Az OpenOffice.org irodai program nyelvi eszközei

Németh László

BME – Média Oktató és Kutató Központ
1111 Budapest, Stoczek u. 2.
Nemeth@MOKK.BME.hu

Kivonat: Az OpenOffice.org irodai programcsomag Lingucomponent modulja és az OpenOffice.org natív nyelvi fejlesztései nyelvi eszközöket, nagyobbrészt helyesírási szótárakat biztosítanak jelenleg mintegy 70-80 nyelvhez. Az OpenOffice.org irodai programcsomag képességeinek növekedésével, például az új Thesaurus komponens vagy a Hunmorph morfológiai elemző beépítésével a helyesírási szótárak mellett a szinonimaszótárak és morfológiai elemzésre is használható erőforrások fejlesztése is elkezdődött. Az OpenOffice.org növekvő népszerűségére, a nyílt forráskódú fejlesztési modellre, valamint a nyelvi erőforrások elkészítésének és karbantartásának automatizálására alapozva lehetővé válhat a nyelvi eszközök és erőforrások folyamatos fejlesztése.

1. Bevezetés

Az OpenOffice.org [8] a legjelentősebb nyílt forráskódú irodai programcsomag. Az OpenOffice.org fejlesztése sok különböző projektre különül el. A Lingucomponent projekt foglalkozik a számítógépes szövegbevitelt és szövegfeldolgozást segítő nyelvtechnológiai fejlesztésekkel. A Lingucomponent projekt célkitűzése, hogy versenyképes nyelvi eszközöket (szó- és mondatszintű helyesírás-ellenőrző, elválasztó program, szinonimaszótár, stb.) biztosítson minél több nyelvhez [1]. A feladat tehát kettős: az alkalmazások és a nyelvi erőforrások fejlesztését is jelenti. Mintegy 70-80 nyelvhez érhető el OpenOffice.org nyelvi erőforrás. Ezek az erőforrások legalább a helyesírás-ellenőrzéshez használt egynyelvű szótárat artalmazzák, legtöbbször a nyelv morfológiáját tükröző *affixumtömörítéssel* [6].

A magyar nyelv támogatása lényegesen javulni fog az OpenOffice.org 2.0-s változatában, mivel az irodai programcsomag helyesírás-ellenőrzőjét felváltja a magyar fejlesztésű Hunspell [4]. A Hunspell támogatja a bonyolult morfológiát és összetettszó-kezelést, valamint az Unicode karakterkódolást is, így nemcsak a magyar, hanem sok más nyelv kezelését is lehetővé teszi. A Hunspell-lel tucatnyi ázsiai és afrikai Aspell erőforrás válik elérhetővé az OpenOffice.org számára. Már az OpenOffice.org integráció befejeződése előtt elkezdődött az új Hunspell erőforrások fejlesztése (például a nepáli, moszi, akan nyelvekhez), illetve a meglévő erőforrások javítása (a finn a morfológia, a német és az afrikaans az összetettszó-kezelés terén használja ki a Hunspell képességeit).

2. OpenOffice.org

Az OpenOffice.org irodai programcsomag a Microsoft Office kiváló ingyenes alternatívájaként ismert hazánkban. Az OpenOffice.org kódbázisára épül számos egyéb kereskedelmi irodai programcsomag is, mint a Sun StarOffice vagy az EuroOffice (korábban MagyarOffice), illetve része az IBM Workplace programcsomagjának is.

Az OpenOffice.org jelentősége azonban túlmutat az alternatíva és a kódbázis szerepen [5]: a platformfüggetlenségnek, a nyitott szabványok támogatásának, a nyílt forráskódnak és fejlesztési modellnek, továbbá a natív nyelvi projektek támogatásának köszönhetően jóval szélesebb közönséget céloz meg, mint az egyes kereskedelmi programváltozatok.

3. Natív nyelvi projektek

Körülbelül 60 natív nyelvi projektje van az OpenOffice.org-nak, pár ezer közreműködővel. További negyven natív nyelvi projekt létrejöttére számít a közösség egy éven belül.

4. Lingucomponent projekt

A Lingucomponent projekthez tartozó jelenlegi OpenOffice.org alkalmazások: elválasztó komponens (AltLinux), szószintű helyesírás-ellenőrző (MySpell, a 2.0.1-es változattól Hunspell) és szinonimaszótár (OpenOffice.org Thesaurus). Az elválasztó komponens D. E. Knuth nevezetes szedőprogramja, a TeX elválasztási algoritmusán alapul. A Hunspell program az Ispell helyesírás-ellenőrző családba tartozik, így képes az Ispell, Aspell és MySpell szótárak használatára.

Az alkalmazásokat érintő jelentősebb tervezett fejlesztések: morfológiai elemző illesztése az elválasztó modulhoz és a szinonimaszótárhoz, a helyesírás-ellenőrző összetettszó-kezelésének általánosítása és a kiejtési hasonlóságon alapuló javaslattevés implementálása, mondatszintű ellenőrzés megvalósítása.

Több más fejlesztés is kapcsolódik közvetve a Lingucomponent projekthez. Ilyen az OpenThesaurus projekt [7], amely egy nyílt keretrendszer Wordnettel szinkronizált szinonimaszótárak fejlesztéséhez (már 5-6 nyelv esetében használják OpenOffice.org erőforrások létrehozására), valamint az An Crúbadán projekt [10], amely már 144 nyelv webkorpuszát gyűjti folyamatosan az Internetről. A webkorpuszok megfelelő kiindulási alapot jelentenek az OpenOffice.org nyelvi erőforrások készítéséhez.

Míg nyelvtechnológiai vonatkozásban a hiányzó alkalmazások és képességek megtervezése és kifejlesztése a feladat, a nyelvi erőforrások esetében a natív nyelvi projektek munkájának összehangolása, minőségbiztosítása, a nyelvészeti munkát egységes keretbe foglaló fejlesztői környezet kialakítása a cél, ami jelentősen megkönnyíti a több száz nyelvierőforrás-fejlesztő munkáját.

A Lingucomponent projekt eredményei nem csak az irodai programcsomag felhasználói, hanem a kutatói közösség számára is fontosak, így a jövőben számítunk az alapkutatásban érintettek (nyelvészek, számítógépes nyelvészek, szoftverergonómusok) fokozott részvételére.

Magyar vonatkozású fejlesztések

Bár az elválasztó modul morfológiailag nem elemzi még a szavakat, a Huhyphn elválasztási szabálygyűjtemény [2] tartalmazza az eddigi legteljesebb magyar elválasztási kivételszótárat.

A magyar Hunspell erőforrás [3] új, unicode-os változata régi hiányosságot pótol: a helyesírási szabályzat előírja az idegen ékezetes latin betűk használatát az idegen szavakban, de ezt eddig nem támogatták a magyar helyesírás-ellenőrzők. Az Unicode és a bonyolult szóösszetételek (például földrajzi nevek) támogatása fontos előrelépést jelent a különféle szakszövegek gondozásában.

A magyar morfológiai elemző [12] a toldalékolt szavak kezeléséhez és a mondatszintű elemzéshez nélkülözhetetlen. Az OpenOffice.org 2.0.1-es változata a magyar morfológiai elemzőt tövezésre, és a szinonimák toldalékolt alakjainak előállítására fogja használni.

Az OpenOffice.org magyar szinonimaszótárának OpenThesaurus alapú fejlesztése elkezdődött a tervezés szintjén. A közösségi fejlesztési modell csak aktív résztvevők megléte esetén működőképes, de a német OpenThesaurus [9] és a magyar Wikipédia [13] sikere biztató a magyar OpenThesaurus jövőjére nézve is.

Bibliográfia

1. Lingucomponent projekt: http://lingucomponent.openoffice.org
2. Nagy Bence: Huhyphn – Magyar elválasztás TeX-hez, Scribushoz, OpenOffice.org-hoz, 2003, http://www.tug.org/tex-archive/language/hungarian/**huhyphn**.pdf
3. Németh László: Magyar Ispell – Válasz a Helyes-e?-re, IV. GNU/Linux szakmai konferencia, LME, Budapest, 2002, 99–107. o., http://mek.oszk.hu/01200/01240/
4. Németh László: A Szószablya fejlesztés, 2003, V. GNU/Linux szakmai konferencia, LME, Budapest, 2003, 103–108. o., http://mek.oszk.hu/02200/02230/
5. Miguel de Icaza: The Global Importance of OpenOffice.org (nyitóbeszéd és prezentáció), OpenOffice.org konf., Berlin, 2004, http://marketing.openoffice.org/conference/thursday.html
6. Geoff Kuennings: International Ispell: http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html
7. Daniel Naber: OpenThesaurus: Building a Thesaurus with a Web Community, 2004, http://www.openthesaurus.de/download/openthesaurus.pdf
8. OpenOffice.org: http://www.openoffice.org
9. OpenThesaurus: http://www.openthesaurus.de
10. Kevin P. Scannel: An Crúbadán, Corpus building for minority languages http://borel.slu.edu/crubadan/
12. Trón, V, Németh, L, Halácsy, P, Kornai, A, Gyepesi, G, and Varga, D. Hunmorph: open source word analysis, In: Proceeding of ACL. ACL. (2005)
13. Wikipédia: http://hu.openoffice.org