalakon lévő helyes szóalakok legalább 96%-t felismeri.

A munka kezdetekor már rendelkezésünkre állt a hunspell első verziója, hiszen az a 2002 óta fejlesztett Magyar MySpell rendszer továbbfejlesztett változata. A hunmorph és a hunstem programok is ennek a kódjára alapulnak majd, tervezésük folyamatban van. Az első verziókat a projekt befejeztével, 2004 májusában adjuk közre. Míg a hunspell szigorúan betartja a helyesírási szabályokat, addig a hunmorph jellemzője, hogy képes elemezni a nem (teljesen) helyes alakokat is.

[5] A projektet a Budapest Műszaki Egyetem Média Oktató és Kutató Központja vezeti, az Informatikai és Hírközlési Minisztérium (az ITEM 2002 pályázat keretében), a MATÁV Rt. és az [origo] támogatja.